ENCYCLOPEDIA OF
MODERN OPTICS
SECOND EDITION

EDITORS IN CHIEF

Bob D. Guenther

Duke University, Durham, NC, United States

Duncan G. Steel

University of Michigan, Ann Arbor, MI, United States

VOLUME 2

Spectroscopy ■ Terahertz ■ Optics of Semiconductors and 2D Materials ■ Nonlinear Optical Spectroscopy ■ Metamaterials and Plasmonics ■ Lasers, UV Lasers, Random lasers

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

ISBN 978-0-12-809283-5

Publisher: Oliver Walter
Acquisition Editor: Ruth Ireland
Content Project Manager: Sean Simms
Associate Content Project Manager: Marise Willis
Designer: Greg Harris

Printed and bound in the United Kingdom

EDITORIAL BOARD

CONTENTS OF VOLUME 2

LIST OF CONTRIBUTORS FOR VOLUME 2

Ladan Arissian
University of New Mexico, Albuquerque, NM, United States

Marc Assmann
Technical University of Dortmund, Dortmund, Germany

Manfred Bayer
Technical University of Dortmund, Dortmund, Germany

Andreas Becker
University of Colorado, Boulder, CO, United States

S Bhandarkar
Alfred University, Alfred, NY, USA

P Blood
Cardiff University, Cardiff, UK

A Brenier
University of Lyon, Villeurbanne, France

WB Bridges
California Institute of Technology, Pasadena, CA, USA

Lauren E Buchanan
Vanderbilt University, Nashville, TN, United States

Francesca Calegari
Center for Free-Electron Laser Science, Hamburg, Germany; University of Hamburg, Hamburg, Germany; and Institute for Photonics and Nanotechnologies CNR-IFN, Milano, Italy

Giulio Cerullo
Institute of Photonics and Nanotechnology, Milano, Italy

CR Chatwin
University of Sussex, Brighton, UK

JJ Coleman
University of Illinois, Urbana, IL, USA

Kankan Cong
Rice University, Houston, TX, United States

A Costela
Consejo Superior de Investigaciones Cientificas, Madrid, Spain

DW Coutts
University of Oxford, Oxford, UK

Juraj Darmo
Vienna University of Technology, Vienna, Austria

Alberto de la Torre
California Institute of Technology, Pasadena, CA, United States

Sandro De Silvestri
Institute of Photonics and Nanotechnology, Milano, Italy

Jean-Claude Diels
University of New Mexico, Allmquerque, NM, United States

FJ Duarte
Eastman Kodak Company, New York, NY, USA

Thomas Elsaesser
Max Born Institute for Nonlinear Optics and Short-Pulse Spectroscopy, Berlin, Germany

JJ Ewing
Ewing Technology Associates, Inc., Bellevue, WA, USA

I Galbraith
Heriot-Watt University, Edinburgh, UK

John Harter
University of California, Santa Barbara, CA, United States

Martin R Hofmann
Ruhr University Bochum, Bochum, Germany

David Hsieh
California Institute of Technology, Pasadena, CA, United States

Rupert Huber
University of Regensburg, Regensburg, Germany

Ulrich Huttner
Philipps University of Marburg, Marburg, Germany

Agnieszka Jaroń-Becker
University of Colorado, Boulder, CO, United States

Yushi Kaneda
The University of Arizona, Tucson, AZ, United States

Mackillo Kira
University of Michigan, Ann Arbor, MI, United States

Stephan W Koch
Philipps University of Marburg, Marburg, Germany

AG Kofman
Weizmann Institute of Science, Rehovot, Israel

Junichiro Kono
Rice University, Houston, TX, United States

G Kurizki
Weizmann Institute of Science, Rehovot, Israel

PT Landsberg
The University of Southampton, Southampton, UK

Mikhail Larionov
Dausinger + Giesen GmbH, Stuttgart, Germany

Alexandre Laurain
University of Arizona, Tucson, AZ, United States

Xiaoqin Li
The University of Texas at Austin, Austin, TX, United States

Bachana Lomsadze
University of Michigan, Ann Arbor, MI, United States

Cristian Manzoni
Institute of Photonics and Nanotechnology, Milano, Italy

SR Meech
University of East Anglia, Norwich, UK

Jerome V Moloney
University of Arizona, Tucson, AZ, United States

Galan Moody
National Institute of Standards & Technology, Boulder, CO, United States

GA Mourou
University of Michigan, Ann Arbor, MI, USA

Gururaj V Naik
Rice University, Houston, TX, United States

G Timothy Noe II
Rice University, Houston, TX, United States

Jennifer P Ogilvie
University of Michigan, Ann Arbor, MI, United States

D Petrosyan
Institute of Electronic Structure and Laser, Heraklion, Greece

Rohit P Prasankumar
Center for Integrated Nanotechnologies, Los Alamos, NM, United States

Klaus Reimann
Max Born Institute for Nonlinear Optics and Short-Pulse Spectroscopy, Berlin, Germany

Marco Santagiustina
University of Padova, Padova, Italy

Kenneth L Schepler
University of Central Florida, Orlando, FL, United States

J Shah
Arlington, VA, USA

Mark Siemens
University of Denver, Denver, CO, United States

Yin Song
University of Michigan, Ann Arbor, MI, United States

Antoinette J Taylor
Center for Integrated Nanotechnologies, Los Alamos, NM, United States

James R Taylor
Imperial College London, London, United Kingdom

Darius Torchinsky
Temple University, Philadelphia, PA, United States

Karl Unterrainer
Vienna University of Technology, Vienna, Austria

E Vauthey
University of Geneva, Geneva, Switzerland

Caterina Vozzi
Institute for Photonics and Nanotechnologies CNR-IFN, Milano, Italy

L Wang
Chinese Academy of Sciences, Beijing, China

Michael Woerner
Max Born Institute for Nonlinear Optics and Short-Pulse Spectroscopy, Berlin, Germany

Wei Xiong
University of California, San Diego, CA, United States

Dmitry A Yarotski
Center for Integrated Nanotechnologies, Los Alamos, NM, United States

Harumasa Yoshida
Mie University, Tsu, Mie, Japan

W Zawadzki
Polish Academy of Sciences, Warsaw, Poland

John J Zayhowski
Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, MA, United States

Liuyan Zhao
University of Michigan, Ann Arbor, MI, United States

Xiushan Zhu
University of Arizona, Tucson AZ, United States

EDITORS IN CHIEF

Bob D. Guenther received his undergraduate degree from Baylor University and his graduate degrees in Physics from University of Missouri. He has had research experience in condensed matter and optical physics. For 9 years he was active in research management as a Senior Executive in the Army, responsible for the physics research sponsored by the Army. After retiring from the government he held the position of Interim Director of the Free Electron Laser Laboratory and helped establish Duke's Fitzpatrick Center for Photonics and Communication Systems and served as Executive Director of the Center until his retirement. In a continuation of the retirement process, he moved to Applied Quantum Technology and has just retired from that company. He is author of the textbook, *Modern Optics*, second edition. He is now composing an elementary book in optics.

Duncan G. Steel is the Robert J. Hiller Professor of Electrical and Computer Engineering, as well as Professor of Physics and Biophysics. Prior to joining the faculty of the University of Michigan in 1985, he was a senior research scientist at Hughes Aircraft Company at the Hughes Research Laboratories. At the University of Michigan, he was Area Chair and Director of the Optical Sciences Laboratory for 20 years until 2007 when he took over the position of Chair of the Biophysics Research Division during its transition to an academic unit. As an educator and teacher, he has chaired or cochaired over 60 doctoral committees. His research includes coherent optical studies of semiconductors and their application to quantum information. His work also included 30 years of studies on age-related modifications in proteins where he exploited numerous optical techniques including single molecule studies in neurons in his studies on Alzheimer's Disease. He was a Guggenheim Scholar and received the 2010 Isakson Prize from the American Physical Society.

INTRODUCTION

This is the second, updated edition of the *Encyclopedia of Modern Optics*. There are 197 entries, many of them new or updated, reflecting the enormous progress in the optical sciences and technology and the ever-expanding impact since the publication of the first edition. Some of the new topics are:

- Nano-photonics and Plasmonics
- Quantum Optics
- Quantum Information
- Optical Interconnects
- Photonic Crystals and Their Applications
- High Efficiency LED's
- Displays
- Transformation Optics
- Fiber Lasers
- Terahertz
- Multidimensional Spectroscopy
- Organic Optoelectronics
- Gravitational Wave Detectors
- Meta Materials and Plasmonics

Selection of article topics and recruiting authors for those topics in this edition has been the work of the topical editors listed in the prologue. They were able to solicit contributions from internationally recognized leaders in their field.

The entries of the encyclopedia are arranged by subject as best as possible, a task made difficult because the field is now highly interdisciplinary and there are many subjects that have an impact in many different areas. We have added references to other articles so that readers can obtain a deeper understanding of the material or understand how a specific discussion on basic science may impact an application or technology.

Transient Holographic Grating Techniques in Chemical Dynamics

E Vauthey, University of Geneva, Geneva, Switzerland

Nomenclature

C concentration [mol L^{-1}]
C_v heat capacity [J K^{-1} kg^{-1}]
d sample thickness [m]
D_{th} thermal diffusivity [m^2 s^{-1}]
I_{dif} diffracted intensity [W m^{-2}]
I_{pr} intensity of the probe pulse [W m^{-2}]
I_{pu} intensity of a pump pulse [W m^{-2}]
$\vec{k}_{ac}$ acoustic wavevector [m^{-1}]
k_r rate constant of a heat releasing process [s^{-1}]
k_{re} rate constant of reorientation [s^{-1}]
k_{th} rate constant of thermal diffusion [s^{-1}]
K attenuation constant
n refractive index
$\tilde{n}$ complex refractive index
N number of molecule per unit volume [m^{-3}]
r polarization anisotropy
R fourth rank response tensor [m^2 V^{-2}]
v_s speed of sound [m s^{-1}]
V volume [m^3]
α_{ac} acoustic attenuation constant [m^{-1}]
β cubic volume expansion coefficient [K^{-1}]
γ angle between transition dipoles
Λ fringe spacing [m]
Δn_d variation of refractive index due to density changes
Δn_d^e Δn_d due to electrostriction
Δn_K variation of refractive index due to optical Kerr effect
Δn_d^p Δn_d due to volume changes
Δn_p variation of refractive index due to population changes
Δn_d^t Δn_d due to temperature changes
Δx peak to null variation of the parameter x
ε molar decadic absorption coefficient [cm L mol^{-1}]
ζ angle of polarization of the diffracted signal
η diffraction efficiency
θ_B Bragg angle (angle of incidence of the probe pulse)
θ_{pu} angle of incidence of a pump pulse
λ_{pr} probe wavelength [m]
λ_{pu} pump wavelength [m]
ρ density [kg m^{-3}]
τ_{ac} acoustic period [s]
υ_{ac} acoustic frequency [s^{-1}]

Introduction

Over the past two decades, holographic techniques have proved to be valuable tools for investigating the dynamics of chemical processes. The aim of this chapter is to give an overview of the main applications of these techniques for chemical dynamics. The basic principle underlying the formation and the detection of elementary transient holograms, also called transient gratings, is first presented. This is followed by a brief description of a typical experimental setup. The main applications of these techniques to solve chemical problems are then discussed.

Basic Principle

The basic principle of the transient holographic technique is illustrated in Fig. 1. The sample material is excited by two laser pulses at the same wavelength and crossed at an angle $2\theta_{pu}$. If the two pump pulses have the same intensity, I_{pu}, the intensity distribution at the interference region, assuming plane waves, is

$$I(x) = 2I_{\mathrm{pu}}\left[1 + \cos\left(\frac{2\pi x}{\Lambda}\right)\right] \tag{1}$$

where $\Lambda = \lambda_{pu}/(2\sin\theta_{pu})$ is the fringe spacing, λ_{pu} is the pump wavelength, and I_{pu} is the intensity of one pump pulse.

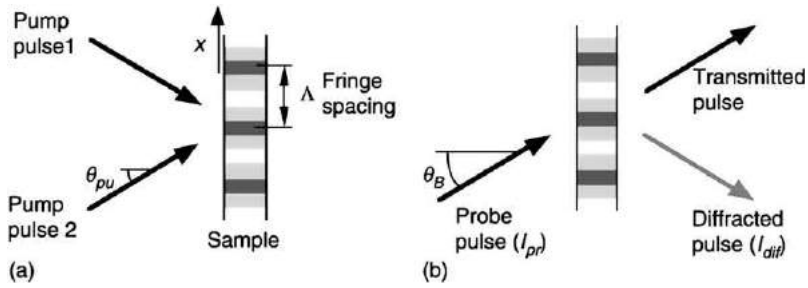

Fig. 1 Principle of the transient grating technique: (a) grating formation (pumping), (b) grating detection (probing).

As discussed below, there can be many types of light–matter interactions that lead to a change in the optical properties of the material. For a dielectric material, they result in a spatial modulation of the optical susceptibility and thus of the complex refractive index, $\tilde{n}$ The latter distribution can be described as a Fourier cosine series:

$$\tilde{n}(x) = \tilde{n}_0 + \sum_{m=1}^{\infty} \tilde{n}_m \cos\left(\frac{m2\pi x}{\Lambda}\right) \tag{2}$$

where $\tilde{n}_0$ is the average value of $\tilde{n}$ In the absence of saturation effects, the spatial modulation of $\tilde{n}$ is harmonic and the Fourier coefficients with $m>1$ vanish. In this case, the peak to null variation of the complex refractive index, $\Delta\tilde{n}$, is equal to the Fourier coefficient $\tilde{n}_1$ The complex refractive index can be split into its real and imaginary components:

$$\tilde{n} = n + iK \tag{3}$$

where n is the refractive index and K is the attenuation constant.

The hologram created by the interaction of the crossed pump pulses consists in periodic, one-dimensional spatial modulations of n and K. Such distributions are nothing but phase and amplitude gratings, respectively. A third laser beam at the probe wavelength, λ_{pr}, striking these gratings at Bragg angle, $\theta_B = \arcsin(\lambda_{pr}/2\Lambda)$, will thus be partially diffracted (Fig. 1(b)). The diffraction efficiency, η, depends on the modulation amplitude of the optical properties. In the limit of small diffraction efficiency ($\eta<0.01$), this relationship is given by

$$\eta = \frac{I_{\text{dif}}}{I_{\text{pr}}} \cong \left[\left(\frac{\ln 10\Delta A}{4\cos\theta_B}\right)^2 + \left(\frac{\pi d\Delta n}{\lambda_{\text{pr}}\cos\theta_B}\right)^2\right] \times \exp\left(-\frac{\ln 10A}{\cos\theta_B}\right) \tag{4}$$

where I_{dif} and I_{pr} are the diffracted and the probe intensity respectively, d is the sample thickness, and $A = 4\pi dK/(\lambda \ln 10)$ is the average absorbance.

The first and the second terms in the square bracket describe the contribution of the amplitude and phase gratings respectively, and the exponential term accounts for the reabsorption of the diffracted beam by the sample.

The main processes responsible for the variation of the optical properties of an isotropic dielectric material are summarized in Fig. 2.

The modulation of the absorbance, ΔA, is essentially due to the photoinduced concentration change, ΔC, of the different chemical species i (excited state, photoproduct, …):

$$\Delta A(\lambda_{\text{pr}}) = \sum_i \varepsilon_i(\lambda_{\text{pr}})\Delta C_i \tag{5}$$

where ε_i is the absorption coefficient of the species i.

The variation of the refractive index, Δn, has several origins and can be expressed as

$$\Delta n = \Delta n_K + \Delta n_p + \Delta n_d \tag{6}$$

Δn_K is the variation of refractive index due to the optical Kerr effect (OKE). This nonresonant interaction results in an electronic polarization (electronic OKE) and/or in a nuclear reorientation of the molecules (nuclear OKE) along the direction of the electric field associated with the pump pulses. As a consequence, a transient birefringence is created in the material. This effect is usually discussed within the framework of nonlinear optics in terms of intensity dependent refractive index or third order nonlinear

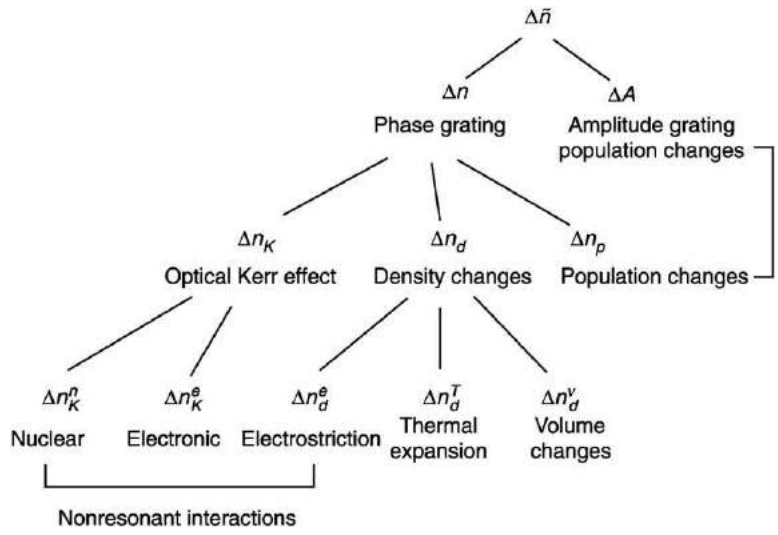

Fig. 2 Classification of the possible contributions to a transient grating signal.

susceptibility. Electronic OKE occurs in any dielectric material under sufficiently high light intensity. On the other hand, nuclear OKE is mostly observed in liquids and gases and depends strongly on the molecular shape.

Δn_p is the change of refractive index related to population changes. Its magnitude and wavelength dependence can be obtained by Kramers–Kronig transformation of $\Delta A(\lambda)$ or $\Delta K(\lambda)$:

$$\Delta n_p(\lambda) = \frac{1}{2\pi^2}\int_0^\infty \frac{\Delta K(\lambda')}{1-(\lambda'/\lambda)^2}\,\mathrm{d}\lambda' \tag{7}$$

Δn_d is the change of refractive index associated with density changes. Density phase gratings can have essentially three origins:

$$\Delta n_d = \Delta n_d^t + \Delta n_d^v + \Delta n_d^e \tag{8}$$

Δn_d^t is related to the temperature-induced change of density. If a fraction of the excitation energy is converted into heat, through a nonradiative transition or an exothermic process, the temperature becomes spatially modulated. This results in a variation of density, hence to a modulation of refractive index with amplitude Δn_d^t. Most of the temperature dependence of n originates from the density. The temperature-induced variation of n at constant density is much smaller than Δn_d^t.

Δn_d^v is related to the variation of volume upon population changes. This volume comprises not only the reactant and product molecules but also their environment. For example, in the case of a photodissociation, the volume of the product is larger than that of the reactant and a positive volume change can be expected. This will lead to a decrease of the density and to a negative Δn_d^v.

Finally, Δn_d^e is related to electrostriction in the sample by the electric field of the pump pulses. Like OKE, this is a nonresonant process that also contributes to the intensity dependent refractive index. Electrostriction leads to material compression in the regions of high electric field strength. The periodic compression is accompanied by the generation of two counterpropagating acoustic waves with wave vectors, $\vec{k}_{ac} = \pm(2\pi/\Lambda)\vec{i}$, where $\vec{i}$ is the unit vector along the modulation axis. The interference of these acoustic waves leads to a temporal modulation of Δn_d^e at the acoustic frequency υ_{ac}, with $2\pi\upsilon_{ac}=k_{ac}v_s$, v_s being the speed of sound. As Δn_d^e oscillates between negative and positive values, the diffracted intensity, which is proportional to $\left(\Delta n_d^e\right)^2$, shows at temporal oscillation at twice the acoustic frequency. In most cases, Δn_d^e is weak and can be neglected if the pump pulses are within an absorption band of the sample.

The modulation amplitudes of absorbance and refractive index are not constant in time and their temporal behavior depends on various dynamic processes in the sample. The whole point of the transient grating techniques is precisely the measurement of the diffracted intensity as a function of time after excitation to deduce dynamic information on the system.

In the following, we will show that the various processes shown in Fig. 2, that give rise to a diffracted signal, can in principle be separated by choosing appropriate experimental parameters, such as the timescale, the probe wavelength, the polarization of the four beams, or the crossing angle.

Experimental Setup

A typical experimental arrangement for pump–probe transient grating measurements is shown in Fig. 3. The laser output pulses are split in three parts. Two parts of equal intensity are used as pump pulses and are crossed in the sample. In order to ensure time coincidence, one pump pulse travels along an adjustable optical delay line. The third part, which is used for probing, can be frequency converted using a dye laser, a nonlinear crystal, a Raman shifter, or white light continuum generation. The probe pulse is sent along a motorized optical delay line before striking the sample at the Bragg angle. There are several possible beam configurations for transient grating and the two most used are illustrated in Fig. 4. When the probe and pump pulses are at different wavelengths, they can be in the same plane of incidence as shown in Fig. 4(a). However, if the pump and probe wavelengths are the same, the folded boxcars geometry shown in Fig. 4(b) has to be used. The transient grating technique is background free and the diffracted signal propagates in a well-defined direction. In a pump–probe experiment, the diffracted signal intensity is measured as a function of the time delay between the pump and probe pulses. Such a setup can be used to probe dynamic processes occurring in timescales going from a few fs to a few ns, the time resolution depending essentially on the duration of the pump and probe pulses. For slower processes, the grating dynamics can be probed in real time with a *cw* laser beam and a fast photodetector.

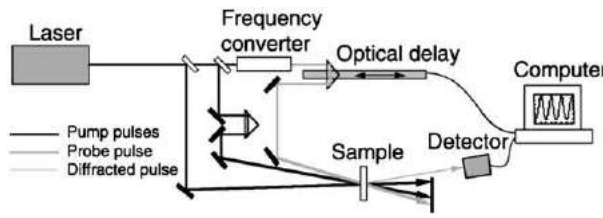

Fig. 3 Schematic of a transient grating setup with pump–probe detection.

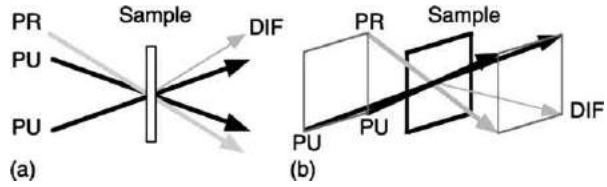

Fig. 4 Beam geometry for transient grating: (a) in plane; (b) boxcars.

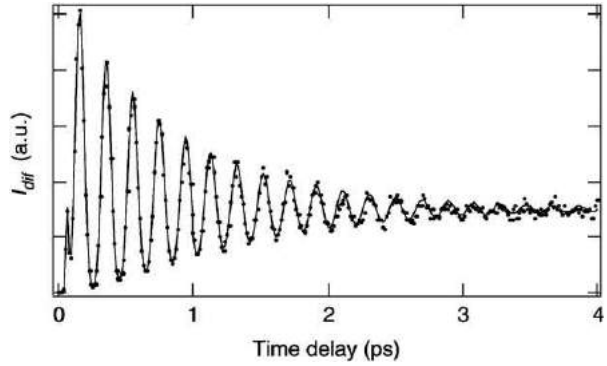

Fig. 5 Time profile of the diffracted intensity measured with a solution of malachite green after excitation with two pulses with close to counterpropagating geometry.

Applications

The Transient Density Phase Grating Technique

If the grating is probed at a wavelength far from any absorption band, the variation of absorbance, $\Delta A(\lambda_{pr})$, is zero and the corresponding change of refractive index, $\Delta n_p(\lambda_{pr})$, is negligibly small. In this case, **Eq. (4)** simplifies to

$$\eta \cong \left(\frac{\pi d}{\lambda_{\mathrm{pr}} \cos\theta_B}\right)^2 \times \Delta n_d^2 \tag{9}$$

In principle, the diffracted signal may also contain contributions from the optical Kerr effect, Δn_K, but we will assume here that this non-resonant and ultrafast response is negligibly small. The density change can originate from both heat releasing processes and volume differences between the products and reactants, the former contribution usually being much larger than the latter. Even if the heat releasing process is instantaneous, the risetime of the density grating is limited by thermal expansion. This expansion is accompanied by the generation of two counterpropagating acoustic waves with wave vectors, $\vec{k}_{\mathrm{ac}} = \pm(2\pi/\Lambda)\vec{i}$ One can distinguish two density gratings: 1) a diffusive density grating, which reproduces the spatial distribution of temperature and which decays by thermal diffusion; and 2) an acoustic density grating originating from the standing acoustic wave and whose amplitude oscillates at the acoustic frequency υ_{ac}.

The amplitudes of these two gratings are equal but of opposite sign. Consequently, the time dependence of the modulation amplitude of the density phase grating is given by

$$\Delta n_d(t) = \left(\frac{\beta Q}{\rho C_v} + \Delta V\right)\rho\left(\frac{\partial n}{\partial \rho}\right)R(t) \tag{10a}$$

with

$$R(t) = 1 - \cos(2\pi\upsilon_{\mathrm{ac}}t) \times \exp(-\alpha_{\mathrm{ac}}\upsilon_s t) \tag{10b}$$

where ρ, β, C_v, and α_{ac} are the density, the volume expansion coefficient, the heat capacity, and the acoustic attenuation constant of the medium, respectively, Q is the amount of heat deposited during the photoinduced process, and ΔV is the corresponding volume change. As the standing acoustic wave oscillates, its corresponding density grating interferes with the diffusive density grating. Therefore, the total modulation amplitude of the density and thus Δn_d exhibits the oscillation at υ_{ac}. **Fig. 5** shows the time profile of the diffracted intensity measured with a solution of malachite green. After excitation to the S_1 state, this dye relaxes nonradiatively to the ground state in a few ps. For this measurement, the sample solution was excited with two 30 ps laser pulses at 532 nm crossed with an angle close to 180°. The continuous line is the best fit of **Eqs. (9)** and **(10)**. The damping of the oscillation is due to acoustic attenuation. After complete damping, the remaining diffracted signal is due to the diffusive density grating only.

$R(t)$ can be considered as the response function of the sample to a prompt heat release and/or volume change. If these processes are not instantaneous compared to an acoustic period $(\tau_{ac} = \upsilon_{ac}^{-1})$, the acoustic waves are not created impulsively and the time dependence of Δn_d is

$$\Delta n_d(t) = \left(\frac{\beta Q}{\rho C_v} + \Delta V\right)\rho\left(\frac{\partial n}{\partial \rho}\right)F(t) \tag{11a}$$

with

$$F(t) = \int_{-\infty}^{t} R(t - t') \cdot f(t')\mathrm{d}t' \tag{11b}$$

where $f(t)$ is a normalized function describing the time evolution of the temperature and/or volume change. In many cases, $f(t) = \exp(-k_r t)$, k_r being the rate constant of the process responsible for the change. **Fig. 6** shows the time profile of the diffracted intensity calculated with **Eqs. (9)** and **(11)** for different values of k_r.

If several processes take place, the total change of refractive index is the sum of the changes due to the individual processes. In this case, Δn_d should be expressed as

$$\Delta n_d(t) = \sum_i \left(\frac{\beta Q_i}{\rho C_v} + \Delta V_i\right)\rho\left(\frac{\partial n}{\partial \rho}\right)F_i(t) \tag{12}$$

The separation of the thermal and volume contributions to the diffracted signal is problematic. Several approaches have been proposed, the most used being the measurement in a series of solvents with different expansion coefficient β. This method requires that the other solvent properties like refractive index, dielectric constant, or viscosity, are constant or that the energetics of the system investigated does not depend on them. This separation is easier when working with water because its β value vanishes at 4 °C. At this temperature, the density variations are due to the volume changes only.

The above equations describe the growth of the density phase grating. However, this grating is not permanent and decays through diffusive processes. The phase grating originating from thermal expansion decays via thermal diffusion with a rate constant k_{th} given by

$$k_{\mathrm{th}} = D_{\mathrm{th}}\left(\frac{2\pi}{\Lambda}\right)^2 \tag{13}$$

where D_{th} is the thermal diffusivity. **Table 1** shows k_{th} values in acetonitrile for different crossing angles.

The decay of the phase grating originating from volume changes depends on the dynamics of the population responsible for ΔV (vide infra).

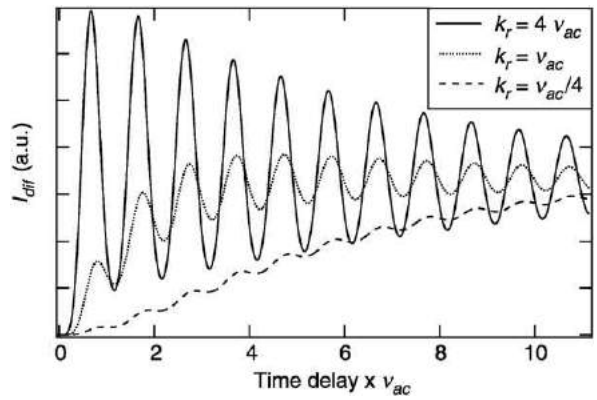

Fig. 6 Time profiles of the diffracted intensity calculated using **Eqs. (9)** and **(11)** with various k_r values.

Table 1 Fringe spacing, Λ, acoustic frequency, v_{ac}, thermal diffusion rate constant, k_{tr}, for various crossing angles of the pump pulses, 2 θ_{pu}, at 355 nm and in acetonitrile

$2\,\theta_{\mathrm{pu}}$	Λ (μm)	$\upsilon_{\mathrm{ac}}(\mathrm{s}^{-1})$	k_{th} (s^{-1})
0.5°	40.7	3.2×10^7	4.7×10^3
50°	0.42	3.1×10^9	4.5×10^7
180°	0.13	9.7×10^9	4.7×10^8

Time resolved optical calorimetry

A major application of the density phase grating in chemistry is the investigation of the energetics of photo-induced processes. The great advantage of this technique over other optical calorimetric methods, like the thermal lens or the photoacoustic spectroscopy, is its superior time resolution. The time constant of the fastest heat releasing process that can be time resolved with this technique is of the order of the acoustic period. The shortest acoustic period is achieved when forming the grating with two counter-propagating pump pulses ($2\theta_{pu}=180^{\circ}$). In this case the fringe spacing is $\Lambda=\lambda_{pu}/2n$ and, with UV pump pulses, an acoustic period of the order of 100 ps can be obtained in a typical organic solvent.

For example, Fig. 7(a) shows the energy diagram for a photoinduced electron transfer (ET) reaction between benzophenone (BP) and an electron donor (D) in a polar solvent. Upon excitation at 355 nm, ^{1}BP* undergoes intersystem-crossing (ISC) to ^{3}BP* with a time constant of about 10 ps. After diffusional encounter with the electron donor, ET takes place and a pair of ions is generated. With this system, the whole energy of the 355 nm photon ($E=3.49$ eV) is converted into heat. The different heat releasing processes can be differentiated according to their timescale. The vibrational relaxation to ^{1}BP* and the ensuing ISC to ^{3}BP* induces an ultrafast release of 0.49 eV as heat. With donor concentrations of the order of 0.1 M, the heat deposition process due to the electron transfer is typically in the ns range. Finally, the recombination of the ions produces a heat release in the

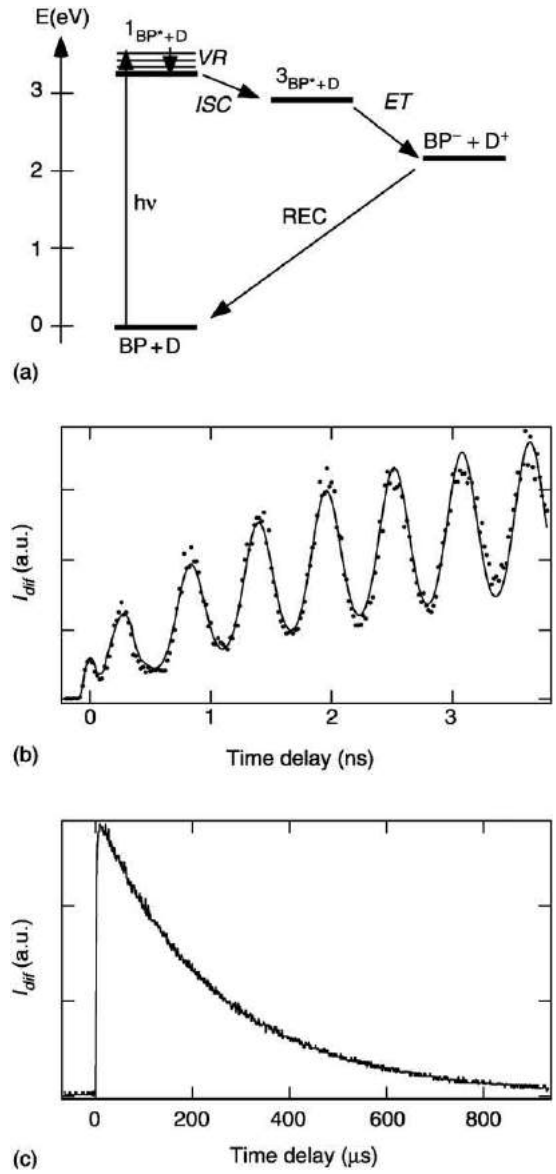

Fig. 7 (a) Energy diagram of the states involved in the photo-induced electron transfer reaction between benzophenone (BP) and an electron donor (D) (VR: vibrational relaxation, ISC: intersystem crossing; ET: electron transfer; REC: charge recombination). (b) Time profile of the diffracted intensity measured at 590 nm after excitation at 355 nm of a solution of BP and 0.05 M D. (c) Same as (b) but measured at 1064 nm with a *cw* laser.

microsecond timescale. Fig. 7(b) shows the time profile of the diffracted intensity measured after excitation at 355 nm of BP with 0.05 M D in acetonitrile. The 30 ps pump pulses were crossed at 27° (τ_{ac} = 565 ps) and the 10 ps probe pulses were at 590 nm. The oscillatory behavior is due to the ultrafast heat released upon formation of ^{3}BP*, while the slow dc rise is caused by the heat dissipated upon ET. As the amount of energy released in the ultrafast process is known, the energy released upon ET can be determined by comparing the amplitudes of the fast and slow components of the time profile. The energetics as well as the dynamics of the photoinduced ET process can thus be determined. Fig. 7(c) shows the decay of the diffracted intensity due to the washing out of the grating by thermal diffusion. As both the acoustic frequency and the thermal diffusion depends on the grating wavevector, the experimental time window in which a heat releasing process can be measured, depends mainly on the crossing angle of the pump pulses as shown in Table 1.

Determination of material properties

Another important application of the transient density phase grating technique is the investigation of material properties. Acoustics waves of various frequencies, depending on the pump wavelength and crossing angle, can be generated without physical contact with the sample. Therefore, acoustic properties, such as the speed of sound and the acoustic attenuation of the material, can be easily obtained from the frequency and the damping of the oscillation of the transient grating signal.

Similarly, the optoelastic constant of a material, $\rho\partial n/\partial\rho$, can be determined from the amplitude of the signal (see Eq. (11)). This is done by comparing the signal amplitude of the material under investigation with that obtained with a known standard.

Finally, the thermal diffusivity, D_{th}, can be easily obtained from the decay of the thermal density phase grating, such as that shown in Fig. 7(c). This technique can be used with a large variety of bulk materials as well as films, surfaces and interfaces.

Investigation of Population Dynamics

The dynamic properties of a photogenerated species can be investigated by using a probe wavelength within its absorption or dispersion spectrum. As shown by Eq. (4), either ΔA or Δn_p has to be different from zero. For this application, Δn_K and Δn_d should ideally be equal to zero. Unless working with ultrashort pulses (< 1 ps) and a weakly absorbing sample, Δn_K can be neglected. Practically, it is almost impossible to find a sample system for which some fraction of the energy absorbed as light is not released as heat. However, by using a sufficiently small crossing angle of the pump pulses, the formation of the density phase grating, which depends on the acoustic period, can take as much as 30 to 40 ns. In this case, Δn_d is negligible during the first few ns after excitation and the diffracted intensity is due to the population grating only:

$$I_{\mathrm{dif}}(t) \propto \sum_i \Delta C_i^2(t) \tag{14}$$

where ΔC_i is the modulation amplitude of the concentration of every species i whose absorption and/or dispersion spectrum overlaps with the probe wavelength.

The temporal variation of ΔC_i can be due either to the dynamics of the species i in the illuminated grating fringes or to processes taking place between the fringes, such as translational diffusion, excitation transport, and charge diffusion. In liquids, the decay of the population grating by translational diffusion is slow and occurs in the microsecond to millisecond timescale, depending on the fringe spacing. As thermal diffusion is typically hundred times faster, Eq. (14) is again valid in this long timescale. Therefore if the population dynamics is very slow, the translational diffusion coefficient of a chemical species can be obtained by measuring the decay of the diffracted intensity as a function of the fringe spacing. This procedure has also been used to determine the temperature of flames. In this case however, the decay of the population grating by translational diffusion occurs typically in the sub-ns timescale.

In the condensed phase, these interfringe processes are of minor importance when measuring the diffracted intensity in the short timescale, i.e., before the formation of the density phase grating. In this case, the transient grating technique is similar to transient absorption, and it thus allows the measurement of population dynamics. However, because holographic detection is background free, it is at least a hundred times more sensitive than transient absorption.

The population gratings are usually probed with monochromatic laser pulses. This procedure is well suited for simple photoinduced processes, such as the decay of an excited state population to the ground state. For example, Fig. 8 shows the decay of the diffracted intensity at 532 nm measured after excitation of a cyanine dye at the same wavelength. These dynamics correspond to the ground state recovery of the dye by non-radiative deactivation of the first singlet excited state. Because the diffracted intensity is proportional to the square of the concentration changes (see Eq. (14)), its decay time is twice as short as the ground state recovery time. If the reaction is more complex, or if several intermediates are involved, the population grating has to be probed at several wavelengths. Instead of repeating many single wavelength experiments, it is preferable to perform multiplex transient grating. In this case, the grating is probed by white light pulses generated by focusing high intensity fs or ps laser pulses in a dispersive medium. If the crossing angle of the pump pulses is small enough (< 1°), the Bragg angle for probing is almost independent on the wavelength. A transient grating spectrum is obtained by dispersing the diffracted signal in a spectrograph. This spectrum consists in the sum of the square of the transient absorption and transient dispersion spectra. Practically, it is very similar to a transient absorption spectrum, but with a much superior signal to noise ratio. Fig. 9 shows the transient grating spectra measured at various time delays after excitation of a solution of chloranil (CA) and methylnaphthalene (MNA) in acetonitrile. The reaction

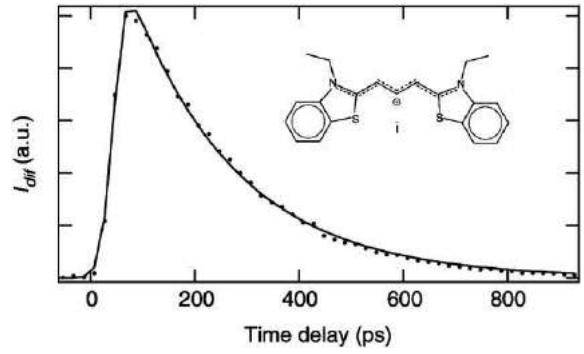

Fig. 8 Time profile of the diffracted intensity at 532 nm measured after excitation at the same wavelength of a cyanine dye (inset) in solution and best fit assuming exponential ground state recovery.

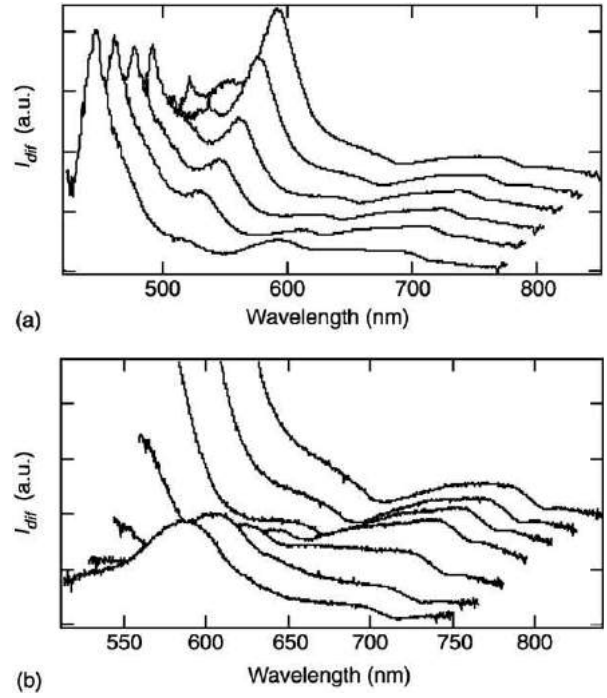

Fig. 9 Transient grating spectra obtained at various time delays after excitation at 355 nm of a solution of chloranil and 0.25 M methylnaphthalene (a): from top to bottom: 60, 100, 180, 260, 330 and 600 ps; (b): from top to bottom: 100, 400, 750, 1100, 1500 ps).

that takes place is

$$^3CA^* \cdot MNA \xrightarrow{\text{diffusion}} (^3CA^* \cdot MNA) \xrightarrow{ET} {}^3(CA^- \cdot MNA^+) \xrightarrow{MNA} {}^3(CA^- \cdot MNA^+)$$

After its formation upon ultrafast ISC, $^3CA^*$, which absorbs around 510 nm, decays upon ET with MNA to generate CA^-(450 nm) and MNA^+ (690 nm). The ion pair reacts further with a second MNA molecule to form the dimer cation (580 nm).

The time profile of the diffracted intensity reflects the population dynamics as long as these populations follow first- or pseudo-first order kinetics. Higher order kinetics leads to inharmonic gratings and in this case the time dependence of the diffracted intensity is no longer a direct measure of the population dynamics.

Polarization Selective Transient Grating

In the above applications, the selection between the different contributions to the diffracted signal was essentially made by choosing the probe wavelength and the crossing angle of the pump pulses. However, this approach is not always sufficient. Another important parameter is the polarization of the four waves involved in a transient grating experiment. For example, when measuring population dynamics, the polarization of the probe beam has to be at magic angle (54.7°) relatively to that of the

pump pulses. This ensures that the observed time profiles are not distorted by the decay of the orientational anisotropy of the species created by the polarized pump pulses.

The magnitude of ΔA and Δn depends on the pump pulses intensity, and therefore the diffracted intensity can be expressed by using the formalism of nonlinear optics:

$$I_{\mathrm{dif}}(t) = C \int_{-\infty}^{+\infty} \mathrm{d}t\, I_{\mathrm{pr}}(t - t'') \left[\int_{-\infty}^{t} \mathrm{d}t' |R_{\mathrm{ijkl}}(t'' - t')| I_{\mathrm{pu}}(t') \right]^2 \tag{15}$$

where C is a constant and R_{ijkl} is an element of the fourth rank tensor **R** describing the nonlinear response of the material to the applied optical fields. In isotropic media, this tensor has only 21 nonzero elements, which are related as follows:

$$R_{1111} = R_{1122} + R_{1212} + R_{1221} \tag{16}$$

where the subscripts are the Cartesian coordinates. Going from right to left, they design the direction of polarization of the pump, probe, and signal pulses. The remaining elements can be obtained by permutation of these indices ($R_{1122}=R_{1133}=R_{2211}=\ldots$). In a conventional transient grating experiment, the two pump pulses are at the same frequency and are time coincident and therefore their indices can be interchanged. In this case, $R_{1212}=R_{1221}$ and the number of independent tensor elements is further reduced:

$$R_{1111} = R_{1122} + 2R_{1212} \tag{17}$$

The tensor **R** can be decomposed into four tensors according to the origin of the sample response: population and density changes, electronic and nuclear optical Kerr effects:

$$\boldsymbol{R} = \boldsymbol{R}(p) + \boldsymbol{R}(d) + \boldsymbol{R}(K, e) + \boldsymbol{R}(K, n) \tag{18}$$

As they describe different phenomena, these tensors do not have the same symmetry properties. Therefore, the various contributions to the diffracted intensity can, in some cases, be measured selectively by choosing the appropriate polarization of the four waves.

Table 2 shows the relative amplitude of the most important elements of these four tensors.

The tensor $\mathbf{R}(p)$ depends on the polarization anisotropy of the sample, r, created by excitation with polarized pump pulses:

$$r(t) = \frac{N_{\parallel}(t) - N_{\perp}(t)}{N_{\parallel}(t) + 2N_{\perp}(t)} = \frac{2}{5} \mathrm{P}_2[\cos(\gamma)] \exp(-k_{\mathrm{re}} t) \tag{19}$$

where $N_{\parallel}$ and $N_{\perp}$ are the number of molecules with the transition dipole oriented parallel and perpendicular to the polarization of the probe pulse, respectively, γ is the angle between the transition dipoles involved in the pump and probe processes, P_2 is the second Legendre polynomial, and k_{re} is the rate constant for the reorientation of the transition dipole, for example, by rotational diffusion of the molecule or by energy hopping.

The table shows that $R_{1212}(d)=0$, i.e., the contribution of the density grating can be eliminated with the set of polarization (0°,90°,0°,90°), the so-called crossed grating geometry. In this geometry, the two pump pulses have orthogonal polarization, the polarization of the probe pulse is parallel to that of one pump pulse and the polarization component of the signal that is orthogonal to the polarization of the probe pulse is measured. In this geometry, $R_{1212}(p)$ is nonzero as long as there is some polarization anisotropy, ($r \neq 0$). In this case, the diffracted intensity is

$$I_{dif}(t) \propto |R_{1212}(p)|^2 \propto [\Delta C(t) \times r(t)]^2 \tag{20}$$

The crossed grating technique can thus be used to investigate the reorientational dynamics of molecules, through $r(t)$, especially when the dynamics of r is faster than that of ΔC. For example, **Fig. 10** shows the time profile of the diffracted intensity measured in the crossed grating geometry with rhodamine 6G in ethanol. The decay is due to the reorientation of the molecule by rotational diffusion, the excited lifetime of rhodamine being about 4 ns.

On the other hand, if the decay of r is slower than that of ΔC, the crossed grating geometry can be used to measure the population dynamics without any interference from the density phase grating. For example, **Fig. 11** shows the time profile of the diffracted intensity after excitation of a suspension of TiO_2 particles in water. The upper profile was measured with the set of

Table 2 Relative value of the most important elements of the response tensors $\mathbf{R}(i)$ and polarization angle ζ of the signal beam, where the contribution of the corresponding process vanishes for the set of polarization (ζ, 45 °,0°, 0°).

Process	R_{1111}	R_{1122}	R_{1212}	ζ
Electronic OKE	1	1/3	1/3	− 71.6°
Nuclear OKE	1	− 1/2	3/4	63.4°
Density	1	1	0	− 45°
Population:				
$\gamma=0°$	1	1/3	1/3	− 71.6°
$\gamma=90°$	1	2	− 1/2	− 26.6°
No correlation: $r=0$	1	1	0	− 45°

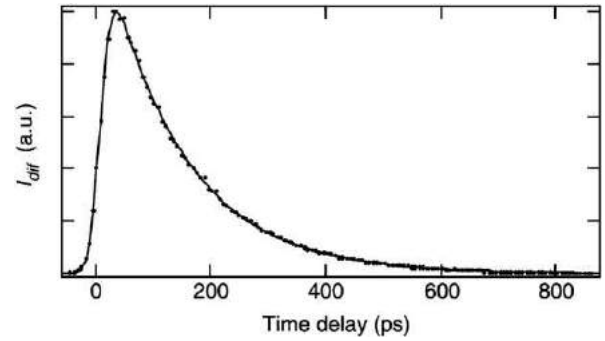

Fig. 10 Time profile of the diffracted intensity measured with a solution of rhodamine 6G with crossed grating geometry.

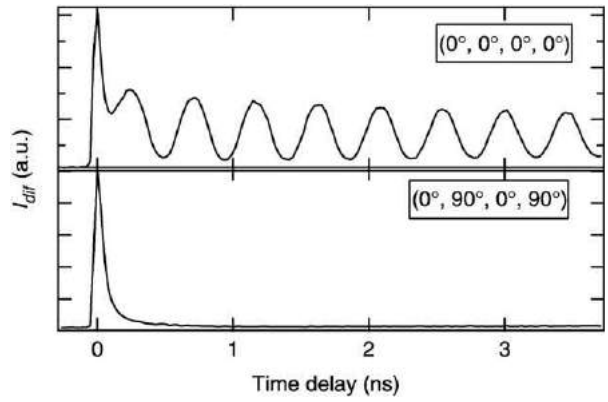

Fig. 11 Time profiles of the diffracted intensity after excitation at 355 nm of a suspension of TiO_2 particles in water and using different set of polarization of the four beams.

polarization $(0^\circ,0^\circ,0^\circ,0^\circ)$ and thus reflects the time dependence of $R_{1111}(p)$ and $R_{1111}(d)$. $R_{1111}(p)$ is due to the trapped electron population, which decays by charge recombination, and $R_{1111}(d)$ is due to the heat dissipated upon both charge separation and recombination. The lower time profile was measured in the crossed grating geometry and thus reflects the time dependence of R_{1212} and in particular that of $R_{1212}(p)$.

Each contribution to the signal can be selectively eliminated by using the set of geometry $(\zeta,45^\circ,0^\circ,0^\circ)$ where the value of the 'magic angle' ζ for each contribution is listed in **Table 2**. This approach allows for example the nuclear and electronic contributions to the optical Kerr effect to be measured separately.

Concluding Remarks

The transient grating techniques offer a large variety of applications for investigating the dynamics of chemical processes. We have only discussed the cases where the pump pulses are time coincident and at the same wavelength. Excitation with pump pulses at different wavelengths results to a moving grating. The well-known CARS spectroscopy is such a moving grating technique. Finally, the three pulse photon echo can be considered as a special case of transient grating where the sample is excited by two pump pulses, which are at the same wavelength but are not time coincident.

See also: Atomic Physics

Further Reading

Bräuchle, C., Burland, D.M., 1983. Holographic methods for the investigation of photochemical and photophysical properties of molecules. Angewandte Chemie International Edition Engl 22, 582–598.
Eichler, H.J., Günter, P., Pohl, D.W., 1986. Laser-Induced Dynamics Gratings. Berlin: Springer.
Fleming, G.R., 1986. Chemical Applications of Ultrafast Spectroscopy. Oxford: Oxford University Press.
Fourkas, J.T., Fayer, M.D., 1992. The transient grating: a holographic window to dynamic processes. Accounts of Chemical Research 25, 227–233.

Hall, G., Whitaker, B.J., 1994. Laser-induced grating spectroscopy. Journal of the Chemistry Society Faraday Trans 90, 1–16.
Levenson, M.D., Kano, S.S., 1987. Introduction to Nonlinear Laser Spectroscopy, revised edn. Boston: Academic Press.
Mukamel, S., 1995. Nonlinear Optical Spectroscopy. Oxford: Oxford University Press.
Rullière, C. (Ed.), 1998. Femtosecond Laser Pulses. Berlin: Springer.
Terazima, M., 1998. Photothermal studies of photophysical and photochemical processes by the transient grating method. Advances in Photochemistry 24, 255–338.

Atomic Physics

G Kurizki and AG Kofman, Weizmann Institute of Science, Rehovot, Israel
D Petrosyan, Institute of Electronic Structure and Laser, Heraklion, Greece

Nomenclature

$|\ldots\rangle$ Atomic or photonic eigenstates [dimensionless]
p Complex polarizability [cm^{-1}]
κ, $C^{2/3}$, B Coupling constant [s^{-1}]
γ Decay rate [s^{-1}]
ρ Density of modes [s]
ω, Δ, s Frequency [s^{-1}]
ϕ Phase [dimensionless]
α, β, c Probability amplitudes [dimensionless]
G Reservoir spectral response [s^{-1}]
a_0 Resonant absorption coefficient [cm^{-1}]
W Spontaneous emission spectrum [s]
k Wave vector [m^{-1}]

Introduction

The fabrication and study of dielectric structures whose refractive index is periodically modulated on a micron or submicron scale, known as photonic crystals (PCs), are attracting considerable interest at present. One-, two-, or three-dimensional (1D-, 2D-, or 3D-) periodic PCs exhibit photonic band gaps (PBGs), analogously to electronic band gaps in ordinary crystals, and can incorporate defects, designed to form localized narrow-linewidth (high-Q) modes at PBG frequencies. The advent of such structures opens up new perspectives in atomic physics and quantum optics, since they are expected to allow an unprecedented control over the spectral density of modes (DOM) and the spatial modulation of narrow-linewidth (high-Q) modes, in the microwave, infrared, and optical domains. An interesting situation arises when foreign atoms or ions – dopants – with transition frequencies within the PBG, are implanted in the PC. Then light near one of these frequencies resonantly interacts with the dopants and is concurrently affected by the PBG dispersion. Consequently, highly nonlinear processes with a rich variety of unusual PC-related features are anticipated. Such nonlinear optical processes, involving near-resonant transitions in PCs, undergo basic modifications as compared to the corresponding processes in free space, which are attributed to the strong suppression of the DOM within PBGs, to sharp bandedges and to intra-gap narrow lines associated with high-Q defect modes.

Results are detailed below for spontaneous emission of radiation in PCs and for the resonant interaction of atoms with the field of a single high-Q defect mode. These results stem from the failure of perturbation theory and the onset of strong field-atom coupling near sharp bandedges or narrow defect-mode lines in 2D and 3D PCs. 3D-PBGs are needed, in order to extinguish spontaneous emission in all possible directions of propagation and dipole orientations. If only one polarized atomic transition is involved in the spontaneous emission, 2D-PCs suffice for its suppression. For controlling strictly unidirectional field propagation, it is sufficient to resort to PBGs in 1D-periodic structures (Bragg reflectors or dielectric multi-layer mirrors).

As a further example of field–atom interactions in doped PCs, we will show that two ultraweak electromagnetic fields (or photons), coupled to appropriate transitions of the dopants in a PC, can mutually induce large phase shifts or drastic changes in absorption. Appreciable photon–photon correlations can then be established, which can be utilized in both classical and quantum optical communications.

Radiative Decay and Photon–Atom Binding in a PC

The interaction of a two-level atom with an arbitrary field-mode continuum can be described by the following second-quantized Hamiltonian in the rotating-wave approximation (RWA)

$$H = \hbar\omega_a|e\rangle\langle e| + \hbar\sum_{\vec{k}}\omega_{\vec{k}}a^\dagger_{\vec{k}}a_{\vec{k}} + \hbar\sum_{\vec{k}}\left[\kappa(\omega_{\vec{k}})a^\dagger_{\vec{k}}|g\rangle\langle e| + \text{hermitian constant}\right] \quad (1)$$

Here $|e\rangle$ and $|g\rangle$ are the excited and ground atomic states, respectively, ω_a is the atomic transition frequency, $a^\dagger_{\vec{k}}$ and $a_{\vec{k}}$ are the creation and annihilation operators of a field mode with wavevector $\vec{k}$, and frequency $\omega_{\vec{k}}$, and $\hbar\kappa(\omega_{\vec{k}})$ is the resonant coupling energy of this mode with the atomic dipole. The $\vec{k}$-summation can be replaced by an integral over the continuum DOM in the frequency domain, $\sum_{\vec{k}} \to \int_0^\infty d\omega\rho(\omega)$, avoiding, for the sake of simplicity, the study of direction-(angle-)dependent DOM effects.

We consider a two-level atom that is excited at time $t=0$ in an empty, periodic dielectric structure. The wavefunction of the combined system field + atom can be cast in the general form:

$$|\Psi(t)\rangle = \alpha(t)|e,\{0_\omega\}\rangle + \int \beta_\omega(t)|g,1_\omega\rangle\rho(\omega)d\omega \quad (2)$$

 doi:10.1016/B978-0-12-809283-5.00935-6

where $\{0_\omega\}$ signifies the completeness of field modes in the vacuum state and $|1\omega\rangle$ denotes single-photon occupation of the ω-mode. The corresponding Schrödinger equation can be solved with the initial condition $|\Psi(0)\rangle = |e, \{0_\omega\}\rangle$ by means of the continuum spectral response:

$$G(\omega) = |\kappa(\omega)|^2 \rho(\omega) \tag{3}$$

The analysis of **Eq. (2)**, with $G(\omega)$ appropriate for photonic bandstructures, is aimed at revealing the prominent features of the atomic excitation decay $\alpha(t)$ and the corresponding emission spectrum.

In what follows, we consider a photonic bandstructure with several PBGs, separated by allowed bands. Each PBG is labeled by index i and has lower and upper cutoff frequencies ω_{Li} and ω_{Ui}, respectively. Then $G(\omega) = 0$ for $\omega_{Li} < \omega < \omega_{Ui}$. Naively, one might expect that an excited atom would either be stable against single-photon decay, if ω_a is within a PBG, or decay completely at $t \to \infty$, if ω_a is anywhere in an allowed band. However, both statements turn out to be inaccurate.

Incomplete decay of $\alpha(t \to \infty)$ occurs if there is a stable eigenvalue (energy level) $\hbar\omega_i$ of the total (field-atom) Hamiltonian **Eq. (1)**. Such an energy level is possible only if $G(\omega_i) = 0$, i.e., for ω_i in a PBG. It must satisfy:

$$\omega_i = \omega_a + \Delta(\omega_i) \tag{4}$$

$$\Delta(\omega_i) = \int_0^{\omega_{Li}} \frac{G(\omega')}{\omega_i - \omega'} d\omega' + \int_{\omega_{Ui}}^{\infty} \frac{G(\omega')}{\omega_i - \omega'} d\omega' \tag{5}$$

Here the integrals are the frequency shifts of the atomic resonance $\hbar\omega_a$ induced by the spectral parts of the reservoir situated, respectively, below and above the ith PBG (**Fig. 1**). If **Eq. (4)** holds, then the corresponding term in $\alpha(t)$ is proportional to $\exp(-i\omega_i t)$, and does not decay. Physically, such a stable level can be interpreted as representing the binding of the photon to the atom (photon-dressed atom), without the ability to leave the atomic vicinity, due to Bragg reflection in the PBG.

Assuming that ω_i occurs in the ith PBG, we observe that the first shift is positive whereas the second shift is negative, i.e., each part of the continuum repels the atomic level from its PBG edge. **Eq. (4)** has a solution if ω_a falls between the minimum and maximum values assumed by its left-hand side in the ith PBG, i.e., if

$$\omega_{Li} - \Delta(\omega_{Li}) < \omega_a < \omega_{Ui} - \Delta(\omega_{Ui}) \tag{6}$$

Incomplete decay occurs if this condition holds at least for one PBG. There may be, at most, one stable state in a single PBG (**Fig. 1**). However, a two-level atom in a periodic structure can have several stable dressed states simultaneously, when conditions hold for several PBGs. It is quite extraordinary that the conditions (**Eq. (6)**) for incomplete decay can be fulfilled with ω_a, not only in a PBG, but even in an allowed band, e.g., when $\Delta(\omega_{Ui})$ is negative or when $\Delta(\omega_{Li})$ is positive (**Fig. 1(a)** and **(b)**).

Equally counterintuitive is the converse possibility, of complete decay for ω_a within a PBG. When there is a single PBG and one of the inequalities in **Eq. (6)** is violated, complete decay will occur for $\omega_{Li} < \omega_a < \omega_{Li} - \Delta(\omega_{Li})$, if $\Delta(\omega_{Li}) < 0$, or $\omega_{Ui} - \Delta(\omega_{Ui}) < \omega_a < \omega_{Ui}$, if $\Delta(\omega_{Ui}) > 0$.

The possibility that one or several discrete, stable, states exist yields the corresponding wavefunction in the form:

$$|\Psi(t)\rangle = \sum_i \sqrt{c_i} |\psi_i\rangle e^{-i\omega_i t} + |\Psi_c(t)\rangle \tag{7}$$

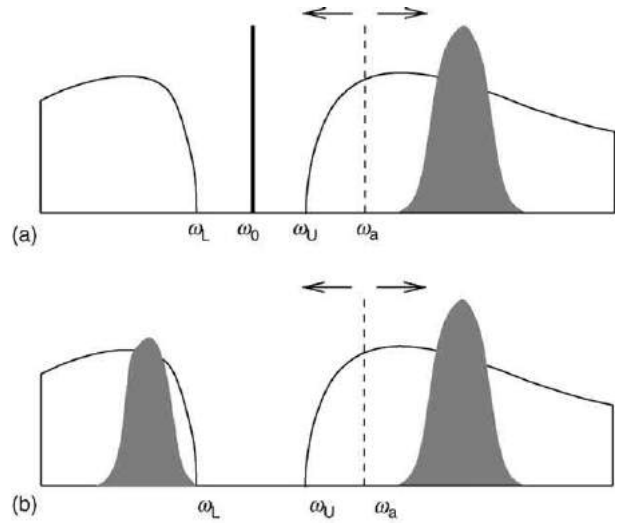

Fig. 1 Frequency shifts of the atomic resonance ω_a (dashed line) due to 'repulsion' (arrows) by asymmetric spectral parts of the continuum below and above the PBG: (a) Incomplete decay for ω_a in allowed band. The interaction with the continuum splits the state into a superposition of a stable part in the PBG (solid line) and a decaying part above ω_a (shaded peak). (b) Complete decay for same ω_a, due to larger shifts that split the state into decaying parts in two allowed bands.

Here the summation is over all discrete atomic photon-dressed states with energies $\hbar\omega_i$, $|\psi_i\rangle$ is a discrete-(dressed-)state eigenfunction of the Hamiltonian (Eq. (1)) normalized to unity, and weighted by the amplitude $\sqrt{c_i} = \left[1 + \int_0^\infty d\omega G(\omega)/(\omega - \omega_i)^2\right]^{-1/2}$. The dressed state $|\psi_i\rangle$ consists of an excited-state component and a photon-bound ground-state component.

The population of the excited state for long times:

$$|\alpha(t)|^2 = \sum_{i,i'} c_i c_{i'} \cos(\omega_i - \omega_{i'})t \quad \text{for} \quad t \to \infty \tag{8}$$

is a nonzero time-constant if there is one discrete stable state, whereas in the case of several such states, it undergoes beats at the frequencies corresponding to their energy differences. The splitting of an excited state $|e\rangle$ into superposed stable states, oscillating at bandgap frequencies ω_i, whose amplitudes c_i, and eigenfrequencies ω_i are controllable by the atomic transition detuning from cutoff, constitutes spontaneous coherence control.

If, however, ω_a is far from the cutoff, then Eq. (2) results in an exponentially decaying amplitude:

$$\alpha(t) \approx e^{-(\gamma_a + i\tilde{\omega}_a)t} \tag{9}$$

where $\tilde{\omega}_a = \omega_a + \Delta_a$, Δ_a, and γ_a being the effective spectral shift and width of the decaying atom. This regime holds for a locally smooth $G(\omega \approx \omega_a)$ such that

$$|\gamma_a'|, |\Delta_a'| \ll 1; |\gamma_a' \Delta_a| \ll \gamma_a; \gamma_a, |\Delta_a| \ll |\omega_a - \omega_g| \tag{10}$$

where ω_g is the bandedge frequency nearest to ω_a and the primes denote differentiation with respect to ω_a

We now apply the foregoing general results to a model DOM distribution. This distribution is derived on keeping the lowest term in the Taylor expansion of the dispersion relation $\omega(\vec{k})$ near a photonic bandedge ω_U (the effects of the further PBG edge are neglected). This yields the effective mass approximation:

$$\omega \approx \omega_U + \sum_{i=x,y,z} (k - k_U)_i^2 / m_i \tag{11}$$

with $1/m_i = (1/2)(\partial^2\omega/\partial k_i^2)|_{\omega=\omega_U}$. In a structure with period L, k_U satisfies the Bragg condition $k_U = \pi/L$. The corresponding DOM in a 3D-periodic structure with an allowed symmetry may be approximated as

$$\rho(\omega) \sim (\omega - \omega_U)^{(1-D)/2} \theta(\omega - \omega_U) \tag{12}$$

where θ is the step function and D is the dimension of the Brillouin-zone surface spanned by bandedge modes with vanishing group velocity. In realistic photonic structures $D \leq 2$, $D=2$ corresponding to completely isotropic dispersion (spherical Brillouin-zone surface) and $D=0$ corresponding to an anisotropic three-dimensional Brillouin zone. Both cases can be represented, respectively, as the limits $\varepsilon \to 0$ and $\varepsilon \to \infty$ of the function:

$$G(\omega) = \frac{C}{\pi} \frac{\sqrt{\omega - \omega_U}}{\omega - \omega_U + \varepsilon} \theta(\omega - \omega_U) \tag{13}$$

where ε is the 'cutoff-smoothing' parameter and C is the continuum coupling constant. Depending on whether $C^{2/3}/\varepsilon$ is greater or less than one, the continuum is close to the case $\varepsilon = 0$ ($D=2$) or $\varepsilon \to \infty$ ($D=0$), respectively.

Using the properties of the model DOM (Eq. (13)), we can infer the criteria for the two regimes discussed above:

(i) The conditions for incomplete decay, Eq. (6), are now (Fig. 2) $\Delta_c \equiv \omega_a - \omega_U < C\varepsilon^{-1/2}$. Hence, the abrupt, singular cutoff of the DOM with $D=2(\varepsilon \to 0)$ implies the existence of a discrete state for any ω_a, either inside or outside the PBG. The energy of the discrete state, $\hbar\omega_0$ which must lie in the PBG, is found to be a real and positive root of the equation $\omega_a - \omega_0 = C/(\sqrt{\omega_U - \omega_0} + \sqrt{\varepsilon})$.

(ii) The conditions for nearly exponential decay, Eq. (10), can be shown to reduce now to the requirement that ω_a be in the allowed zone sufficiently far from cutoff, $\Delta_c \gg \min\{C^{2/3}, C\varepsilon^{1/2}\}$. Under this condition, the intermediate-time exponential decay of the excited state (Fig. 3) is given to first approximation by Eq. (9), with $\gamma_a = C\sqrt{\Delta_c}/(\varepsilon + \Delta_c)$ and $\Delta_a = -C\sqrt{\varepsilon}/(\varepsilon + \Delta_c)$.

The resulting atomic frequency shift Δ_a is negative in the present model, vanishing for $\varepsilon \to 0$ ($D=2$). The long-time behavior of $\alpha(t)$ can be shown to exhibit a tail decaying as $t^{-3/2}$ (or $t^{-1/2}$ at $\Delta_c = C\varepsilon^{-1/2}$) and oscillating at the cutoff frequency ω_U (Fig. 3).

With the increase of ε, the smoothing inhibits the decay more strongly for ω_a at cutoff, $\Delta_c = 0$, because $G(\omega_a \simeq \omega_U)$ is now weaker and the stable-state probability c_0 (Eq. (7)) is correspondingly larger. By contrast, at a large detuning from cutoff Δ_c the large-ε smoothing allows the decay to become complete, and the corresponding spectrum is entirely Lorentzian.

A defect in the periodic structure can produce a narrow-linewidth local mode in the PBG, whose spectral response is describable by a Lorentzian:

$$G_d(\omega) = \frac{\gamma_d}{\pi} \frac{\Gamma_d^2}{\Gamma_d^2 + (\omega - \omega_d)^2} \tag{14}$$

where γ_d characterizes the coupling strength of the atomic dipole with the defect field, whereas ω_d and Γ_d stand for the line center and width, respectively. The presence of a nonvanishing DOM in the PBG, due to a defect, causes spontaneous emission in the

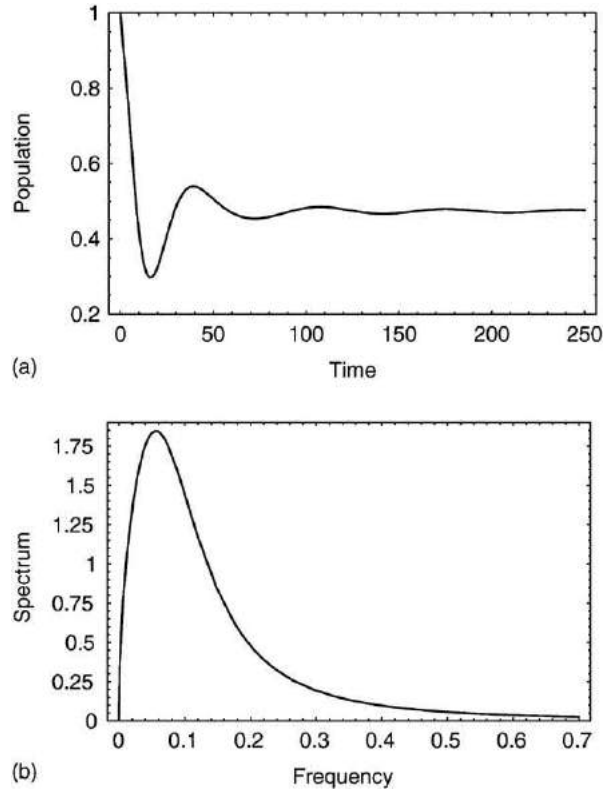

Fig. 2 (a) Incomplete decay of the excited state population as a function of $\gamma_c t$, where $\gamma_c = C\varepsilon^{-1/2}$, for $\varepsilon = 10^{-3}$, at cutoff $\Delta_c = 0$. The beat period is $2\pi/(\omega_U - \omega_0)$ and the nondecaying probability is $c_0^2 = 4/9$. (b) The corresponding spectrum. The frequency is normalized to γ_c.

PBG spectral range. This broadens the discrete state ω_0, which becomes metastable (see **Fig. 4**). The oscillator strength of the line at ω_0 is then proportional to c_0.

EIT and Cross-Coupling of Photons in Doped PCs

Nonlinear effects, whereby one light beam influences another, require strong fields or else light confinement in a high-Q cavity. The analysis above has shown that by choosing an appropriate detuning of an atomic transition from the PBG cutoff, the spontaneous coherence is established between the states of an initially excited atom. The ability to control this coherence, by varying the detuning, opens interesting perspectives for optical processes in PCs containing multilevel atoms with a resonant transition near a bandedge. From among such processes, we discuss here electromagnetically induced transparency (EIT) and its applicability to nonlinear photon switching or giant cross-phase modulation.

Let us examine the nonlinear coupling of two weak (single-photon) optical fields E_a and E_b with the frequencies ω_a and ω_b, respectively, propagating along the z-axis in a PC dilutely doped with identical four-level atoms (**Fig. 5(a)**). These fields interact with the atoms via the transitions $|1\rangle \rightarrow |2\rangle$ and $|3\rangle \rightarrow |4\rangle$, respectively, while the transition $|2\rangle \rightarrow |3\rangle$ is coupled to the structured mode-continuum $\rho(\omega)$ in the PC (**Fig. 5(b)**). Initially the atoms are in the ground state $|1\rangle$ and the continuum is in the vacuum state $\{0_\omega\}$. Then the wavefunction of the system at the position z_l of the lth atom reads:

$$|\Psi(z_l, t)\rangle = \alpha_1 |1, \{0_\omega\}\rangle + \alpha_2 |2, \{0_\omega\}\rangle + \int \beta_{3,\omega} |3, 1_\omega\rangle \rho(\omega) \mathrm{d}\omega + \int \beta_{4,\omega} |4, 1_\omega\rangle \rho(\omega) \mathrm{d}\omega \tag{15}$$

The consecutive terms in **Eq. (15)** denote the atom being in states $|1\rangle$, $|2\rangle$, $|3\rangle$, or $|4\rangle$, with zero $\{0_\omega\}$ or one $|1_\omega\rangle$ photon in mode ω whose DOM is $\rho(\omega)$, and α_1, α_2, $\beta_{3,\omega}$, or $\beta_{4,\omega}$ are the corresponding probability amplitudes. Upon making the weak-field linear-response approximation, we can set $\alpha_1 \simeq 1$ and solve the set of equations for the slowly varying (compared to an optical cycle) probability amplitudes α and β, using the perturbation theory. Under these conditions, we effectively obtain a free-space propagation of the E_b field. By contrast, the evolution of the E_a field in the slowly varying envelope approximation, is given by $E_a(z, t) = E_a(0, t - z/v_g)\exp(ipz)$. We thus see that the real part of the macroscopic complex polarizability p is responsible for the phase shift ϕ_a of the E_a field, $\phi_a = \mathrm{Re}(p)z$, while the probability of the absorption $\mathscr{A}$ of the field depends on the imaginary part of p, $\mathscr{A} = 1 - \exp[-2\mathrm{Im}(p)z]$. The polarizability is expressed by

$$p = a_0 \frac{i\gamma_2/2}{\gamma_2/2 - i\Delta_a + I(\Delta_a)} \tag{16}$$

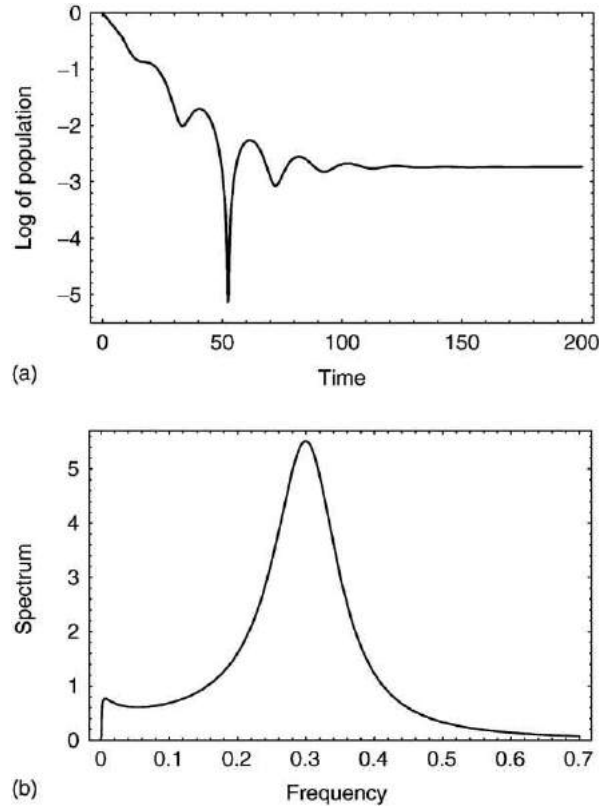

Fig. 3 As in **Fig. 2**, for a large detuning from the sharp cutoff, $\varepsilon = 10^{-3}$, $\Delta_c = 0.3$. (a) $\log_{10}|\alpha|t^2$ is plotted as a function of t. The nearly complete decay ($c_0^2 \sim 10^{-3}$) is modulated by beats with frequency Δ_c. A power tail obtains for $t \gg \gamma_a^{-1} \approx \Delta_c^{1/2}/C$, decreasing as $t^{-1/2}$ for $t \ll \Delta_c^2/C^2$ and as $t^{-3/2}$ for $t \gg \Delta_c^2/C^2$. (b) The nearly Lorentzian spectrum, $w(\omega)$, has a small peak near ω_U, due to the sharply peaked DOM near the cutoff.

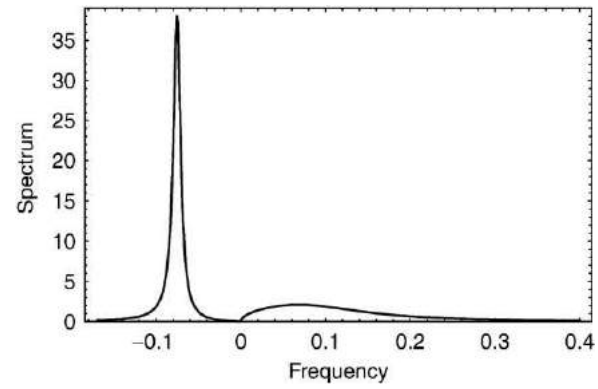

Fig. 4 Spectrum $w(\omega)$(same units as in **Figs. 2** and **3**) for ω_a at cutoff, $\Delta_c = 0$, in the presence of a defect in the PBG, $\omega_d = -10$, $\gamma = 0.1$, $\Gamma_d = 3$, $\varepsilon = 10^{-3}$. The oscillator strength is now roughly equally shared between the distribution above cutoff (same as in **Fig. 2(b)**) and the defect peak at ω_0.

where $a_0 \equiv \sigma_0 N$ is the linear resonant absorption coefficient on the atomic transition $|1\rangle \rightarrow |2\rangle$, with σ_0 the resonant absorption cross-section and N the density of doping atoms, γ_2 is the radiative width of state $|2\rangle$, $\Delta_a = \omega_a - \omega_{21}$ is the detuning from the atomic resonance ω_{21}, and $I(\Delta_a)$ is the integral of the saturation factor over the structured DOM. The group velocity v_g is expressed as $v_g = \partial_{k_a}\omega_a = [n_a/c + \partial_{\omega_a}\mathrm{Re}(p)]^{-1}$, where n_a is the (averaged) refraction index at the frequency ω_a.

To calculate $I(\Delta_a)$, we assume the isotropic PBG model, **Eq. (13)** with $\varepsilon \rightarrow 0$, with the atoms doped at the positions of the local defects in the PC separated by a distance d from each other. These defect modes in the PBG are localized around each atomic site in a volume $V_d \simeq (rL)^3$ of several $(r)^3$ lattice cells L^3, with $L \simeq \pi c/\omega_{23}$, and serve as effective high-Q cavities, **Eq. (14)** with $\Gamma_d \ll 1$. Assuming $|\Delta_b| \gg |\Delta|, |\Delta_a|, \gamma_4$, where $\Delta_b = \omega_b - \omega_{43}$ and $\Delta = \omega - \omega_{23}$, the integration leads to

$$I(\Delta_a) = \frac{B_d^2}{\gamma_{31} - i(\Delta_a - \Delta_d - s_3)} - \frac{B_U^{3/2}}{\sqrt{i\gamma_{31} + (\Delta_a - \Delta_U - s_3)}} \tag{17}$$

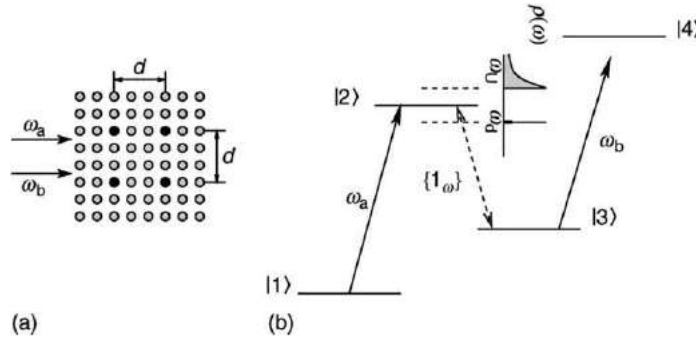

Fig. 5 (a) Photonic crystal dilutely doped with atoms located at black dots. (b) Four-level atom coupled to a structured continuum $\rho(\omega)$ near the band-edge or defect mode frequencies (DOM plotted) via the intermediate transition $|2\rangle \rightarrow |3\rangle$ and interacting with two weak fields E_a and E_b at the sideband transitions $|1\rangle \rightarrow |2\rangle$ and $|3\rangle \rightarrow |4\rangle$, respectively.

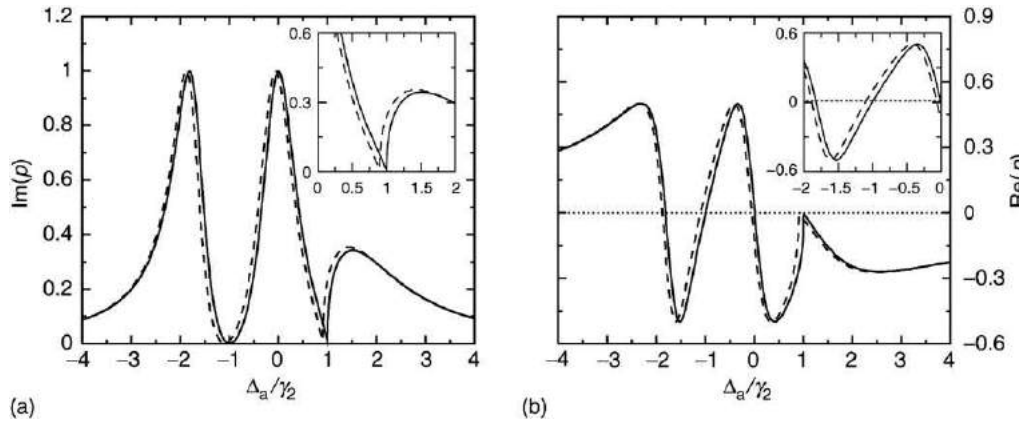

Fig. 6 (a) Imaginary and (b) real part of the complex polarizability p as a function of the detuning Δ_a for the case $E_b=0$ (solid lines) and $E_b \neq 0$ (dashed lines). The parameters (normalized by γ_2) are: $\Delta_d=-1$, $\Delta_U=1$, $\gamma_{31}=0.001$, $B_d=B_U=1$, $s_3=-0.1$, and $a_0=1$ cm^{-1}. The insets magnify the important frequency regions.

where $\Delta_{d,\,U}=\omega_{d,\,U}-\omega_{23} \ll \omega_{23}$ are the detunings of the defect-mode and PBG-edge frequencies from the atomic resonance ω_{23}, γ_{31} the $|1\rangle \leftrightarrow |3\rangle$ decoherence rate, $s_3=(\mu_{34}^2/\hbar^2\Delta_b)|E_b|^2$ is the E_b field-induced ac Stark shift of level $|3\rangle$ (μ_{ij} is the atomic dipole matrix element on the transition $|i\rangle \rightarrow |j\rangle$), and B_d and B_U are the coupling constants of the atom with the structured reservoir, whose main contributions are near ω_d and ω_U.

To illustrate the results of the foregoing analysis, we plot in **Fig. 6** the imaginary and real parts of the polarizability (**Eq. (16)**). Consider first the case of one incident field $E_a(E_b=0)$. Clearly, two frequency regions, $\Delta_a \sim \Delta_d$ and $\Delta_a \sim \Delta_U$, where the absorption vanishes and, at the same time, the dispersion slope is steep, are of particular interest. One can see that there is, however, a substantial difference between the spectra in the foregoing frequency regions, for the following physical reasons. First, in the vicinity of Δ_d, the atom interacts with the defect mode as in a high-Q cavity. This strong interaction 'dresses' the atomic states $|2\rangle$ and $|3\rangle$, thereby splitting the spectrum around $\Delta_a \simeq \Delta_d$ by the amount equal roughly to $2B_d$. Near the two-photon Raman resonance $\Delta_a=\Delta_d$, the two alternative transition paths $|1\rangle \rightarrow |2\rangle$ (the direct transition) and $|1\rangle \rightarrow |2\rangle \rightarrow |3\rangle \rightarrow |2\rangle$ (transition via state $|3\rangle$) interfere distractively with each other, cancelling thus the absorption of the E_a field and the medium becomes transparent to the radiation. This effect has been widely studied in atomic vapors, where the transition $|2\rangle \rightarrow |3\rangle$ is strongly driven by a coherent laser field, and is called electromagnetically induced transparency (EIT). The transparency window is rather broad and is given by the inverse Lorentzian (see the first term on the right-hand side of **Eq. (17)**). Due to the steepness of the dispersion curve, the corresponding group velocity is much smaller than the speed of light, $v_g \simeq 2B_d^2/(\gamma_2 a_0) \ll c$, which leads to a large delay time $T_{del}=\zeta/v_g$ at the exit $z=\zeta$ from the medium. One has to keep in mind, however, that the absorption-free propagation time is limited by the EIT decoherence time $T_{del} < \gamma_{31}^{-1}$, which imposes a limitation on the length ζ of the active PC medium. Second, in the vicinity of Δ_U, the strong interaction of the atom with the continuum near the bandedge ω_U causes the Autler–Townes splitting of level $|2\rangle$ into a doublet with a separation equal roughly to B_U. One component of this doublet is shifted out of the PBG, while the other one remains within the gap and forms the photon–atom bound state. Consequently, there is vanishing absorption and rapid variation of the dispersion at $\Delta_a \simeq \Delta_U$. Since the transparency region is very narrow with a width $\delta\omega \sim \gamma_{31} (\ll \gamma_2, B_U)$, for an absorption-free propagation of the E_a pulse, its temporal width τ_a should satisfy the condition $\tau_a > \pi/\delta\omega$. Simultaneously, a small deviation from the condition $\Delta_a=\Delta_U$ will lead to a strong increase in the absorption of the E_a field.

Let us now switch on the E_b field. As seen from Eq. (17), its effect is merely to shift the spectrum by the amount equal to s_3 (Fig. 6). This shift, however, will have different implications in the two frequency regions distinguished above: if $s_3 \ll B_d$, i.e., the Stark shift is smaller than the width of the EIT window at Δ_d, the medium will still remain transparent for an E_a field with the detuning $\Delta_a = \Delta_d$, but its phase will experience an appreciable nonlinear shift ϕ_a, given by $\phi_a = Re(p)z \simeq -\gamma_2 a_0 s_3 z/(2B_d^2)$. On the other hand, for an E_a field with the detuning $\Delta_a = \Delta_U$, the medium, which is transparent for $E_b = 0$, $\mathrm{Im}(p) \ll a_0$, will become highly absorptive (opaque) even for such a small frequency shift as s_3 (provided $s_3 < 0$ and $|s_3| > \Delta\omega$), $\mathrm{Im}(p) \simeq \gamma_2 a_0 \sqrt{|s_3|}\Big/\left(2B_U^{3/2}\right)$ and thus acting as an ultrasensitive, effective switch.

The remaining question is how to maximize the interaction between the E_a pulse, which propagates with a small group velocity, and the E_b pulse, which propagates with a velocity close to the speed of light. The interaction between the fields is maximized if: they enter the medium simultaneously; the transverse shapes of their envelopes overlap completely; and the pulse length l_b of the E_b field satisfies the condition $(l_b + \zeta)/c \leq \zeta/v_g$. Then the E_b pulse leaves the medium not later than the E_a pulse. The effective interaction length between the two pulses is, therefore, $z_{\mathrm{eff}} \sim l_b\, v_g/c \leq \zeta$, after which the two pulses slip apart. Thus, the presence of the E_b (control) field induces either strong absorption or a large phase shift of the E_a (signal) field, depending on the frequency region employed.

The effects surveyed above reveal unusual features of spontaneous emission and photon–atom binding in PCs, atomic interaction with the field of a high-Q defect mode, and nonlinear coupling of two fields via four-level dopant atoms in PCs. These effects are of fundamental interest. In addition, they can serve as the basis for highly efficient optical communications and data processing, in either the classical or the quantum domain, by providing two key elements: ultrasensitive nonlinear phase-shifters and photon switches.

Further Reading

Joannopoulos, J.D., Meade, R.D., Winn, J.N., 1995. Photonic Crystals: Molding the Flow of Light. Princeton: Princeton University Press.
John, S., Wang, J., 1991. Quantum optics of localized light in a photonic band gap. Physics Review B 43, 12772–12789.
Kofman, A.G., Kurizki, G., Sherman, B., 1994. Spontaneous and induced atomic decay in photonic band structures. Journal of Modern Optics 41, 353–384.
Lambropoulos, P., Nikolopoulos, G.M., Nielsen, T.R., Bay, S., 2000. Fundamental quantum optics in structured reservoirs. Reports on Progress in Physics 63, 455–503.
Loudon, R., 1983. The Quantum Theory of Light. Oxford, UK: Clarendon Press.
Meystre, P., Sargent III, M., 1991. Elements of Quantum Optics. Berlin: Springer Verlag.
Petrosyan, D., Kurizki, G., 2001. Photon–photon correlations and entanglement in doped photonic crystals. Physics Review A 64.023810-1-6.
Sakoda, K., 2001. Optical Properties of Photonic Crystals., Springer Series in Optical Sciences, 80., Berlin: Springer Verlag.
Scully, M.O., Zubairy, M.S., 1997. Quantum Optics. Cambridge, UK: Cambridge University Press.

Terahertz Physics of Semiconductor Heterostructures

Juraj Darmo and Karl Unterrainer, Vienna University of Technology, Vienna, Austria

Introduction

Semiconductor nanostructures are very attractive quantum systems which allow engineering of wavefunctions and transition energies. By changing size and geometry the optical and electronic properties of nanostructures can be designed in a wide range. This makes nanostructures very interesting for possible applications and for fundamental questions about light matter interactions. The quantum confinement leads to the occurrence of quantized states. The optical matrix element for transitions between quantized states can be orders of magnitude larger than that of atoms and the selection rules can be adjusted to meet the requirements of specific experiments. Thus, semiconductor nanostructures are in the center of solid state physics and device related research. Regarding application, optoelectronic devices have been targeted due to the possibility to design the wavelength especially in the infrared and THz spectral region. The THz spectral region is scientifically very interesting since many fundamental resonances (co)exist there: Collective excitations of electrons, phonons, impurity transitions and quantized transitions. Consequently, THz spectroscopy was developed to study these elementary processes. This had been quite difficult due to the lack of sources, but the realization of femtosecond laser generated THz pulses provided an ideal tool to study quantized transitions. Modern THz time-domain spectroscopy (TDS) allows performing time-resolved and non-linear spectroscopy which was previously impossible or restricted to few large scale facilities. Semiconductor nanostructures allow the realization of quantum optical systems, for example, by coupling several different quantum wells which can then be studied by TDS. Important questions are connected to the lifetimes of the electrons of excited states, to dephasing times and to collective effects. In addition, the possibility to externally control the quantized electrons by an electrical field is very attractive for experiments as well as for applications.

The development of THz quantum cascade lasers (Köhler *et al.*, 2002) has shown in a very impressive way that employing quantum structure can lead to conceptual new optoelectronic devices. The targeted spectral range spans from the THz range, through the infrared up to the telecommunication band. In the THz and infrared the quest is for sources and amplifiers while fast modulators are needed for telecommunications. For all these proposals the population dynamics and the dephasing times are the most crucial parameters. Ultrafast spectroscopy has enabled to study this times in a very direct way. For the measurement of carrier dynamics in nanostructures, pulsed excitation is combined with ultrashort THz probe pulses to perform time-resolved intersubband absorption spectroscopy. With the rapid developments in THz-technology a new class of non-equilibrium phenomena in nanostructures can be studied. With the obtained knowledge the realization and further development of monolithic semiconductor THz devices has become possible.

In Section Few Cycle THz Spectroscopy of Semiconductor Quantum Structures of this article, we discuss experiments with semiconductor quantum wells. THz Time-domain spectroscopy is employed for investigating the optical response of an intersubband transition. The study of parabolic quantum wells shows how the intersubband transitions are governed by collective effects and how these collective excitations react to short THz pulses. A detailed study of relaxation times is performed for a coupled quantum well system. High intensity THz pulses allow the observation of non-linear processes in a multi-level quantum well.

In Section Few Cycle THz Spectroscopy of Quantum Cascade Heterostructures, time-resolved THz spectroscopy is employed to study THz Quantum Cascade (QC) heterostructures and lasers with a focus on gain dynamics, dispersion and broadband amplification. In addition, few cycle THz pulses are used to injection seed a broadband THz QCL resulting in very short THz pulses.

Few Cycle THz Spectroscopy of Semiconductor Quantum Structures

There is a large number of studies that use THz pulses for the investigation of semiconductor bulk materials. The complex index of refraction in the THz regime gives insight in the conductivity (i.e., mobility of free carriers) of the material. Free carriers can be either provided by doping or by photoexcitation across the bandgap. There are compelling reasons to use THz pulses to excite and probe carriers in semiconductor quantum structures. The photon energies are comparable to the subband spacing, kinetic energies of carriers, impurity transition and phonon energies. In addition, THz radiation does not generate electron-hole pairs so that the experiments directly probe free carriers rather than excitons.

Linear Response

In an initial experiment (Heyman *et al.*, 1998) THz pulses were used to investigate a modulation-doped GaAs/AlGaAs multiple quantum well ($d=51$ nm) with a transition energy between the two lowest subbands of 1.5 THz and a carrier density of $n_s=2.75\times10^{10}\,\mathrm{cm}^{-2}$ only populating the lowest subband. The THz pulses are coupled into the cleaved edge of the quantum well sample so that the electric field is parallel to the growth direction which is necessary to excite the intersubband transition. On top

of the sample an aluminum Schottky gate is evaporated and the well is contacted with ohmic contacts. The wells can be depleted of carriers by applying a negative gate voltage to the Schottky gate (− 10 V gate bias). By modulating the voltage between 0 and − 10 V the response of the intersubband system can be measured. This modulation technique allows separating the intersubband signal form the response of the SI GaAs substrate that causes a substantially chirped THz pulse due to the frequency-dependent index of refraction. The intersubband response is measured with a cross-correlation technique: Femtosecond laser pulses (100 fs) from a Ti:Sapphire laser are split into two pulse trains. The first one generates THz "driving" pulses which are transmitted through the quantum well. The transmitted pulses are then mixed at the detector with short THz "probe" pulses generated in the second path by photoexcitation of a low-temperature grown GaAs chip. The duration of the probe pulses (~ 100 fs) determines the time resolution and the probe delay controls the time difference. This allows studying the time dependence of amplitude and phase of the transmitted THz pulse. **Fig. 1** shows the cross-correlations signal of the QW. The signal from the carriers rises during the first 2 ps in response to the exciting THz pulses and then continues to oscillate at a constant frequency of 1.5 THz. Due to the femtosecond time resolution in this experiment it is possible to observe the intersubband electrons even during the process of excitation. At first the incident THz pulse generates a coherent superposition between the first and second subband. After the THz driving pulse is over (> 2 ps) the carriers continue to oscillate and radiate at the intersubband transition frequency and the signal is damped out by the free induction decay. Since the amplitude and phase of the radiated electric field is recorded, not only the absorption due to the QW electrons but also the phase delay associated with the absorption can be extracted (**Fig. 1**).

In a subsequent experiment (Kersting *et al.*, 2000) the capability of studying amplitude and phase of THz pulses was used to study a modulation-doped parabolic quantum well. The advantage of parabolic quantum wells is that the transition frequency is independent of the carrier concentration (Kohn's theorem) or the applied electric field. The transition frequency for the electrons in the parabolic quantum well is given by

$$\omega_0 = \sqrt{\frac{8\Delta}{L^2 m^*}}$$

where Δ is the energetic depth of the well, L is the well width, and m* the effective mass. The THz radiation was coupled to the intersubband transition via a metallic Schottky grating. **Fig. 2(a)** and **(b)** shows the response of the QW electrons to *resonant* THz radiation. The electrons follow the driving pulse and continue to radiate at the parabolic quantum well frequency. The response to a non-resonant THz pulse is shown in **Fig. 2(c)–(e)**. First, the electrons inside the QW follow the driving pulse but when the driving pulse is over a phase jump occurs and the electrons lock to their eigenfrequency. The electrons oscillate and emit at the parabolic quantum well resonance frequency and slowly lose their coherence. This behavior was explained by using time-dependent perturbation theory for a two level system with the THz pulse as perturbation. For the time-dependent wave function the ansatz $\Psi(t) = a_1(t)\Psi_1 + a_2(t)\Psi_2$ was used. The time-dependent $a_1(t)$ and $a_2(t)$ coefficients were computed using the measured driving THz

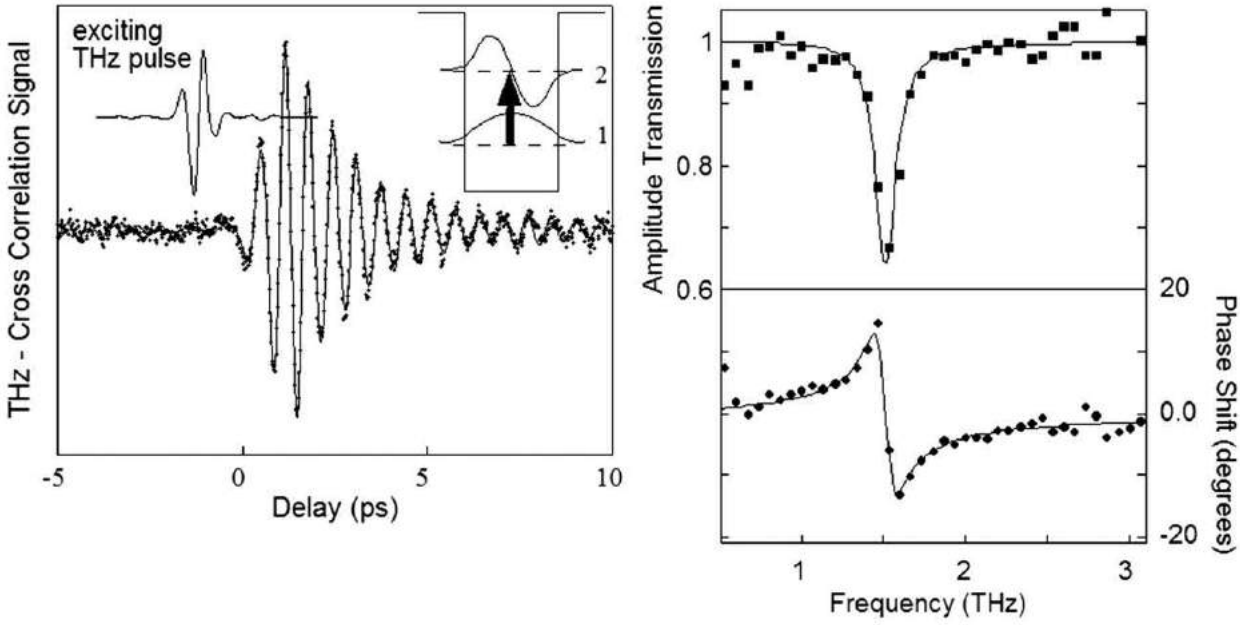

Fig. 1 Left: Measured cross-correlation signal obtained by modulating the charge density in the 51 nm quantum well sample, and recording the change in signal with a lock-in amplifier. The THz pulse excites electrons into a coherent superposition of the n = 1 and n = 2 subband states. The resulting oscillating polarization radiates, producing the oscillations visible in the signal. The upper trace shows the signal recorded with no sample (Adopted from Heyman, J., Kersting,R., Unterrainer,K., 1998. Time-domain measurement of intersubband oscillations in a quantum well. Applied Physics Letters 72, 644.). Right: Absorption coefficient and index of refraction in the quantum well sample due to the intersubband transition. Solid points are calculated from the time-domain data.

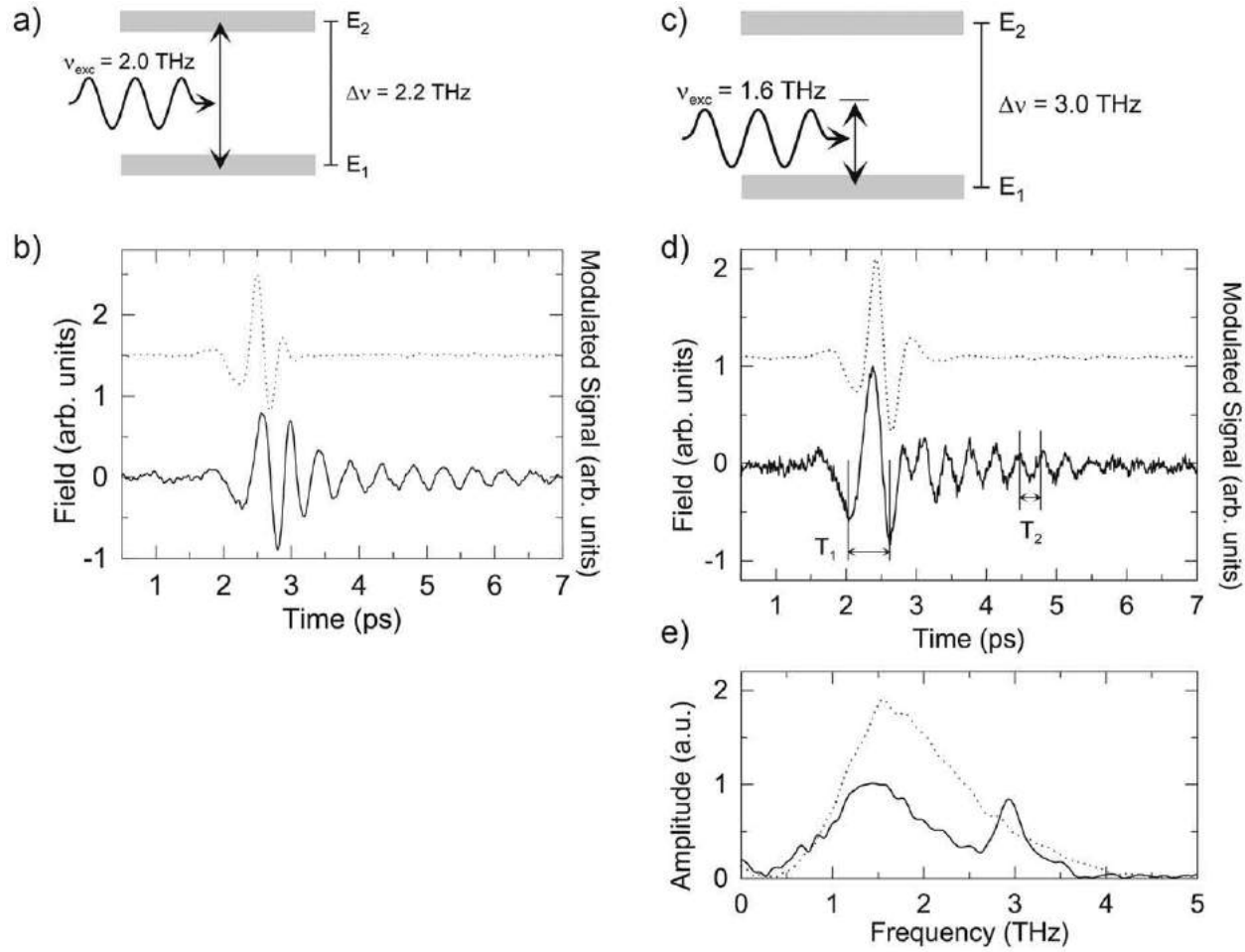

Fig. 2 Resonant and non-resonant excitation of a parabolic quantum well. Adopted from Kersting, R., Bratschitsch, R., Strasser, G., Unterrainer, K., Heyman, J., 2000. Sampling a THz dipole transition with sub-cycle time-resolution. Optics Letters 25, 272.

pulse as input. The comparison to the observed THz transmission is very good and explains the abrupt frequency change as well as the observed phase jump at the end of the driving pulse. This shows that Kohn´s theorem even holds for impulsive and broadband excitation.

Carrier Density Dependence

With time-resolved THz spectroscopy we have also the possibility of using synchronized laser pulses with photon energies above the bandgap. These laser pulses can be used to excite electrons into the subbands of undoped quantum wells. The electrons in the subbands can then be studied by THz pulses (Müller *et al.*, 2004) providing a powerful method to access the dynamical processes in semiconductor quantum wells. It was used to investigate the carrier density dependence of the intersubband relaxation in a double quantum well structure. In this experiment an interband pump pulse excites electrons into the first and second subband of a double QW with a $|1> \rightarrow |2>$ level spacing smaller than the LO phonon energy. The time evolution of the electron population in these two subbands is monitored by probing the intersubband absorption to a third (empty) level. Ultrabroadband THz pulses are generated by phase-matched difference frequency mixing in 30 μm-thick GaSe crystal (Huber *et al.*, 2000) and detected with an interferometric cross-correlation setup (Heyman *et al.*, 1998). The resulting time-dependent THz absorption spectrum exhibits two resonances, corresponding to the $|1> \rightarrow |3>$ and $|2> \rightarrow |3>$ intersubband absorption, respectively. On basis of the absorption spectra the carrier density dependence of the relaxation time between states $|2>$ and $|1>$ can be obtained.

Time-resolved absorption spectra, recorded at a photo-excited sheet carrier density of 1×10^{10} cm^{-2}, are shown in **Fig. 3**. The spectra exhibit absorption peaks at 27 THz and 30 THz, corresponding to the $|2> \rightarrow |3>$ and $|1> \rightarrow |3>$ intersubband transition, respectively. The amplitude of the low-energy peak shows a monotonous decrease with time-delay after excitation due to intersubband relaxation. The amplitude of the second peak, however, rises slightly in the beginning and decays afterwards due to carrier recombination. Since the spectrally integrated absorption is directly proportional to the subband population, it is possible to determine the population dynamics on the basis of the time-resolved absorption spectra. About 40% of the photoexcited electrons

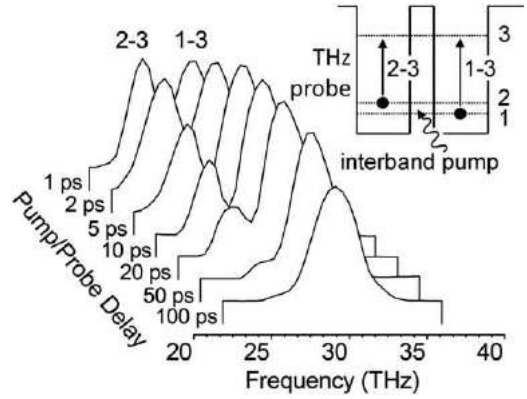

Fig. 3 Intersubband absorption spectra at different time-delays after the interband pump pulse for an excitation density of $n_S = 1 \times 10^{10}$ cm^{-2}. The inset shows the double quantum well and the two intersubband transitions used for probing the populations. Adopted from Müller, T., Parz, W., Strasser, G., Unterrainer, K., 2004. Influence of carrier–carrier interaction on the time-dependent intersubband absorption in a semiconductor quantum well. Physical Review B 70, 155324.

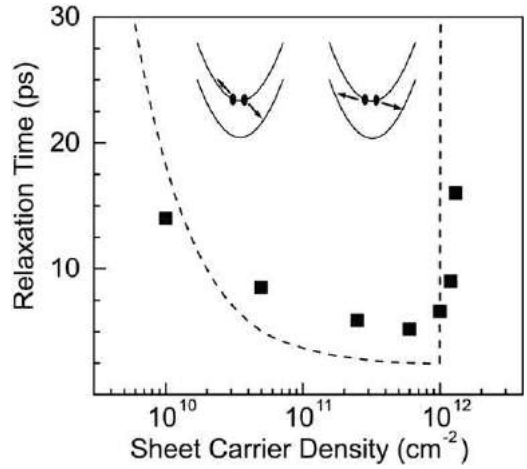

Fig. 4 Experimental intersubband relaxation times (symbols) as function of excitation density. The calculation (dashed line) confirms the experimental results. The inset shows possible two-electron scattering processes. Adopted from Müller, T., Parz, W., Strasser, G., Unterrainer, K., 2004. Influence of carrier–carrier interaction on the time-dependent intersubband absorption in a semiconductor quantum well. Physical Review B 70, 155324.

are injected into the second subband, while the remaining 60% are injected into the first subband at higher k-value. The population in the second subband shows an exponential decay. The electrons which relax down add to the population in the ground level. Subsequently, the population in the ground level drops because of carrier recombination. By fitting the results from a phenomenological rate-equation model to the experimental data, an intersubband relaxation time of 14 ps is obtained for the excitation density of 1×10^{10} cm^{-2} . At a higher excitation density of 3×10^{11} cm^{-2} the dynamics changes a lot. First, the $|2> \rightarrow |3>$ absorption decays much faster, and a relaxation time of only 5.9 ps is obtained. Second, a blue-shift is observed of the $|1> \rightarrow |3>$ absorption as the time evolves, i.e., as electrons in the second subband relax into the first one. The shift can be understood as follows: The depolarization shift of the $|1> \rightarrow |3>$ transition is proportional to the population in subband $|1>$. Thus, as the population in $|1>$ increases, a shift of the resonance to higher energy occurs.

In Fig. 4 the excitation density dependence of the intersubband relaxation time is shown for varying the excitation sheet carrier density between 1×10^{10} cm^{-2} and 1.3×10^{12} cm^{-2}. With increasing carrier density a significant shortening of the $|2> \rightarrow |1>$ relaxation time is observed. In the high density regime a sudden increase of the relaxation time occurs. In order to understand the physical mechanism behind this dependence the possible scattering mechanisms have to be considered and numerical estimates of scattering times have to be compared with the experimental results. Since the energy spacing between the two lowest subbands of our quantum well structure is smaller than the LO phonon energy, the electrons in the second subband do not possess sufficient energy to emit directly LO phonons. Relaxation can thus only be due to acoustic phonon emission and/or carrier-carrier scattering. Using the model described in (Ferreira and Bastard, 1989) the acoustic phonon scattering time was estimated to be $\sim$300 ps for this structure. This time is too long to explain the experimental findings. Earlier time-resolved photoluminescence experiments have shown that the intersubband carrier-carrier scattering rates can be very high, almost approaching in some circumstances the

intersubband scattering rate due to LO phonon emission. Thus carrier-carrier scattering has to be considered and scattering rates were calculated in the Born-approximation (Smet *et al.*, 1996). The results of this calculation are shown in **Fig. 4** as dashed line. At an excitation density of $\sim 1 \times 10^{12}$ cm^{-2} the Fermi-level in the ground state approaches the bottom of the second subband and the intersubband relaxation drastically slows down due to Pauli-blocking. This is illustrated by the vertical dashed line. The strong dependence of the intersubband relaxation time on the carrier density and the good agreement with the calculation shows the importance of carrier-carrier scattering for intersubband relaxation. As a result, even for intersubband transitions below the optical phonon energy the relaxation times will be much shorter than the acoustic phonon scattering time which is very important to consider for all device concepts based on intersubband transitions.

Nonlinear Response

Going to the non-linear regime promises additional very interesting insights. So far, the spectroscopy of intersubband excitations in quantum wells subject to intense THz irradiation has been mainly limited to narrowband pulses from free-electron lasers (Craig *et al.*, 1996) or has been based on optical probing of intraband transitions. The direct observation of intersubband dynamics using few-cycle pulses has been either restricted to the linear regime or has been carried out at higher frequencies in the mid-infrared region (Luo *et al.*, 2004; Dynes *et al.*, 2005). The THz spectral range offers unique possibilities to study novel non-linear physics. For instance, the vacuum Rabi splitting can become a significant fraction of the intersubband transition frequency, giving rise to intersubband polaritons and the so-called ultrastrong coupling regime (Günter *et al.*, 2009; Ciuti *et al.*, 2005). In addition, the large bandwidth of single-cycle THz pulses allows the coupling of adjacent intersubband transitions. This opens the road to novel physics beyond the two-level atom model. One very important ingredient for quantum optics and quantum information processing in general is the ability to prepare the system in a certain quantum state. The simplest example would be the controlled transfer of population from the ground state to any of the excited states, for example, to achieve a complete population inversion between two subbands. Apart from the pure scientific interest, such an inverted system may also be useful as a switchable THz gain medium in real world applications, such as THz amplifiers or Q-switches. Dietze *et al.* (2012) reported the first direct observation of intersubband dynamics in a modulation doped multiple quantum well (MQW) sample subject to intense few-cycle THz pulses. Single cycle THz pulses with a bandwidth of 5.5 THz and maximum electric field amplitude of 20 kV/cm are focused on the cleaved facet of the multi quantum well sample (52 nm well width, 10 periods). By applying a bias voltage, the wells can be depleted which allows using an electrical modulation technique to selectively measure the effect of the electronic polarization on the transmitted THz pulses. **Fig. 5** shows the level scheme of the QWs. The transition frequencies have been calculated self-consistently including the depolarization shift. At 5 K, all electrons are assumed to be in the ground state. Thus, with the bandwidth of the driving pulse indicated by the blue bar, only the first three states can be accessed. The pump spectrum peaks around 1.4 THz, which is close to the first transition $|1> \rightarrow |2>$. For sufficiently high electric field strengths, Rabi oscillations between the ground and first excited state will lead to a level splitting, which can be probed by the next higher transition $|2> \rightarrow |3>$. Furthermore, the direct transition $|1> \rightarrow |3>$ is dipole forbidden, but state $|3>$ can be populated via two interfering pathways, either through the intermediate state $|2>$ or directly by a 2-photon resonant transition (indicated by red arrows). This quantum interference might lead to an enhanced transmission similar to electromagnetically induced transparency (Luo *et al.*, 2004). In addition, the absorption strength of each transition is dependent on the relative occupation numbers of the involved levels. Thus, for strong driving, we expect to observe saturation of the intersubband transitions.

Thus we expect that the non-linear effects are thereby strongly dependent on the electric field amplitude of the THz driving pulse. **Fig. 6** shows the measured modulation signal, ΔE, for different values of the pump power incident on the GaP crystal. For the chosen experimental parameters, the THz peak field scales almost linearly with the optical pulse energy. At the lowest pump energy of 125 μJ, the incident THz peak field is estimated to be 0.8 kV/cm. The modulation signal shown in **Fig. 6(a)** corresponds to the expected free-induction decay of a weakly driven two-level system (black curve). The oscillations are monochromatic with a center frequency of 1.5 THz and a dephasing time of 2.3 ps. When the THz amplitude is increased, a beating pattern gradually emerges, indicated by the black arrows in **Fig. 6(b)**. This beating is a clear indication for a second frequency component in the spectrum. We attribute this to increased occupation of the first excited level, which activates the next transition $|2> \rightarrow |3>$.

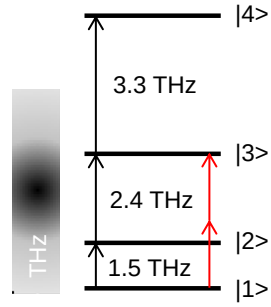

Fig. 5 Level scheme of the quantum well and possible transitions.

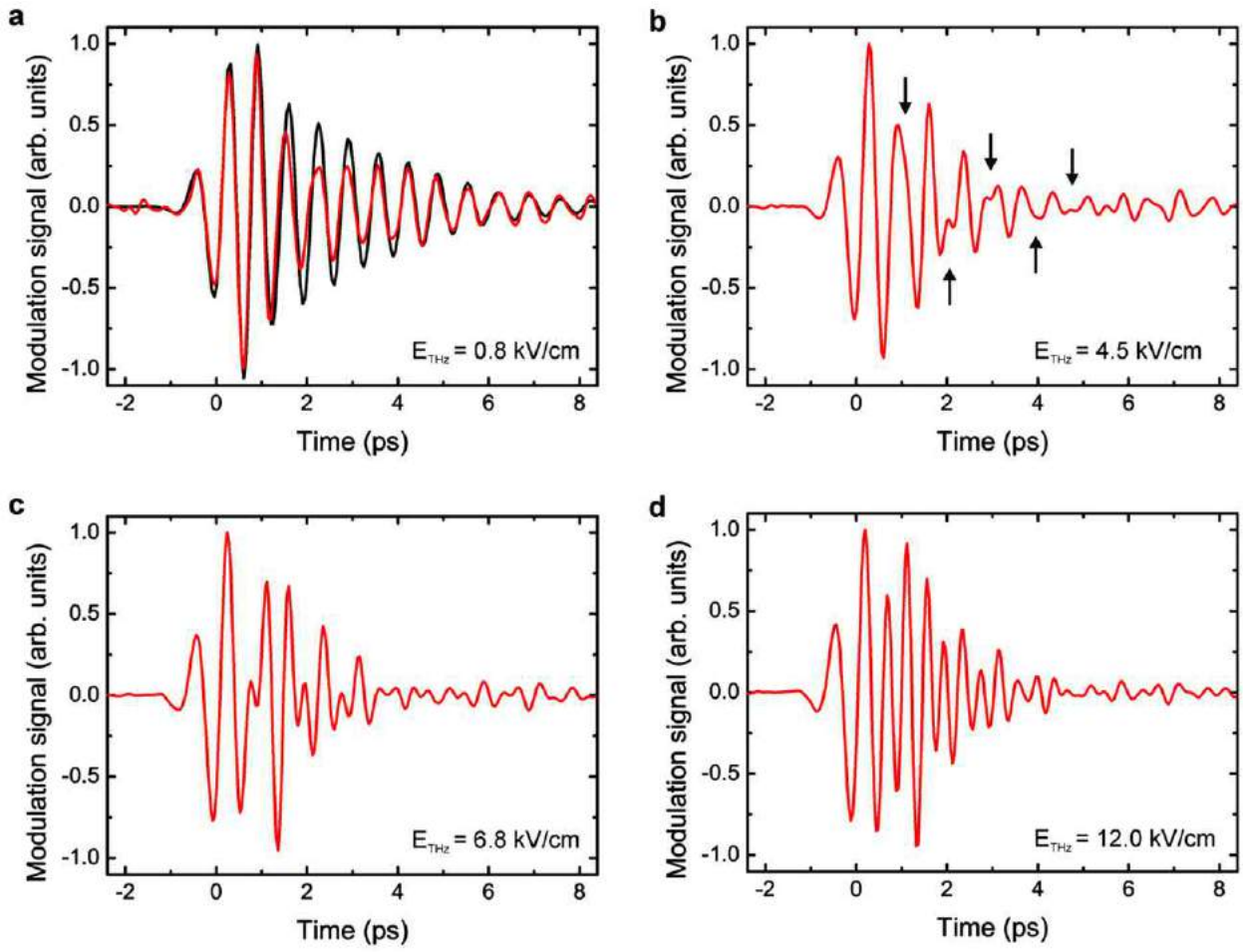

Fig. 6 Modulation signals for different pump powers. (a) For the lowest THz field amplitude, the free induction decay is almost monochromatic (red). The black curve shows the results of FDTD simulations for a two-level system. (b) For 4.5 kV/cm, a beating becomes obvious. The temporal positions of the beat nodes are indicated by the arrows. (c) For higher peak fields, the total pulse becomes shorter and shows echo-like features. (d) Above 12 kV/cm, the free induction decay contains several higher frequencies. Adopted from Dietze, D., Darmo, J., Unterrainer, K., 2012. THz-driven nonlinear intersubband dynamics in quantum wells. Optics Express 20, 23053.

In addition, the modulation signal decays faster indicating that additional dephasing mechanisms occurring compared to the simple two-level model.

Around 6.8 kV/cm (**Fig. 6(c)**), the beating pattern gets more pronounced and resembles the photon echoes produced by the free-induction decay. This pulse-like structure is an indication for a broadened modulation spectrum which contains many frequency components. The character of the modulation signal changes again for very high pump powers (**Fig. 6(d)**). Both the oscillation period and the decay time are much faster than in **Fig. 6(a)**. **Fig. 7** shows the modulation spectra obtained by Fourier transforming the time domain signals. The signal originating from the free induction decay of the three allowed intersubband transitions are marked by 1, 2, and 3. For the lowest driving field, there is only a single feature present around 1.5 THz. For increasing pump powers, electrons are efficiently transferred to higher lying states and the higher transitions start to appear in the spectra: the $|2> \rightarrow |3>$ transition around 2.4 THz and the $|3> \rightarrow |4>$ transition at 3.3 THz. The 2.4 THz peak shows a doublet structure, which vanishes for higher pump powers. The observed behavior is in principle consistent with the scenario of sequential pumping of electrons from the ground state to higher excited states by the broadband THz pulse. For intermediate peak electric fields around 6 kV/cm, an additional feature with a broad frequency content appears centered in between the 1.5 and 2.4 THz transition (marked by asterisk). To determine whether this effect is related to a real electron current (in-phase), we have calculated the (single-pass) absorption according to:

$$\alpha(\omega) \sim \mathrm{Im}(-\mathrm{i}\Delta E(\omega)/E_{\mathrm{ref}}(\omega))$$

which is valid for the assumption, that the depleted sample is non-absorbing and $\Delta E \ll E_{\mathrm{ref}}$. **Fig. 7(b)** shows the absorption spectra associated to **Fig. 7(a)**. The shaded area is excluded from the discussion due to the limited signal-to-noise ratio. The broadband spectral feature has disappeared, which indicates that it is indeed related to an in-phase polarization. It resembles the ponderomotive contribution from free carriers in the quantum well plane (Golde *et al.*, 2009). However, the spectrum of the broad

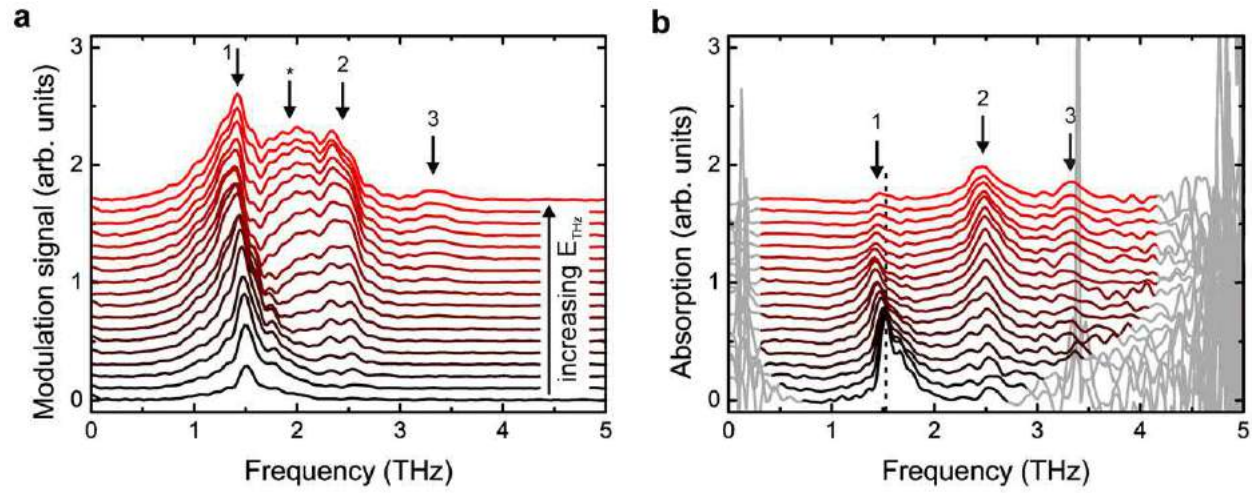

Fig. 7 Modulation and absorption spectra for different pump powers. (a) Modulation spectrum. The three allowed transitions are marked by (1, 2, 3). There appears also a broadband feature for higher pump powers marked by the asterisk. (b) Absorption spectrum. The broadband feature has disappeared. The fundamental transition exhibits a clear red shift for higher pump powers. The curves are vertically offset for clarity. Adopted from Dietze, D., Darmo, J., Unterrainer, K., 2012. THz-driven nonlinear intersubband dynamics in quantum wells. Optics Express 20, 23053.

feature does not coincide with the spectrum of the driving pulse, and, in addition, the chosen experimental geometry excludes the in-plane ponderomotive current because the THz electric field is oriented perpendicular to the quantum well plane. As an alternative explanation it is suggested that this coherent signal is a consequence of the non-linear mixing of the THz pulse with the induced in-phase polarization of the quantum well (Dietze *et al.*, 2013). The intersubband absorption shows a clear dependence on the pump power. The decrease of the absorption peak associated to the $|1> \rightarrow |2>$ transition is a sign for THz induced saturation of the intersubband transition. In addition, the transition frequency is shifted to lower values for increased amplitudes of the driving field. This effect has been predicted theoretically to be related to an undressing of the collective intersubband excitation which is causing the depolarization shift of the resonance frequency (Zaluzny, 1993). Previously it has only been observed experimentally using intense narrowband THz radiation from a free electron laser with peak powers on the order of 1 kW. The absorption line of the $|2> \rightarrow |3>$ transition shows a doublet-like structure for optical 6 pump energies below 1 mJ, which is reminiscent of the Autler-Townes effect in a three-level system (Zaks *et al.*, 2011). The disappearance of the splitting with increasing field strengths, however, is counter intuitive. A possible explanation could be based on the scaling properties of the vacuum Rabi splitting in an ensemble of two-level atoms: $\Omega_R \sim (n_{2d})^{-1}$, where n_{2d} is the areal density of electrons (Ciuti *et al.*, 2005). In the present case, electrons are shared among all levels. Thus, the level splitting is proportional only to the number of electrons participating in the coherent population transfer between the ground and first excited state. For higher pump field strengths, electrons are efficiently transferred to the higher lying states and are not accessible for the Rabi oscillations. As a consequence, the level splitting disappears.

Few Cycle THz Spectroscopy of Quantum Cascade Heterostructures

Quantum Cascade semiconductor (QC) heterostructures are very powerful structures providing optical gain in the frequency range spanning from near-infrared to far-infrared (Faist *et al.*, 1994; Köhler *et al.*, 2002). Thereby, the optical gain is achieved by the optical transition between appropriately spaced quantized levels of QC heterostructures. Due to the vast space of design and technology parameters, the detailed information on the optical gain value, its bandwidth and profile, the gain recovery time and their dependence on bias and operating temperature, is of great importance. Few-cycle THz spectroscopy is the tool of choice to access this information.

Usually, studying the optical gain using standard detectors recording the power emitted by the laser is limited to operation conditions below the lasing threshold. For the first time, Kröll *et al.* (Kröll *et al.*, 2007) have accessed the internal laser processes in the whole range of operating conditions. For detection of the THz signal electro-optic detection (EOD) (Wu *et al.*, 1996) featuring a sub 100 fs resolution is used. Thereby, the necessary synchronization between the THz signal and the probe is guaranteed by THz photon mutually phase-locked to the probe and injected into the investigated laser structure. Therefore, only the electric field of the original photons and of those generated by a stimulated emission (hence with a timing and phase equal to the injected photons) is detected, while photons generated by amplified spontaneous emission feature a random phase and timing produce a zero averaged EOD signal. This technique originally demonstrated for THz quantum cascade lasers (Kröll *et al.*, 2007), has been later extended for mid-infrared lasers (Parz *et al.*, 2008) and near-infrared coherent sources (Keiber *et al.*, 2016).

Fig. 8 schematically shows the layout of a typical THz time-domain system (TDS) used to investigate THz QC heterostructures. It does not differ from the standard THz-TDS (Smith and Arnold, 2011), except of the signal modulation. In THz-TDS, the lock-in

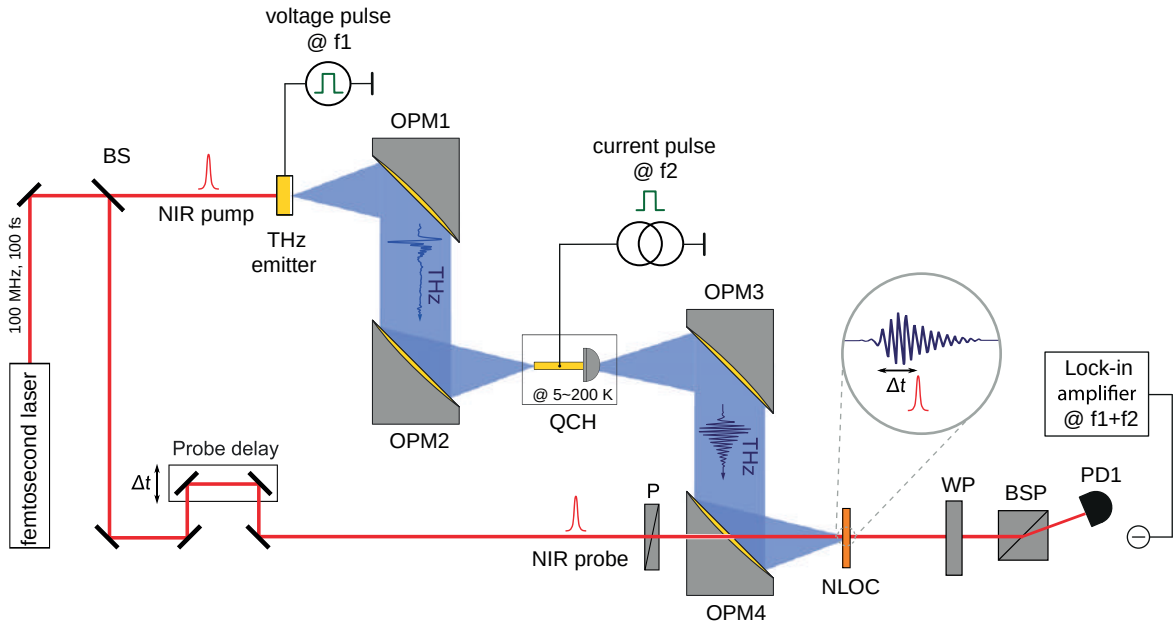

Fig. 8 Time-domain spectroscopy setup for investigation of THz Quantum Cascade heterostructures. BS, beam splitter; NLOC, non-linear optical crystal; OPM, off-axis parabolic mirror; P, NIR polarizer; PBS, polarizing beam splitter; PD, near-infrared photodetector; WP, $\lambda/4$ waveplate.

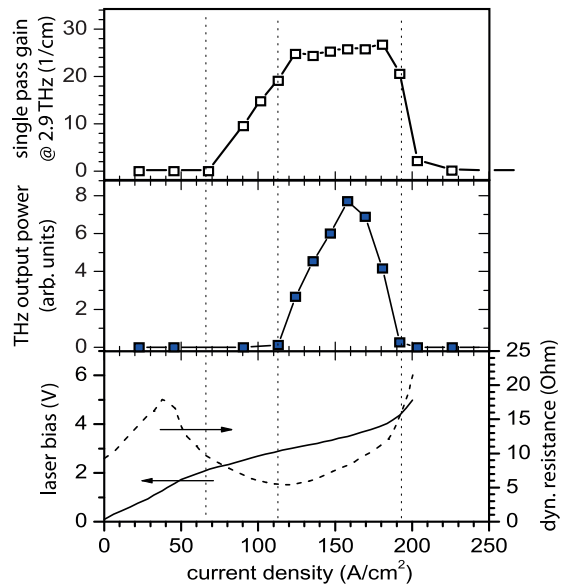

Fig. 9 Basic electrical and optical characteristics of a THz QCL: (bottom panel) the current-voltage dependence of the device (solid line) and corresponding dynamic resistance (dashed line); (middle panel) the current – THz light intensity characteristics; (top panel) the gain observed from the THz injection measurement. Details on THz QCL and method is adopted from Kröll, J., Darmo, J., Dhillon, S.S., *et al.*, 2007. Phase resolved measurements of stimulated emission in a laser. Nature 449, 698–702.

technique is typically used to analyze the signal from the THz detector (photocurrent of a THz antenna or of a balanced photodetector) that inherently requires a modulation of the THz signal. When an active device is tested, an additional modulation of the device is required, so a double–modulation technique is employed for the investigation of THz QC heterostructures. Thereby, the THz emitter is chopped at frequency f_1, the THz QC heterostructure is electrically modulated at frequency f_2 and the signal at the THz electric field detector is demodulated at $f_1 \pm f_2$. Due to the high nonlinearity of both modulations of the detector and the QC heterostructure, the modulation frequencies should preferably be chosen not to be harmonic pairs (Kröll *et al.*, 2007).

Using THz-TDS Kröll *et al.* (Kröll *et al.*, 2007) have investigated a THz QC heterostructure providing a gain at 2.7 THz at different operation conditions (**Fig. 9**). As expected, the THz gain is a function of the electric current flowing through the device and is correlated with the THz power emitted from the laser and with the current-voltage characteristics. The injected THz photons

get amplified by stimulated emission well below the laser threshold (at a current density of $\sim$110 A/cm^2) and the gain steadily increases as the electron injection into upper lasing level is rising with increasing current through the device. Thereby, the onset of the gain (at a current density of $\sim$110 A/cm^2) appears at the applied voltage necessary to reach the correct QC heterostructure alignment. The drop of the dynamic resistance indicates that this condition is reached. When the THz gain is high enough to compensate waveguide and mirror losses, the lasing threshold is reached. Coherent emission starts and the THz gain gets clamped to the value of the total laser loss. This regime is observed for a device current density between 110 and 180 A/cm^2, for which the measured THz gain remains almost constant. The residual weak increase of the gain with the device current is attributed to the gradual increasing total loss due to, for example, heating. Lasing stops when the bias applied to the QC heterostructure is too high and the heterostructure alignment is broken. Accordingly, the observed THz gain drops very quickly (for a current density $>$180 A/cm^2).

The issue of the clamped gain has been addressed in a follow-up work by Jukam *et al.* (Jukam *et al.*, 2009). They used ultrafast switching of the QC heterostructure device current in order to achieve effective THz gain switching. When the switching was synchronized to the optical injection, a temporal increase of the gain has been observed (Oustinov *et al.*, 2010). It is well demonstrated for the amplification of the THz probe pulse injected into the laser device (see **Fig. 10**). If the probe pulse is entering the laser before the gain switching occurs, the pulse experiences a steady-state gain. However, when the gain is switched before the probe pulse is injected, the amplification is significantly increased. The details of the dynamics of the gain rise, persistence and decay depend on the THz pulse length ($\sim$1.5 ps), the transit time ($\sim$18 ps), and the temporal characteristics of the switching pulse itself ($\sim$100 ps). It is worth noting, that after the gain switching event the gain eventually drops to values below the steady-state values, indicating a significant gain depletion due to a temporal burst of the THz emitted power.

When the length of the gain switching pulse is much longer than the QCL cavity round-trip time, a very complex THz pulse dynamics is observed (Bachmann *et al.*, 2015). In **Fig. 11** three different operation modes of a THz QCL are compared. First, the THz pulse train formed in the device driven close to threshold is shown. The injected few-cycle THz pulse gets gradually attenuated as at this operation point the total loss still exceeds the gain in the laser device. For this condition, the cavity mirrors loss dominates. When the laser is cw operated at the point far above threshold, a stable train of THz pulses is generated, although the pulse amplitude is small in comparison to the THz seed amplitude due to the small spectral overlap of the THz seed and THz gain, and the small clamped THz gain. The latter limit is readily overcome by employing gain switching. Voltage pulses as short as 600 fs bring the laser from the sub-threshold operation point to the operation for maximum output power (in the CW regime). This switching has also a dramatic impact on the THz pulse train formation. The pulse amplitude exceeds the amplitude of the THz seed in spite of the very different spectral content which indicates significant amplification of the frequency components within the THz gain spectrum. The pulse amplitude closely follows the temporal shape of the voltage pulse with small lag at the rising and falling edges. At the beginning, the THz pulse needs about two cavity round-trips to reach saturation of amplification, while the slow decay of the pulse amplitude is a typical cavity ring-down behavior. Finally, the time domain output of the gain switched QCL gets structured due to the modal dispersion which is discussed later in this section.

The experimental observations mentioned in the previous paragraphs have been compared with a general model of QC heterostructure based lasers. A single Quantum Cascade (**Fig. 12(a)**) can be considered as a two-level quantum system, where the upper states can be populated by electron injection or by absorption from the occupied ground state. Such a system can be described by a density matrix and its time development is given by the optical Bloch equation (OBE). OBE together with the Maxwell equations form the theoretical base for the complete description of the interaction of the electromagnetic wave with the THz QC heterostructure gain medium. Since the THz time domain data have to be modeled, the features of the THz TDS need to be also considered. In standard THz TDS the changes of the few-cycle THz pulse after interaction with an object are evaluated in terms of the complex permittivity (or permeability) of the object. The injected THz photons, however, can be scattered, absorbed, multiplied or temporarily trapped in the device cavity. Hence, the time domain data have to be evaluated with respect to these

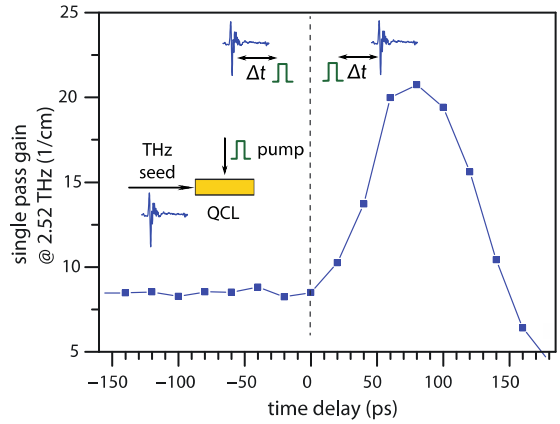

Fig. 10 Gain switching of a THz QC heterostructure with an ultra-fast photoconductive switch. Adopted from Oustinov, D., Jukam, N., Rungsawang, R., *et al.*, 2010. Phase seeding of a terahertz quantum cascade laser. Nature Communications 1, 69–75.

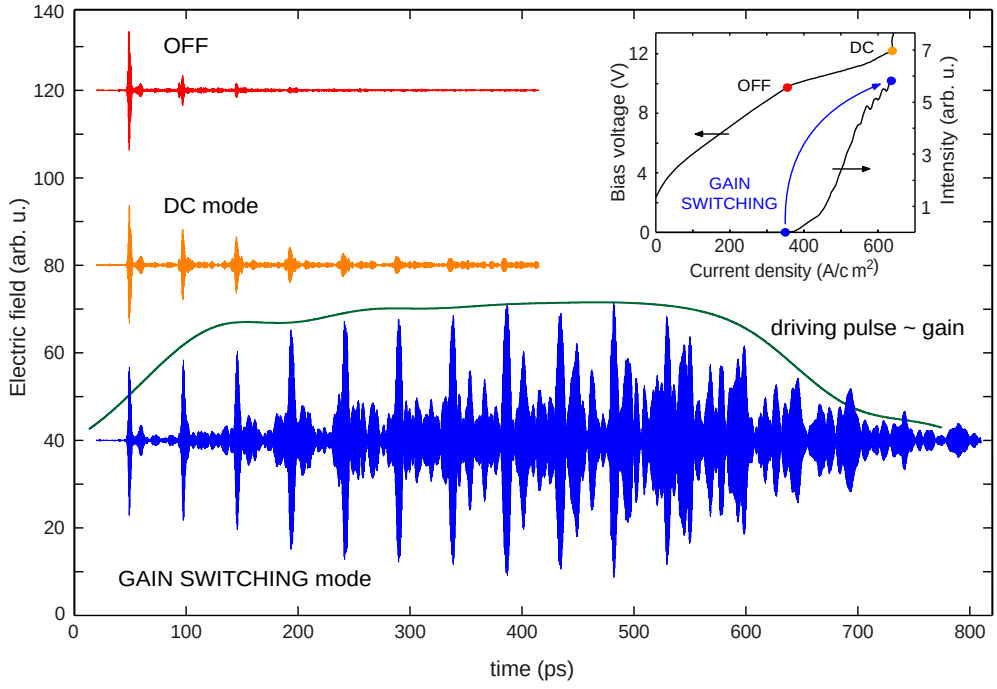

Fig. 11 Dynamics of the gain and THz pulse train formation in a broadband THz QC heterostructure driven with a sub-nanosecond voltage pulse (its envelope indicated with green solid line). Adopted from Bachmann, D., Leder, N., Rösch, M., *et al.*, 2015. Broadband terahertz amplification in a heterogeneous quantum cascade laser. Optics Express 23, 3117–3125.

processes and a detailed model of the device has to be considered. A complete QC heterostructure based laser model has been applied by Kröll *et al.* (Kröll *et al.*, 2007), Darmo *et al.* (Darmo *et al.*, 2006) and Freeman *et al.* (Freeman *et al.*, 2013) to study the THz emission in time-domain with respect to different aspects of gain medium and waveguide. The emission of a driven two-level system is demonstrated in **Fig. 12(b)**, which shows the internal electric field in a QCL waveguide of the length of 2 mm when a THz seed with a bandwidth of 300 GHz is injected. The amplitude of the electric field is not constant and indicates the propagation of a dispersed pulse. This internal amplitude profile corresponds to the externally observed laser emission in the time-domain, hence the internal dynamics of the laser can be monitored. The effect of the laser cavity is demonstrated in **Fig. 12(c)**. Therein the impact of the material and modal dispersion and the modal phase diffusion is shown.

For the purpose of frequency comb generation and the subsequent achievement of pulsed output of the QCL, the knowledge of the dispersion of the QC heterostructure based emitter is crucial. For the THz QCL this issue has been recently addressed by Burghoff *et al.* (Burghoff *et al.*, 2014), Rösch *et al.* (Rösch *et al.*, 2015) and Bachmann *et al.* (Bachmann *et al.*, 2016a). The system dispersion can be assessed from the few cycle THz pulses circulating within the QCL cavity due to partial reflections on the cavity mirrors (Burghoff *et al.*, 2014; Bachmann *et al.*, 2016a). During the cavity round-trip the frequency components of the THz pulse accumulate different phase shifts according to the frequency dependent effective refractive index of the QCL cavity. Bachmann *et al.* (Bachmann *et al.*, 2016a) performed a detailed analysis of this phase shift for a THz QCL with a heterogeneous QC heterostructure (Turčinková *et al.*, 2011). **Fig. 13(a)** shows the observed spectral dependence of the group velocity dispersion (GVD) that exhibits strong oscillations in the whole gain bandwidth of the QC heterostructure. Bachmann *et al.* (2016a) have looked into the origin of these oscillations considering a model in which the dispersion in the THz QCL consists of contributions from the dispersion of the heterostructure material; the waveguide and from the optical gain of the intersubband transitions. The measured GVD can only be fitted when the optical gain induced dispersion is considered (see **Fig. 13(b)**). Thereby a real design of the QC heterostructure has been considered (Turčinková *et al.*, 2011; Bachmann *et al.*, 2014) with three different sections. The result demonstrates unambiguously that the QCL cavity dispersion (in terms of the group velocity dispersion GVD) is dominated by the dispersion caused by the optical gain medium (Bachmann *et al.*, 2016a). Therefore, the dispersion is dependent on the operation temperature and on the bias of the QCL (**Fig. 13(b)**), the fact has to be considered for the correct design of a dispersion compensating structure (Burghoff *et al.*, 2014; Villares *et al.*, 2016).

Injection seeding of a QCL with a few-cycle THz pulse (Kröll *et al.*, 2007) enables also to map the laser cavity modes. The used seeding mechanism (Kröll *et al.*, 2007; Dhillon *et al.*, 2010; Burghoff *et al.*, 2011) does not prefer certain modes, but feeds all possible cavity modes, including higher order lateral modes. Different lateral modes travel at different group velocities leading to an additional dispersion of the injected THz pulse and instead of clean separated pulses a rich dynamic inter-pulse structure is formed (**Fig. 14(a)**). Bachmann *et al.* (Bachmann *et al.*, 2016b) have demonstrated the impact of lateral mode control on THz pulse formation in a THz QC heterostructure operating in the amplifier regime. When the propagation loss of higher-order lateral modes is compromised (e.g., by an additional loss), almost bandwidth limited THz pulses are generated from the few-cycle THz seed

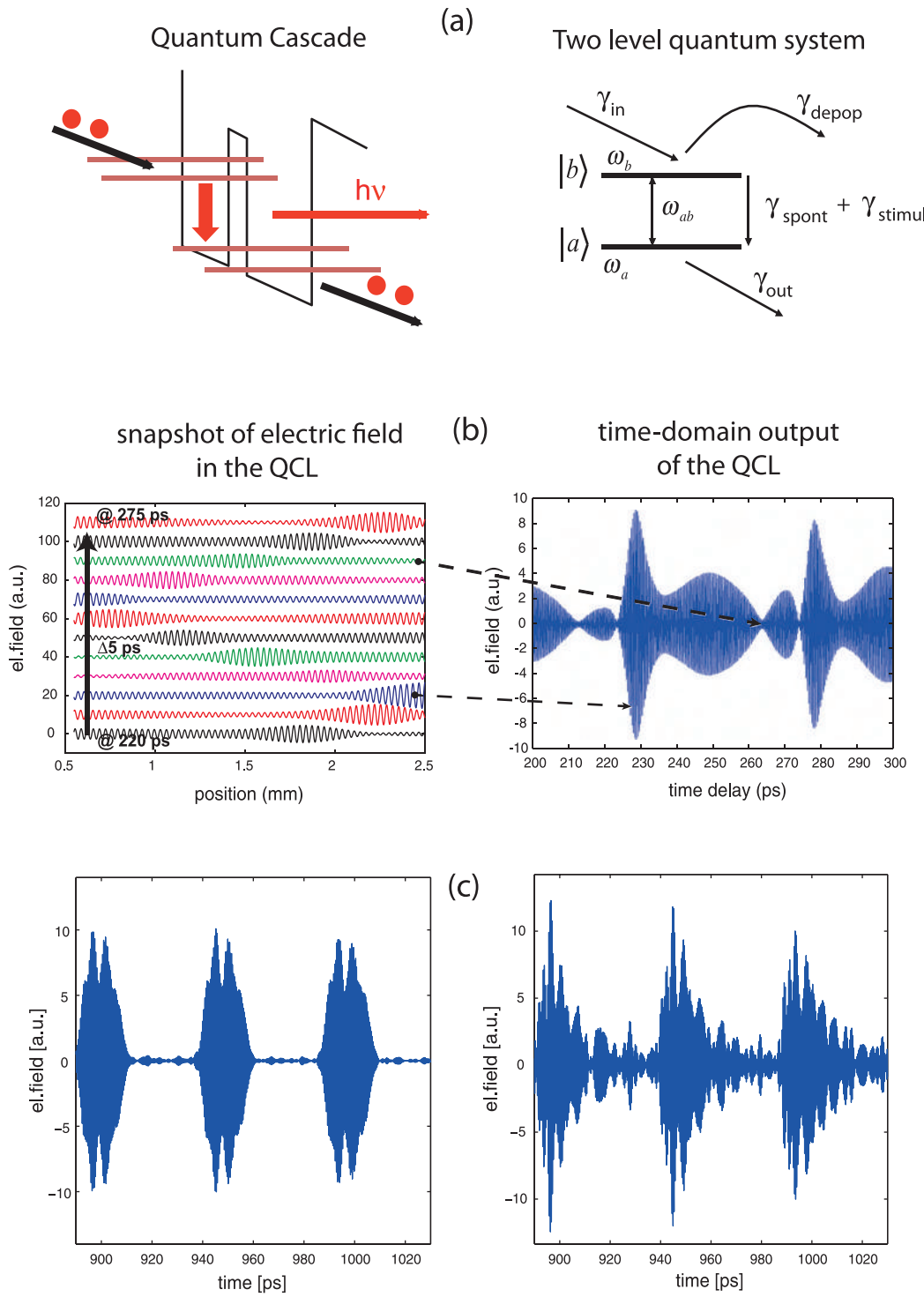

Fig. 12 Modeling the time response of a THz QC heterostructure: (a) schematic drawing of a Quantum Cascade and its two-level quantum-mechanical model; (b) electric field time snapshots in the THz QCL waveguide and their correlation with the laser output waveform; (c) impact of different parameters on the laser output.

(**Fig. 14(b)**). Thereby, the seeded THz gain bandwidth of ~400 GHz guarantees the THz pulse length of ~2.5 ps. Such a significant control of the THz pulse form was achieved by using lateral mode control in the form of absorbing boundary conditions on the edges of the QCL cavity ridge (Bachmann *et al.*, 2016b).

In addition to the very successful application as laser device, THz QC heterostructure based systems can also be used as amplifier of the few-cycle THz pulses. This approach is especially appealing for the THz spectroscopy community due to the THz

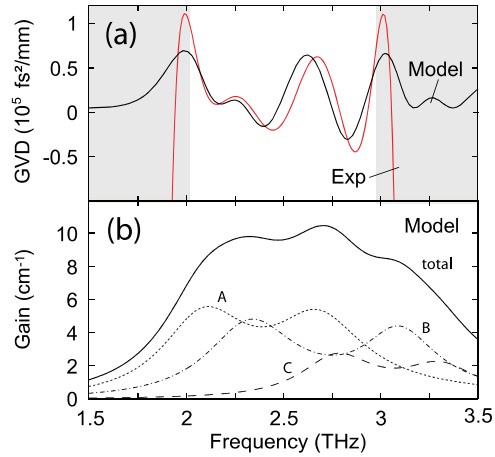

Fig. 13 Dispersion in a THz broadband QCL: (top panel) measured group velocity dispersion GVD (red line) and model GVD (black line); gray shaded areas indicate the spectral range with too low signal-to-noise ratio to extract correct GVD values; (bottom panel) model gain curve assembled from three different QC sections fitting the experimental GVD. Adopted from Bachmann, D., Rösch, M., Scalari, G., *et al.*, 2016a. Dispersion in a broadband terahertz quantum cascade laser. Applied Physics Letters 109, 221107. doi:10.1063/1.4969065.

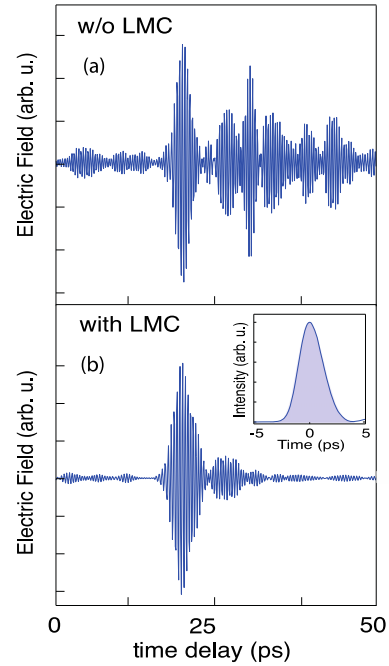

Fig. 14 The electric field of a THz pulse formed in the seeded THz-QCL: (a) device without lateral mode control (LMC); (b) device with lateral mode control. Inset: pulse intensity profile indicating ~2.5 ps pulse length. Adopted from Bachmann, D., Rösch, M., Süess, M.J., *et al.*, 2016b. Short pulse generation and mode control of broadband terahertz quantum cascade lasers. Optica 3, 1087–1094. doi:10.1364/OPTICA.3.001087.

QCL operating frequency window of 2.0–5.0 THz, i.e., the frequency window in which typical THz sources used for the THz time-domain spectroscopy have a severely compromised spectral brightness. First THz amplification demonstrations have been done with a THz QC heterostructure providing a limited gain bandwidth of <200 GHz (Jukam *et al.*, 2009; Dhillon *et al.*, 2010), hence with a low potential for applications. The successful development of broadband THz QC heterostructures (Turčinková *et al.*, 2011; Freeman *et al.*, 2011) enabled a THz seed amplification with a bandwidth as large as 600 GHz (Bachmann *et al.*, 2015). Recently, Bachmann *et al.* (Bachmann *et al.*, 2016b) demonstrated the amplification of the THz electric field strength by >20 dB in the spectral window between 2.1–3.0 THz (**Fig. 15**) in a mode controlled broadband THz QC heterostructure based Fabry-Perot amplifier. Thereby, a train of ~2.5 ps long THz pulses with a total length of ~650 ps is generated per single THz seed pulse. The pulse train length is defined by the gain switching time window (Bachmann *et al.*, 2015).

Finally, THz injection locking (Oustinov *et al.*, 2010) or RF injection locking (Barbieri, 2011) synchronized with a coherent sampling femtosecond laser are key approaches to build a THz system for the remote sensing/imaging applications, that is

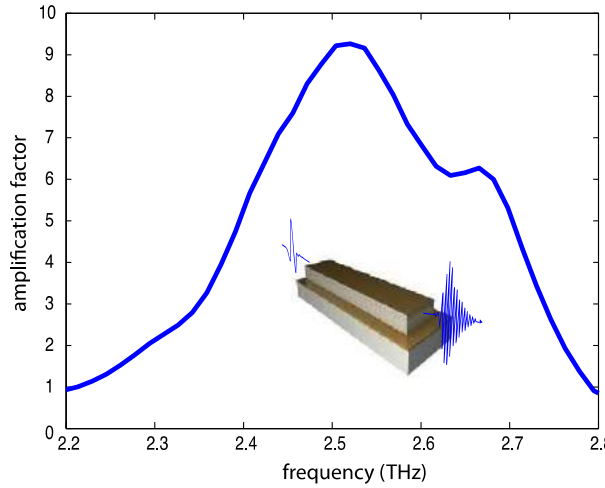

Fig. 15 Amplification factor achieved for the broadband THz QC heterostructure based Fabry-Perot amplifier. The amplification profile follows the gain profile. Adopted from Bachmann, D., Leder, N., Rösch, M., *et al.*, 2015. Broadband terahertz amplification in a heterogeneous quantum cascade laser. Optics Express 23, 3117–3125.

harvesting of the strengths of both THz QC heterostructure technology and THz time-domain spectroscopic technique (Darmo *et al.*, 2010a,b).

Conclusions

Semiconductor nanostructures are very exciting object to study light matter interaction and as a building block for novel optoelectronics devices. The ability to design the energies of the quantized transition, their optical matrix element and the tunnel coupling is unprecedented in solid state physics. Few-cycle THz spectroscopy is a perfectly suited method to study these structures as it provides time-resolution and phase resolution. This allows accessing the full response of the nanostructures to THz radiation. Linear experiments reveal the decay of the intersubband polarization with a decay time of several ps. The carrier lifetime is obtained by optical injection of carriers and monitoring their population dynamics with THz probe pulses. The lifetimes vary as function of their density and range from a few ten ps to a few ps due to carrier-carrier scattering. High intensity THz pulses allow studying the whole field of non-linear optics in semiconductor nanostructures. Higher order transitions, field induced level spitting and complex many body effects are demonstrated. Semiconductor nanostructures enable the realization of Quantum Cascade lasers which are becoming the most powerful compact THz radiation sources. The study of the gain bandwidth, saturation and bias dependence is pursued by few-cycle THz pulses and by coherent electro-optic detection. This unique combination of a designable gain medium and THz time-domain spectroscopy provided for the first time direct insight to the process of stimulated emission and lasing. The formation of spatial hole burning, the influence of the intersubband gain dispersion on the THz propagation, and the saturation of the gain are directly observed. The use of a ultrabroad band QC heterostructure gain medium is enables the realization of octave spanning lasers and frequency combs. THz time domain spectroscopy provides a direct measurement of the spectral gain and of the dispersion. The knowledge and control of the dispersion is crucial for frequency comb operation and also for short pulse generation. Injection seeding of ultrabroad band QC heterostructures resulted in the observation of pulses as sort as 2.5 ps. This successful development will continue by studying, for example, strong coupling of intersubband transition to metamaterials, modelocking and intersubband transition in novel 2D materials.

See also: Foundations of Coherent Transients in Semiconductors. Semiconductor Lasers. Terahertz Lasers

References

Bachmann, D., Leder, N., Rösch, M., *et al.*, 2015. Broadband terahertz amplification in a heterogeneous quantum cascade laser. Optics Express 23, 3117–3125.
Bachmann, D., Rösch, M., Deutsch, C., *et al.*, 2014. Spectral gain profile of a multi-stack terahertz quantum cascade laser. Applied Physics Letters 105, 181118.
Bachmann, D., Rösch, M., Scalari, G., *et al.*, 2016a. Dispersion in a broadband terahertz quantum cascade laser. Applied Physics Letters 109, 221107.
Bachmann, D., Rösch, M., Süess, M.J., *et al.*, 2016b. Short pulse generation and mode control of broadband terahertz quantum cascade lasers. Optica 3, 1087–1094.
Barbieri, S., 2011. Coherent sampling of active mode-locked terahertz quantum cascade lasers and frequency synthesis. Nature Photonics 5, 306–313.
Burghoff, D., Kao, T.-Y., Ban, D., *et al.*, 2011. A terahertz pulse emitter monolithically integrated with a quantum cascade laser. Applied Physics Letters 98, 061112.
Burghoff, D., Kao, T.-Y., Han, N., *et al.*, 2014. Terahertz laser frequency combs. Nature Photonics 8, 462–467.

Ciuti, C., Bastard, G., Carusotto, I., 2005. Quantum vacuum properties of the intersubband cavity polariton field. Physical Review B 72, 115303.
Craig, K., Galdrikian, B., Heyman, J.N., *et al.*, 1996. Undressing a collective intersubband excitation in a quantum well. Physical Review Letters 76, 2382–2385.
Darmo, J., Bachmann, D., Unterrainer, K., 2010a. Temporal and spectral aspects of quantum cascade heterostructure with a broadband gain. In: ITQW 2017 Conference Proceedings No. 190, TU Wien, Vienna, 2015.
Darmo, J., Kröll, J., Unterrainer, K., 2006. Theoretical aspects of time-domain spectroscopy applied to semiconductor terahertz gain medium. In: AIP Conference Proceeding No. 893AIP, New York, 2007, pp. 515–518.
Darmo, J., Martl, M., Dietze, D., Strasser, G., Unterrainer, K., 2010b. Key components of terahertz remote sensing system. IEICE Technical Report ED2010-172, pp. 77 – 81.
Dhillon, S.S., Sawallich, S., Jukam, N., *et al.*, 2010. Integrated terahertz pulse generation and amplification in quantum cascade lasers. Applied Physics Letters 96, 061107.
Dietze, D., Darmo, J., Unterrainer, K., 2012. THz-driven nonlinear intersubband dynamics in quantum wells. Optics Express 20, 23053.
Dietze, D., Darmo, J., Unterrainer, K., 2013. Efficient population transfer in modulation doped single quantum wells by intense few-cycle terahertz pulses. New Journal of Physics 15, 065014.
Dynes, J.F., Frogley, M.D., Beck, M., Faist, J., Phillips, C.C., 2005. A stark splitting and quantum interference with intersubband transitions in quantum wells. Physical Review Letters 94, 157403.
Faist, J., Capasso, F., Sivco, D.J., *et al.*, 1994. Quantum cascade laser. Science 264, 553–556.
Ferreira, R., Bastard, G., 1989. Evaluation of some scattering times for electrons in unbiased and biased single- and multiple-quantum-well structures. Physical Review B 40, 1074.
Freeman, J.R., Brewer, A., Madeo, J., *et al.*, 2011. Broad gain in a bound-to-continuum quantum cascade laser with heterogeneous active region. Applied Physics Letters 99, 241108.
Freeman, J.R., Maysonnave, J., Khanna, S., *et al.*, 2013. Laser seeding dynamics with few-cycle pulses. Physical Review A 87, 063817.
Golde, D., Wagner, M., Stehr, D., *et al.*, 2009. Fano signatures in the intersubband terahertz response of optically excited semiconductor quantum wells. Physical Review Letters 102, 127403.
Günter, G., Anappara, A.A., Hees, J., *et al.*, 2009. Sub-cycle switch-on of ultrastrong light-matter interaction. Nature 458, 178.
Heyman, J., Kersting, R., Unterrainer, K., 1998. Time-domain measurement of intersubband oscillations in a quantum well. Applied Physics Letters 72, 644.
Huber, R., Brodschelm, A., Tauser, F., Leitenstorfer, A., 2000. Generation and field-resolved detection of femtosecond electromagnetic pulses tunable up to 41 THz. Applied Physics Letters 76, 3191.
Jukam, N., Dhillon, S.S., Osutinov, D., *et al.*, 2009. Terahertz amplifier based on gain switching in a quantum cascade laser. Nature Photonics 3, 715–719.
Keiber, S., Sederberg, S., Schwarz, A., *et al.*, 2016. Electro-optic sampling of near-infrared waveforms. Nature Photonics 10, 159–162.
Kersting, R., Bratschitsch, R., Strasser, G., Unterrainer, K., Heyman, J., 2000. Sampling a THz dipole transition with sub-cycle time-resolution. Optics Letters 25, 272.
Köhler, R., Tredicucci, A., Belham, F., *et al.*, 2002. Terahertz semiconductor-heterostructure laser. Nature 417, 156–159.
Kröll, J., Darmo, J., Dhillon, S.S., *et al.*, 2007. Phase resolved measurements of stimulated emission in a laser. Nature 449, 698–702.
Luo, C.W., Reimann, K., Woerner, M., *et al.*, 2004. Phase-resolved nonlinear response of a two-dimensional electron gas under femtosecond intersubband excitation. Physical Review Letters 92, 047402.
Müller, T., Parz, W., Strasser, G., Unterrainer, K., 2004. Influence of carrier–carrier interaction on the time-dependent intersubband absorption in a semiconductor quantum well. Phys. Rev. B 70, 155324.
Oustinov, D., Jukam, N., Rungsawang, R., *et al.*, 2010. Phase seeding of a terahertz quantum cascade laser. Nature Communications 1, 69–75.
Parz, W., Müller, T., Darmo, J., *et al.*, 2008. Ultrafast probing of light-matter interaction in a mid-infrared quantum cascade laser. Applied Physics Letters 93, 091105.
Rösch, M., Scalari, G., Beck, M., Faist, J., 2015. Octave-spanning semiconductor laser. Nature Photonics 9, 42–47.
Smet, J.H., Fonstad, C.G., Hu, Q., 1996. Intrawell and interwell intersubband transitions in multiple quantum wells for far-infrared sources. Journal of Applied Physics 79, 9305.
Smith, R.M., Arnold, M.A., 2011. Terahertz time-domain spectroscopy of solid samples: principles, applications, and Challenges. Applied Spectroscopy Reviews 46, 636–679.
Turčinková, D., Scalati, G., Castellano, F., *et al.*, 2011. Ultra-broadband heterogeneous terahertz quantum cascade laser emitting from 2.2 to 3.2 THz. Applied Physics Letters 99, 191104.
Villares, G., Riedi, S., Wolf, J., *et al.*, 2016. Dispersion engineering of quantum cascade laser frequency comb. Optica 3, 252–258.
Wu, Q., Litz, M., Zhang, X.-C., 1996. Broadband detection capability of ZnTe electro-optic field detectors. Applied Physics Letters 68, 2924–2926.
Zaks, B., Stehr, D., Truong, T.-A., *et al.*, 2011. THz-driven quantum wells: coulomb interactions and Stark shifts in the ultrastrong coupling regime. New Journal of Physics 13, 083009.
Zaluzny, M., 1993. Influence of the depolarization effect on the nonlinear intersubband absorption spectra of quantum wells. Physical Review B 47, 3995.

Further Reading

Shtrichman, I., Metzner, C., Ehrenfreund, E., *et al.*, 2001. Depolarization shift of the intersubband resonance in a quantum well with an electron-hole plasma. Physical Review B 65, 035310.
Wagner, M., Schneider, H., Stehr, D., *et al.*, 2010. Observation of the intraexcitonic Autler-Townes effect in GaAs/AlGaAs semiconductor quantum wells. Physical Review Letters 105, 167401.

Strong-Field Terahertz Excitations in Semiconductors

Ulrich Huttner, Philipps University of Marburg, Marburg, Germany
Rupert Huber, University of Regensburg, Regensburg, Germany
Mackillo Kira, University of Michigan, Ann Arbor, MI, United States
Stephan W Koch, Philipps University of Marburg, Marburg, Germany

Introduction

Under resonant excitation conditions, already weak electromagnetic fields can induce optical transitions in matter systems. Here, the exciting photon energy matches the energetic separation of two electronic states that have a finite dipole transition matrix element. An example is the hydrogen atom, described by the atomic Coulomb potential schematically shown in **Fig. 1(a)**. Its electronic eigenstates are defined by the Rydberg series with a ground-state energy of $R_y = -13.6$ eV. A light field can only create a transition of an electron from the ground state $n=1$ to a higher state $n=2$ if its photon energy $\hbar\omega_1$ matches the energetic separation of both energy levels. Less energetic light $\hbar\omega_2 < \hbar\omega_1$ cannot be absorbed by the unexcited atom. At the same time, the energetic separation of the electronic states defines the frequency spectrum of the light which can be emitted by the atom, producing, for example, the Balmer series in hydrogen.

The presence of strong electromagnetic fields changes this picture. Such a strong external electric field acts as an electrostatic bias, effectively modifying the atomic potential. This is demonstrated in **Fig. 1(b)**, where an external field with a field strength of 100 MV/cm tilts the hydrogen potential. As a result, the previously non-resonant excitation with photon energy $\hbar\omega_2$ can create a transition to an unbound state in the continuum. Consequently, also non-resonant excitations become possible in the strong-field regime (Boyd, 2008). The required strong electro-magnetic bias can be realized dynamically by the electric field of a light pulse created by a strong laser.

In that case, the oscillating electric field will ionize the atom at peak electric fields and subsequently accelerate the electron in the continuum. Upon a change of the field's polarity, the electron will be accelerated back to the ion such that it can recombine under emission of a high-energy photon. Under these conditions, the emitted radiation contains resonances at integer multiples of the exciting field; this effect is known as *high-harmonic generation* (HHG). Atomic HHG is often explained by the above three-step model (Corkum, 1993). It is routinely used in the creation of ultrashort light sources with pulse durations in the attosecond regime, up to X-ray photon energies (Krausz, 2016). Furthermore, it has been demonstrated that HHG may be used as a spectroscopic tool to study the underlying matter excitations. Recent experiments have applied the same type of strong-field excitations to solids, realizing HHG also in solid-state systems like insulators or semiconductors (Ghimire *et al.*, 2011; Schubert *et al.*, 2014; Luu *et al.*, 2015; Vampa *et al.*, 2015).

Principles

An ideal crystalline semiconductor consists of a periodic arrangement of atoms. In such a periodic lattice, the individual energy levels of the atoms merge into the bandstructure $\varepsilon_{\mathbf{k}}^{\lambda}$ which defines the possible values for the kinetic energy of an electron with crystal momentum $\hbar\mathbf{k}$. Generally, multiple bands λ exist, where the energetically uppermost completely filled band is labelled *valence band* and the energetically lowest fully unoccupied band is the *conduction band*. In semiconductors an energetically forbidden region exists between these bands. Their minimal separation defines the bandgap energy, which is typically in the range of 1–2 eV, corresponding to the photon energy of optical fields (Haug and Koch, 2009). Therefore, the non-resonant excitations

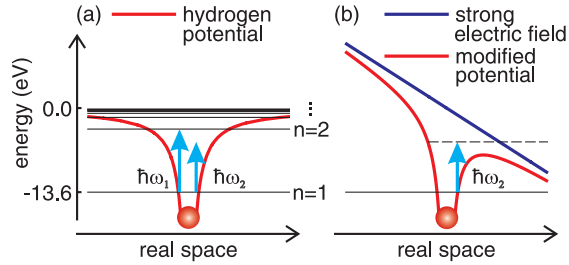

Fig. 1 Resonant vs. non-resonant excitations: (a) Potential energy of a hydrogen atom (red line) together with its electronic eigenstates marked by black lines. A transition of an electron from one energetic state to another can only be realized by a resonant excitation, where the photon energy $\hbar\omega_1$ matches the energetic separation of both states and (b) the presence of a strong electrostatic potential (blue line) creates a modified atomic potential (red line), such that a previously non-resonant excitation with photon energy $\hbar\omega_2$ can create a transition to an unbound state (dashed line).

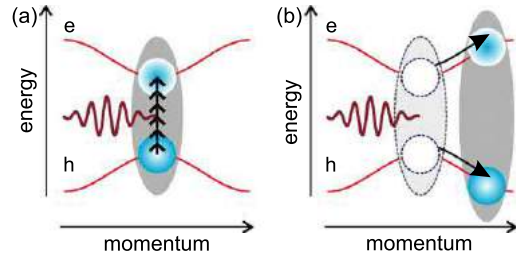

Fig. 2 Fundamental excitations in a semiconductor: (a) Interband excitations induced by a strong THz field (red pulse) lead to the creation of polarization (shaded area) and charge carriers (spheres) also for very non-resonant excitations, where the THz photon energy (black arrows) is much smaller than the bandgap between valence- and conduction band (red lines) and (b) the THz field accelerates charge carriers as well as polarization inside the bands creating intraband currents (black arrows).

needed for HHG can be realized with strong terahertz (THz) or near infrared fields with photon energies of several tens of meV. Since a few years, such strong fields can be experimentally realized (Schubert *et al.*, 2014).

An optical excitation of a semiconductor induces a transition amplitude between a state in the completely filled valence band and the empty conduction band. This microscopic polarization will eventually lead to the creation of charge carriers for strong enough fields, even for very non-resonant excitations (Haug and Koch, 2009). These interband excitations are schematically depicted in **Fig. 2(a)**, where the THz photon energy (black arrows) is much smaller than the bandgap between valence and conduction band (red lines). Simultaneously, the THz field accelerates charge carriers inside the bands as well as the polarization. This leads to the creation of intraband currents, symbolized in **Fig. 2(b)** by the black arrows. The interplay of interband- and intraband excitations, described by the microscopic polarization and carrier distributions, leads to the emission of high-harmonic radiation (Golde *et al.*, 2011). As HHG is a fully coherent process, the emitted radiation not only depends on the intensity of the excitation, but also its phase (Schubert *et al.*, 2014).

Semiconductor-Bloch Equations

A microscopic quantum many-body theory of HHG is provided by the semiconductor Bloch equations (SBEs) (Haug and Koch, 2009; Kira and Koch, 2006, 2012). The SBEs describe electrons in a periodic lattice, which interact via the Coulomb interaction and with an external light field. In addition, also the coupling to lattice vibrations or other external potentials can be included. The solution of the SBEs yields the time dynamics of the microscopic polarizations and the carrier distributions in their respective bands for arbitrary excitations. As the strong field excitations applied for HHG can couple multiple bands, the SBEs are schematically presented below for a multiband situation with an arbitrary number of bands λ. The polarization between a valence band h and a conduction band e is then given by

$$i\hbar\frac{\partial}{\partial t}p_{\mathbf{k}}^{\mathrm{he}}=\left(\varepsilon_{\mathbf{k}}^{\mathrm{he}}+i|e|E(t)\nabla_{\mathbf{k}}\right)p_{\mathbf{k}}^{\mathrm{he}}-\hbar\Omega_{\mathbf{k}}^{\mathrm{eh}}\left(1-f_{\mathbf{k}}^{\mathrm{e}}-f_{\mathbf{k}}^{\mathrm{h}}\right)+\sum_{\lambda\neq e,h}\left(\hbar\Omega_{\mathbf{k}}^{\lambda h}p_{\mathbf{k}}^{\lambda e}-\hbar\Omega_{\mathbf{k}}^{e\lambda}p_{\mathbf{k}}^{h\lambda}\right)+\Gamma_{\mathbf{k}}^{\mathrm{he}} \tag{1}$$

in the electron–hole picture (Haug and Koch, 2009; Schubert *et al.*, 2014). In this description, a missing electron, i.e., a vacancy in the valence band is treated as a separate quasiparticle, a *hole*. The microscopic polarization $p_{\mathbf{k}}^{\mathrm{he}}$ depends on the energetic separation $\varepsilon_{\mathbf{k}}^{\mathrm{he}}$ between bands e and h. It is driven by the electric field via the Rabi frequency $\Omega_{\mathbf{k}}^{\mathrm{eh}}$, which contains a product of the electric field $E(t)$ with the dipole matrix element $d_{\mathbf{k}}^{\lambda\lambda'}$ between the two bands $\lambda=$ e and $\lambda'=$h. Additionally, $p_{\mathbf{k}}^{\mathrm{he}}$ is coupled to all other microscopic polarizations involving either of the two bands e or h via the Rabi frequency. Both, $\varepsilon_{\mathbf{k}}^{\mathrm{he}}$ and $\hbar\Omega_{\mathbf{k}}^{\mathrm{eh}}$ are renormalized by the Coulomb interaction between the charge carriers. The term containing the gradient describes the acceleration of the polarization by the THz field. Higher-order correlations, including interactions between more than two particles, which are created by the Coulomb interaction are included in $\Gamma_{\mathbf{k}}^{\mathrm{he}}$ (Kira and Koch, 2006). In HHG $\Gamma_{\mathbf{k}}^{\mathrm{he}}$ is dominated by scattering terms resulting in a dephasing of the electronic excitations (Golde *et al.*, 2011). The carrier occupation of the conduction band e is determined by

$$\hbar\frac{\partial}{\partial t}f_{\mathbf{k}}^{\mathrm{e}}=-2\,\mathrm{Im}[\sum_{\lambda\neq\mathrm{e}}\hbar\Omega_{\mathbf{k}}^{e\lambda}p_{\mathbf{k}}^{e\lambda}]+|e|E(t)\nabla_{\mathbf{k}}f_{\mathbf{k}}^{\mathrm{e}}+\Gamma_{\mathbf{k}}^{\mathrm{e}} \tag{2}$$

It is driven by the microscopic polarizations between the band e and all other bands, which are connected by dipoles. Similar to the microscopic polarizations, the acceleration of the carrier distributions by the THz field is described by the gradient term, while $\Gamma_{\mathbf{k}}^{\mathrm{e}}$ contains higher-order correlations. Importantly, Eqs. (1) and (2) are mutually coupled, such that the SBEs form a set of coupled differential equations. This coupling leads to a strong intermixing of interband excitations and intraband currents, both contributing to the high-harmonic emission (Golde *et al.*, 2011). In detail, the emitted radiation has a polarization source $P(t)$, which is defined by a summation over all momentum states of the microscopic polarization between all dipole coupled bands and a current source $J(t)$, which is effectively defined by the carrier distributions. The emitted high-harmonic field $E_{\mathrm{HHG}}(t)$ is then obtained from $E_{\mathrm{HHG}}(t)\propto\frac{\partial}{\partial t}P(t)+J(t)$ in the time domain. Its Fourier transform yields the corresponding frequency spectrum $I_{\mathrm{HHG}}(\omega)\propto|\omega P(\omega)+\mathrm{i}J(\omega)|^2$.

Examples

High-Harmonic Generation in Two Versus Three Bands

In order to characterize the basic properties of high harmonic frequency spectra and to demonstrate the influence of the number of involved electronic bands, $I_{\mathrm{HHG}}(\omega)$ is compared for a two-band system and a three-band system in **Fig. 3**.

In detail, the SBEs are evaluated for both bandstructure configurations shown in the inset using an identical THz excitation with central frequency of $f_{\mathrm{THz}}=30$ THz. Using two electronic bands, $I_{\mathrm{HHG}}(\omega)$ in **Fig. 3** shows narrow and well-separated resonances at odd multiples of f_{THz}, where the intensity of the harmonics rapidly decreases for higher harmonic orders. Such a behavior is known from atomic HHG, where the inversion symmetry prevents the emission of even harmonic orders (Lewenstein *et al.*, 1994). In a three-band calculation, however, the $I_{\mathrm{HHG}}(\omega)$ spectrum contains resonances at frequencies corresponding to even and odd harmonic orders. Furthermore, a plateau of harmonic orders with comparable intensity is visible. The appearance of a plateau is a typical feature of high-harmonic spectra, which is well-known from atomic HHG and a hallmark of a non-perturbative process. In the non-perturbative regime, which can only be reached by strong excitations, the intensity I_n of the harmonic order n does no longer obey the perturbative scaling law $I_n \sim E^{2n}$.

Non-Perturbative Quantum Interference

The fundamentally different emission characteristics of a two-band configuration in contrast to a system involving at least three bands can be explained by analyzing the microscopic excitation paths. In a two-band system consisting of one conduction band (e) and one valence band (h), as displayed in **Fig. 4(a)**, an excitation between both bands can only proceed directly from one band to another, as symbolized by the gray arrow. Such transitions are denoted as *direct* excitation paths.

The strength of this transition and its contribution to the high-harmonic emission is determined by the polarization $P_{\mathrm{dir}}(t)$ between both bands. Due to the mutual coupling of microscopic polarizations and carrier occupations, $P_{\mathrm{dir}}(t)$ contains a sequence of terms, which depend only on odd powers of the electric field. Carrying out a Fourier transform in order to compute the corresponding emission spectrum will, thus, produce resonances at odd multiples of its driving frequency. Consequently, $I_{\mathrm{HHG}}(\omega)$ computed with two electronic bands shows only odd harmonic orders in **Fig. 3**. In the presence of strong, non-perturbative excitations, this field dependence of $P_{\mathrm{dir}}(t)$ is then modified into a nonlinear function P_{odd} with odd parity with respect to the driving field $E(t)$, i.e., $P_{\mathrm{odd}}(-E) = -P_{\mathrm{odd}}(E)$.

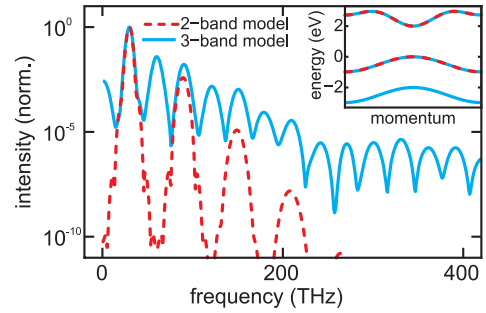

Fig. 3 High harmonic emission spectra: Computed $I_{\mathrm{HHG}}(\omega)$ spectra of a two-band system (red-dahsed line) and a three-band system (blue-solid line), which are excited with strong THz light with central frequency $f_{\mathrm{THz}}=30$ THz. The band dispersion used in both calculations is shown in the inset in corresponding colors. Reproduced from Huttner, U., Schuh, K., Moloney, J.V., Koch, S.W., 2016. Similarities and differences between high-harmonic generation in atoms and solids. J. Opt. Soc. Am. B 33, C22–C29.

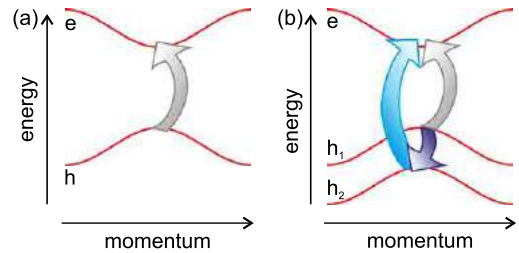

Fig. 4 Direct and indirect transition paths: (a) In a two-band system consisting of one conduction band e and one valence band h, an excitation can only proceed directly from one band to another, symbolized by the gray arrow and (b) considering two valence bands h_1 and h_2, an excitation from one valence band to the conduction band e may either proceed directly (gray arrow) or via the other valence band forming an indirect transition path, symbolized by the combination of both blue arrows.

In a three-band system, as shown in Fig. 4(b), excitations from one band to another may either proceed directly (symbolized by the gray arrow), or via an intermediate step including the third band (blue arrows). These *indirect* transitions are generated by the coupling of a direct transition from the lower valence band to the conduction band to a transition between the valence bands, mediated by the electric field. As a result, the corresponding polarization $P_{\mathrm{indir}}(t)$ consists of terms which depend only on even powers of the electric field. Therefore, $P_{\mathrm{indir}}(t)$ creates the even harmonic orders in the $I_{\mathrm{HHG}}(\omega)$ spectrum computed with three electronic bands. In analogy to the direct transitions, non-perturbative excitations modify this field dependence of $P_{\mathrm{indir}}(t)$ into a nonlinear function P_{even} with even parity, i.e., $P_{\mathrm{even}}(-E)=P_{\mathrm{even}}(E)$.

The total transition yield P_{tot} is defined by an interference of both transition paths according to

$$P_{\mathrm{tot}}(E)=P_{\mathrm{even}}(E)+P_{\mathrm{odd}}(E). \tag{3}$$

This quantum interference is also present in the non-perturbative regime creating even and odd harmonic orders with comparable strength (Hohenleutner *et al.*, 2015).

Time-Resolved HHG

The interference of different excitation paths becomes most prominent in time-resolved studies (Hohenleutner *et al.*, 2015). Therefore, Fig. 5 shows intensity envelopes $I_{\mathrm{HHG}}(t)$ of computed $E_{\mathrm{HHG}}(t)$ time traces. They are calculated for a strong THz excitation, $E_{\mathrm{THz}}(t)$ using two and three bands, respectively. For two electronic bands, $I_{\mathrm{HHG}}(t)$ shows intensity modulations on a time scale which corresponds to the duration of a THz half-cycle. Furthermore, the emission maxima occur delayed with respect to the field crests of $E_{\mathrm{THz}}(t)$.

In contrast to that, the three-band $I_{\mathrm{HHG}}(t)$ yields high-harmonic radiation as a series of ultrashort bursts, which appear only at positive field cycles of $E_{\mathrm{THz}}(t)$. The emission at negative field cycles is strongly suppressed. Furthermore, the emission bursts are synchronized to the crests of the driving field. Non-perturbative quantum interference can explain this observation. In the two-band computation, the high-harmonic emission is exclusively defined by the direct paths producing the polarization $P_{\mathrm{dir}}(E_{\mathrm{THz}})$. The HHG emission intensity, thus, has only a single source term and is independent of the sign of $E_{\mathrm{THz}}(t)$. In the three-band computation, however, direct and indirect paths contribute to the emission according to Eq. (3) via the sum of their amplitudes $P_{\mathrm{indir}}(E_{\mathrm{THz}})+P_{\mathrm{dir}}(E_{\mathrm{THz}})$. For positive half cycles of $E_{\mathrm{THz}}(t)$ they interfere constructively. For negative half cycles of $E_{\mathrm{THz}}(t)$, in contrast, a sign flip of the electric field $E_{\mathrm{THz}}\rightarrow -E_{\mathrm{THz}}$ switches from constructive interference to destructive interference $P_{\mathrm{indir}}(-E_{\mathrm{THz}})+P_{\mathrm{dir}}(-E_{\mathrm{THz}})=P_{\mathrm{indir}}(E_{\mathrm{THz}})-P_{\mathrm{dir}}(E_{\mathrm{THz}})$, where the contributions of direct and indirect paths cancel. Consequently, these half cycles lead to strongly suppressed emission. Additionally, it has been shown, that non-perturbative excitations balance the amplitudes of direct and indirect excitation paths producing efficient interference conditions (Hohenleutner *et al.*, 2015).

New, sophisticated experimental techniques based on frequency-resolved optical gating, allow for the monitoring of the high-harmonic emission $I_{\mathrm{HHG}}(t)$ on the same timescale as the exciting THz waveform (Hohenleutner *et al.*, 2015). Fig. 6(a) shows such a measurement in GaSe in comparison to computations performed using five electronic bands in Fig. 6(b). Both, experiment and theory show an identical unipolar high-harmonic emission, as well as a synchronization of the emission bursts to the positive field crests of the driving field.

Dynamical Bloch Oscillations

When a strong field bias is applied to a semiconductor via a static electric field, charge carriers will be accelerated inside the bands leading to an increase of their momentum $\hbar\mathbf{k}$. If the field is strong enough, it can accelerate an electron wave packet up to the border of the first Brillouin zone (BZ) and beyond. Upon leaving the first BZ on one side, the wave packet re-enters at the opposite side of the BZ with sign-flipped momentum. This corresponds to oscillations of the electron wave packet in real and momentum space. These oscillations are referred to as *Bloch oscillations*. Remarkably, they are created by a constant electric field bias, however, they are hard to observe experimentally due to the fast scattering of the accelerated electrons.

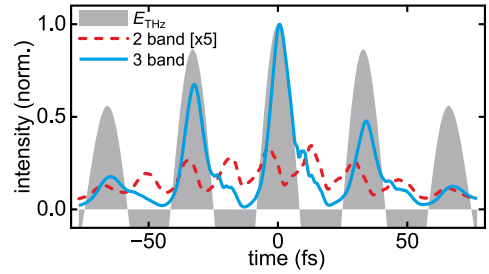

Fig. 5 Time-resolved high-harmonic emission. (a) High-harmonic emission intensity $I_{\mathrm{HHG}}(t)$ as function of time computed for a two-band system (red dashed line) and a three-band system (blue solid line), driven by a strong THz field (shaded area). The emission intensity of the two-band system is magnified by a factor of five. Reproduced from Huttner, U., Schuh, K., Moloney, J.V., Koch, S.W., 2016. Similarities and differences between high-harmonic generation in atoms and solids. J. Opt. Soc. Am. B 33, C22–C29.

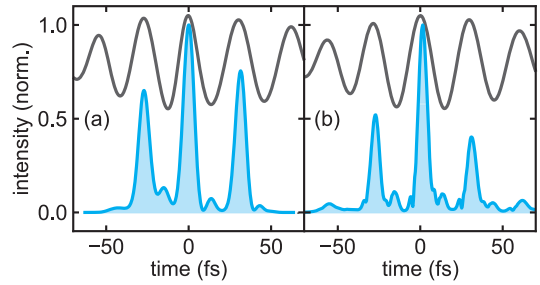

Fig. 6 Theory-experiment comparison of time-resolved high-hamonic emission in GaSe: (a) High-harmonic emission intensity $I_{HHG}(t)$ (blue) as function of time, measured in GaSe for a THz peak field (black line) of 44 MV/cm with central frequency of 33 THz and (b) corresponding, computed high-harmonic emission using five electronic bands. Adapted from Hohenleutner, M., Langer, F., Schubert, O., *et al.*, 2015. Real-time observation of interfering crystal electrons in high-harmonic generation. Nature 523, 572–575.

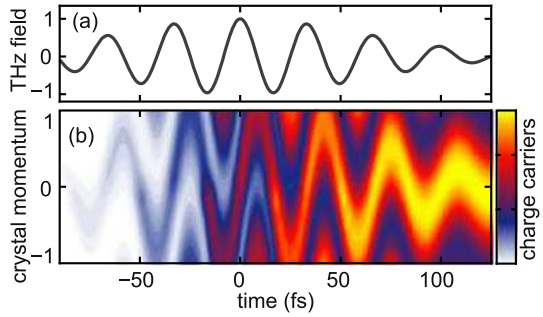

Fig. 7 Dynamical Bloch oscillations: (b) Computed electron wave-packet dynamics in the lowest conduction band of GaSe as function of time and crystal momentum, excited by a strong THz pulse (a) with duration of 100 fs. Adapted from Schubert, O., Hohenleutner, M., Langer, F., *et al.*, 2014. Sub-cycle control of terahertz high-harmonic generation by dynamical Bloch oscillations. Nat. Photon. 8, 119–123.

The situation is different for transient excitation conditions with strong THz fields which can lead to a dynamical version of the Bloch oscillations (Schubert *et al.*, 2014). These oscillations can be visualized in simulations, for example, by monitoring the time-dependent microscopic carrier distribution in a THz excited semiconductor. As an illustration, we show in **Fig. 7** the distribution of electrons in the lowest conduction band of GaSe as function of time and crystal momentum, after the excitation of the system by an ultrashort, strong THz waveform depicted in **Fig. 7(a)**. During one half cycle of the THz excitation, the electronic wave packet is accelerated throughout the whole BZ, leaves the BZ on one side, and enters at the other side. Because they create an oscillatory motion of electrons in a non-parabolic bandstructure, these dynamic Bloch oscillations strongly contribute to the emission of high-harmonic radiation (Schubert *et al.*, 2014).

Electron–Hole Recollisions: High-Order Sideband Generation

It is interesting to extend the strong THz excitation scheme by applying an additional resonant light pulse. For example, this light pulse can resonantly create coherent excitons, i.e., an electron-hole (e–h) polarization oscillating with the energy of the excited exciton resonance. Once created, the THz field will then accelerate electrons and holes in opposite directions due to their opposite charge. In a similar fashion as described by the three-step model of atomic HHG, electrons and holes may eventually recollide and recombine under photon emission.

This process of e–h recollisions is the driving mechanism of high-order sideband generation (HSG) (Zaks *et al.*, 2012; Langer *et al.*, 2016). The HSG emission spectrum produces a series of resonances at both sides of the frequency of the optical excitation – called sidebands, which are separated by twice the THz central frequency.

However, in contrast to atomic HHG, the time separation between optical and THz pulse can precisely be adjusted and controlled. Using an ultrashort optical excitation, this time delay defines and controls at which phase of the THz field E_{THz}, the coherent excitons are created. Consequently, in analogy to atomic recolission models, this creation time determines the degree to which a recollision between electrons and holes is possible within the ultrashort coherence time of the excitonic excitations. As a result, *good* and *bad* excitation times exist, entailing strong or weak sideband emission, respectively. The quasiparticle dynamics for both excitation times are schematically illustrated by the cartoon in **Fig. 8**.

At bad excitation times (**Fig. 8(a)**), coherent excitons are created shortly after a zero-crossing of the THz bias. Consequently, the THz fields separates electrons and holes, such that no recollision occurs. At good excitation times (**Fig. 8(b)**), however, coherent

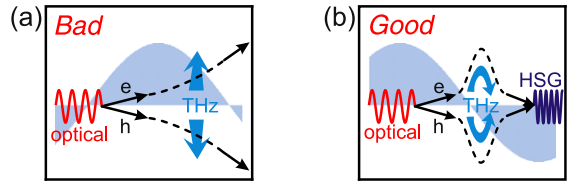

Fig. 8 Schematics of e–h recollisions for good and bad excitation times: An optical excitation (red line) creates coherent e–h pairs, which are subsequently separated and accelerated by a strong THz bias (blue area). For bad excitation times (a), electrons and holes do not recollide within the finite and ultrashort lifetime of the coherent excitons. For good excitation times (b), in contrast, e–h recollisions lead to strong sideband emission (dark blue line). Adapted from Langer, F., Hohenleutner, M.C., Schmid, P., *et al.*, 2016. Lightwave-driven quasiparticle collisions on a subcycle timescale. Nature 533, 225–229.

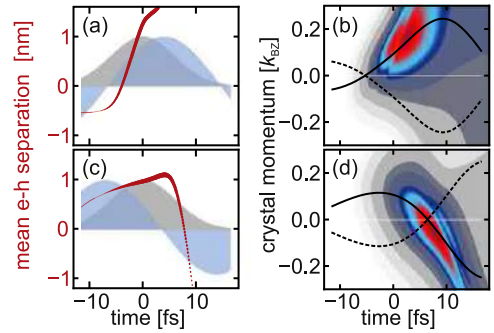

Fig. 9 Microscopic quasi-particle dynamics: (a), (c) Mean e–h separation obtained by evaluating the e–h pair-correlation function for bad (a) and good (c) excitation times. Envelope of the optical excitation (shaded area) and THz waveform (blue area). (b), (d) Electron dynamics in the conduction band as function of time and crystal momentum for bad (b) and good (d) excitation times. Mean electron (solid line) and hole (dashed line) momentum as function of time. Adapted from Langer, F., Hohenleutner, M.C., Schmid, P., *et al.*, 2016. Lightwave-driven quasiparticle collisions on a subcycle timescale. Nature 533, 225–229

excitons are created shortly after a field crest of E_{THz}, such that electrons and holes are accelerated and recollided, producing strong sideband emission.

The underlying microscopic quasi-particle dynamics can be monitored in real space, as well as in momentum space. Therefore, **Fig. 9(a)** and **(c)** shows the mean relative distance between an electron and the corresponding hole (red line) in real space for (a) bad and (c) good excitation times. The relative e–h separation $\langle r \rangle$ is defined by means of the e–h pair-correlation function (Kira and Koch, 2006). While $\langle r \rangle$ grows monotonically for bad excitation times, effectively suppressing recollisions, $\langle r \rangle$ is rapidly crossing $r=0$ after an initial increase for good excitation times. Around $r=0$ the electron and hole wavepackets have maximum overlap enabling an efficient recombination, which leads to strong HSG emission, confirming the interpretation of **Fig. 8**.

A similar picture arises in momentum space by monitoring the electron dynamics in the conduction band in **Fig. 9(b)** and **(d)**. For bad excitation times (b), the electron wavepacket is moved away from the center of the BZ, where e–h recombination is most efficient, while the hole wavepacket is moving in the opposite direction. For good excitation times (d), electron and hole wave packets, which are initially separated from each other, are pushed back towards the BZ center enabling recollisions and producing strong HSG emission.

Consequently, the emitted sideband radiation contains details about the microscopic quasiparticle dynamics and information about the colliding quasiparticles themselves. Monitoring the intensity of the emitted sideband radiation as function of the delay time allows, for example, for an estimate of the excitonic coherence time (Langer *et al.*, 2016).

See also: Coherent Terahertz Sources. Multidimensional Terahertz Spectroscopy. Terahertz Physics of Semiconductor Heterostructures

References

Boyd, R.W., 2008. Nonlinear Optics, third ed. Academic press.
Corkum, P.B., 1993. Plasma perspective on strong field multiphoton ionization. Physical Review Letters 71, 1994.
Ghimire, S., DiChiara, A.D., Sistrunk, E., *et al.*, 2011. Observation of high-order harmonic generation in a bulk crystal. Nature Physics 7, 138.
Golde, D., Kira, M., Meier, T., Koch, S.W., 2011. Microscopic theory of the extremely nonlinear terahertz response of semiconductors. Physica Status Solidi (b) 248, 863–866.
Haug, H., Koch, S.W., 2009. Quantum Theory of the Optical and Electronic Properties of Semiconductors. World Scientific.
Hohenleutner, M., Langer, F., Schubert, O., *et al.*, 2015. Real-time observation of interfering crystal electrons in high-harmonic generation. Nature 523, 572.

Kira, M., Koch, S., 2006. Many-body correlations and excitonic effects in semiconductor spectroscopy. Progress in Quantum Electronics 30, 155.
Kira, M., Koch, S.W., 2012. Semiconductor Quantum Optics. Cambridge: Cambridge University Press.
Krausz, F., 2016. The birth of attosecond physics and its coming of age. Physica Scripta 91, 063011.
Langer, F., Hohenleutner, M., Schmid, C.P., *et al.*, 2016. Lightwave-driven quasiparticle collisions on a subcycle timescale. Nature 533, 225.
Lewenstein, M., Balcou, P., Ivanov, M.Y., LHuillier, A., Corkum, P.B., 1994. Theory of high-harmonic generation by low-frequency laser fields. Physical Review A 49, 2117.
Luu, T.T., Garg, M., Kruchinin, S.Y., *et al.*, 2015. Extreme ultraviolet high-harmonic spectroscopy of solids. Nature 521, 498.
Schubert, O., Hohenleutner, M., Langer, F., *et al.*, 2014. Sub-cycle control of terahertz high-harmonic generation by dynamical Bloch oscillations. Nature Photonics 8, 119.
Vampa, G., Hammond, T.J., Thire, N., *et al.*, 2015. Linking high harmonics from gases and solids. Nature 522, 462.
Zaks, B., Liu, R.B., Sherwin, M.S., 2012. Experimental observation of electron–hole recollisions. Nature 483, 580.

Rydberg States in Semiconductors

Manfred Bayer and Marc Assmann, Technical University of Dortmund, Dortmund, Germany

Introduction

In atomic physics, Rydberg atoms are highly excited atoms with at least one valence electron excited to a state with huge principal quantum number n. In the limit of a simple hydrogen-like system, a classical Bohr-like description of the electron is a good approximation of the system. Some peculiar properties of the system can be derived from this correspondence. For circular motion of an electron around a nucleus carrying charge e, the arising Coulomb force acts as a centripetal force on the electron:

$$\frac{e^2}{4\pi\varepsilon_0 r^2} = \frac{mv^2}{r} \tag{1}$$

In Bohr-like models angular momentum is quantized, so

$$mvr = n\hbar \tag{2}$$

which yields a direct relation between the electron orbit radius and the principal quantum number:

$$r = \frac{4\pi\varepsilon_0 n^2\hbar^2}{me^2} \tag{3}$$

So the radius of an orbit scales quadratically with the principal quantum number, which means that spatial extensions in the micrometer range can be reached for $n \approx 100$. Many other fundamental properties of Rydberg atoms show similar scaling laws with respect to the principal quantum number (Gallagher, 2005). The dipole moment also shows a quadratic scaling, while the radiative lifetime scales as n^3 and the geometric cross section already follows a n^4-dependence. Polarizability even scales as n^7. It is easy to see that such an easily polarizable system, where the largest possible carrier separation is almost equal to the radius of the electron orbit is extremely sensitive to external fields and acts like a miniaturized antenna.

While cold atoms provide an excellent environment for studies of fundamental physics, it is desirable to have a physical system that is more suitable for everyday life when thinking of applications. Semiconductors form the material basis for modern-day electronics and optoelectronics, so one might envisage them as good candidate systems to look for Rydberg physics. The basic excitation of a semiconductor system is the exciton, a pair consisting of a negatively charged conduction band electron and a positively charged valence band hole bound by the attractive Coulomb force. They can be considered as the semiconductor analogues of hydrogen atoms. Indeed highly excited excitons are the most promising candidates for identifying Rydberg physics in a semiconductor setting. However, many characteristic quantities of hydrogen and semiconductor systems differ strongly, which renders the observation of highly excited excitons more complicated. This can already be seen from the binding energy of the states, which is for unperturbed hydrogen-like systems given by $E_n = -\frac{E_{Ryd}}{n^2}$, where the Rydberg energy E_{Ryd} amounts to:

$$E_{Ryd} = \frac{\mu e^4}{32\varepsilon_0^2\varepsilon^2\pi^2\hbar^2} \tag{4}$$

Differences arise due to the presence of the dielectric environment denoted by ε and most importantly due to the difference in effective mass μ. For hydrogen, the proton is about 1836 times heavier than the electron, while electrons and holes have similar mass, which results in the effective mass in atomic and semiconductor systems differing by several orders of magnitude. Therefore, also the Rydberg energy differs strongly. It amounts to 13.6 eV for hydrogen, but is only in the range of meV for typical semiconductors. For most materials such as the prototypical semiconductor GaAs with a Rydberg energy of only 4.2 meV, this means that highly excited states will be spaced so closely that they are not resolvable. Also, the energies of excited states will approach the ionization continuum quickly, so that even at low temperatures the highly excited excitons will be turned into free carriers by thermal activation. On the other hand, many wide band gap materials such as ZnO or GaN might in principle be good candidates for observing Rydberg physics due to the large exciton binding energies in those materials, but they typically suffer from a large number of impurities and inhomogeneous crystals. Monolayers of transition metal dichalcogenides provide another promising candidate system for observing Rydberg physics in a semiconductor setting. They show exciton binding energies of several hundred meV, but currently the linewidths of exciton states in these systems are too large to observe more than the first few excited exciton states.

At the moment, Cuprous Oxide (Cu_2O) shows the most pronounced Rydberg effects in semiconductor systems and the main part of this article will be devoted to Cu_2O. Historically, Cu_2O was the material, in which excitons were observed first by Gross and Karryjew (1952) and Hayashi and Katsuki (1952). It features a moderately large exciton binding energy of about 90 meV and is a direct band gap semiconductor with a band gap of about 2.172 eV. The highest valence and the lowest conduction band are both formed from copper states, namely the 3d and 4s orbitals. Therefore, in contrast to most semiconductor systems, absorption of a photon results in creation of an electron-hole pair at the same atom. The excitons belonging to the transition between these bands are termed the yellow exciton series because of the wavelengths of these transitions are in the yellow range of the electromagnetic spectrum around 590 nm. Other exciton series at higher energies exist and are called the green, blue and violet series, respectively,

 doi:10.1016/B978-0-12-803581-8.09579-5

but we will primarily focus on the yellow series. As both the conduction and the valence band states have the same parity, direct dipole transitions between band states are forbidden. However, for excitons, the relative motion of electron and hole opens up another degree of freedom. The total symmetry of the exciton is thus given by the direct product of the symmetries of the bands and the envelope. Accordingly, excitons with a P-envelope are dipole allowed in optical transitions (Agekyan, 1977), while those with an S-envelope are not.

Rydberg Excitons in Cu_2O

The series of P-excitons has been investigated already in the early days of spectroscopic studies on Cu_2O, where principal quantum numbers up to $n=9$ were identified (Gross and Karryjew, 1952; Gross, 1956). The series could be extended up to $n=12$ over the years (Matsumoto *et al.*, 1996). Going to even higher n results in entering the Rydberg exciton regime, but there are several limiting factors to the highest principal quantum number that can possibly be observed in a sample. Besides temperature, which will result in thermal ionization of states with small binding energy, the sample quality has the biggest influence on the observable exciton states. As the spatial extension of excitons grows with their principal quantum number, impurities, strain or sample imperfections within this spatial region may prevent the formation of large exciton states. While usually in semiconductor physics high quality crystals are created by means of artificial fabrication, artificial Cu_2O crystals are at current still inferior in quality compared to their natural counterparts. Accordingly, the highest principal quantum numbers observed so far for excitons in Cu_2O have been observed in natural crystals cut and polished from a rock mined at the Tsumeb mine in Namibia. The P-exciton absorption spectrum of a 34 μm thick slab mounted free of strain and held at a temperature of 1.2 K is shown in **Fig. 1** (Kazimierczuk *et al.*, 2014). The spectrum shows a large number of lines, so the lower panels show the high-energy parts of the spectrum with increasing resolution. Exciton lines are labeled in terms of their principal quantum number. In total, states up to $n=25$ can be identified. For such large n, it is instructive to estimate the spatial extent of the exciton wavefunction. In analogy to hydrogen, the average radius of an orbital for a given combination of l and n can be calculated as

$$\langle r_{n,l}\rangle = \frac{1}{2}a_B(3n^2 - l(l+1)) \tag{5}$$

where a_B denotes the Bohr radius, which is 1.11 nm for P-excitons in Cu_2O (Kavoulakis and Chang, 1997). For the highest principal quantum number seen in the spectrum one gets $\langle r_{25,1}\rangle = 1.04$ μm, which corresponds to a spatial extension of more than 2 μm. This is equivalent to more than ten times the wavelength of the light in the material or several billion unit cells of the crystal.

A first test, whether Rydberg excitons can be considered akin to Rydberg atoms is to check the n^{-2} scaling law of the binding energies. However, the shape of the exciton resonances shows a characteristic asymmetric Fano-type line shape. This is a consequence of the interference of discrete exciton states with a continuum of states arising due to optical-phonon assisted absorption of the 1S state. The main consequence of this asymmetric line is that the position of the peak maximum does not necessarily coincide with the position of the resonance. The deviation may be larger than 1 meV. Accordingly, it is mandatory to fit the asymmetric line shape given by (Toyozawa, 1964)

$$\alpha_n(E) = C_n \frac{\frac{\Gamma_n}{2} + 2q_n(E - E_n)}{\left(\frac{\Gamma_n}{2}\right)^2 + (E - E_n)^2} \tag{6}$$

to the experimental data in order to extract the real positions of the resonances. Here E_n is the real position of the nth resonance, the amplitude C_n is proportional to the oscillator strength, q_n describes the asymmetry of the line and Γ_n is the spectral width of the resonance. These energies are shown in **Fig. 2**. They match the Rydberg formula $E_n = E_G - \frac{E_{Ryd}}{n^2}$ rather well. As the data shows, the band gap energy E_G and E_{Ryd} can be determined to 2.17208 eV and 90–92 meV, respectively.

A slight deviation from the ideal formula is apparent, but can be explained in terms of a quantum defect as will be discussed in detail later. The linewidths for the different n are shown in **Fig. 3**. For larger n they decrease down to few μeV. Up to about $n=10$, where one can assume that the lines are only homogeneously broadened, the linewidths follow the inverse cubic law expected for Rydberg states. For even larger principal quantum numbers, one encounters some additional broadening that can arise due to crystal imperfections, the influence of the other exciton series or power broadening. Still, the linewidth is as low as 3 μeV for the largest exciton states, which corresponds to lifetimes in the nanosecond range. At first, such long lifetimes might seem surprising, as it might seem that the large spatial extension of these excitons might make it susceptible to scattering. The main recombination pathways besides direct radiative recombination are carrier-carrier scattering, which is negligible at low excitation densities, relaxation into lower exciton states by spontaneous emission of infrared photons, which is unlikely due to the small spontaneous emission rate for low-energy photons and relaxation by emission of an optical phonon, which also shows an inverse cube law scaling as seen in the experiments.

Rydberg Blockade

The final scaling law of interest is the exciton oscillator strength, which is supposed to scale as n^{-3}. The oscillator strengths are proportional to the peak areas of the resonances, which are shown in **Fig. 4** for a laser intensity of 6 μW mm^{-2}.

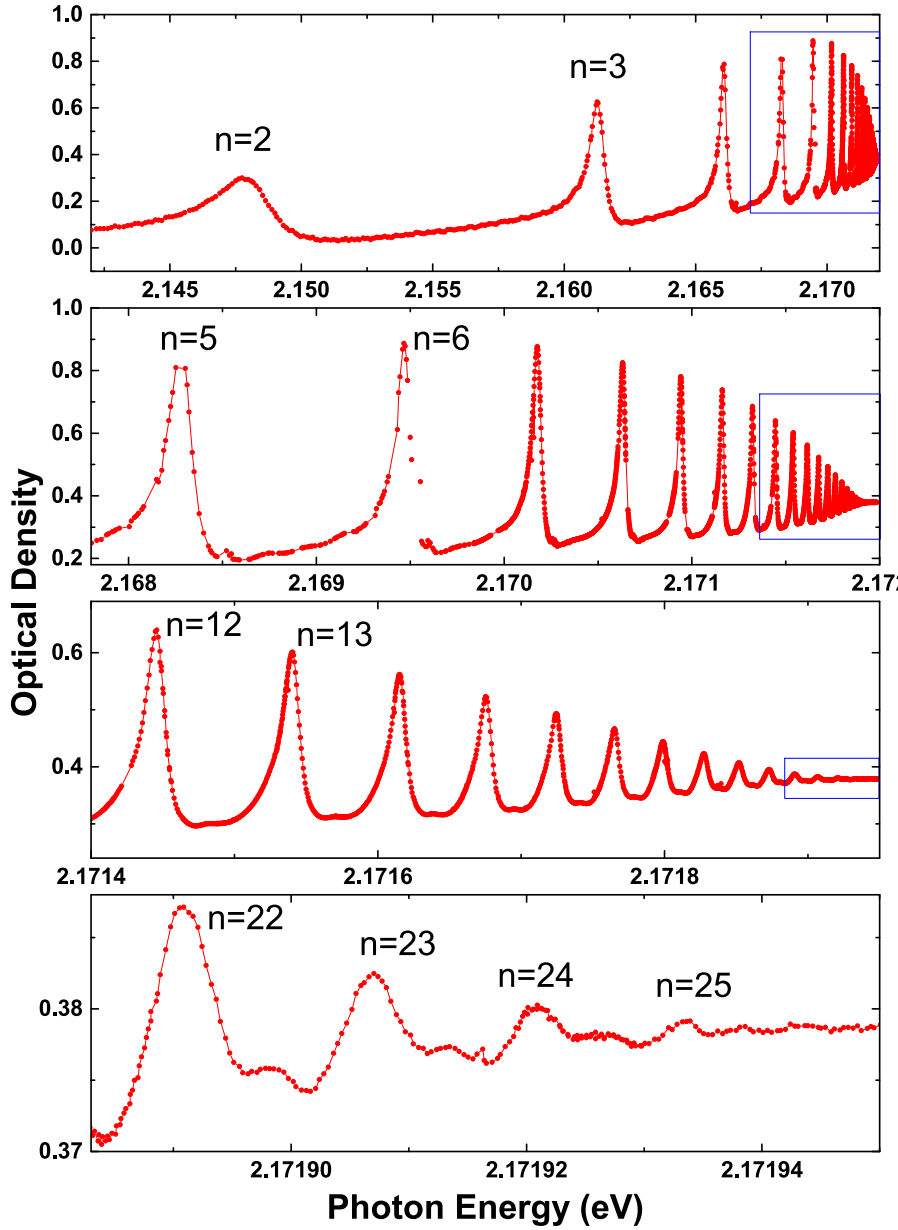

Fig. 1 High-resolution spectrum of the yellow exciton series in Cu_2O. Peaks correspond to P-excitons with principal quantum number n. Blue rectangles mark the region shown in close-up view in the next panel.

The expected inverse cubic law dependence is seen up to $n=17$. Beyond that point, the oscillator strength of larger excitons is significantly reduced. This reduction shows a pronounced intensity dependence as shown in **Fig. 5**. With increased laser power, the resonances start to vanish continuously with the higher-energy states being effected prior to and stronger than the states of lower n. The areas of the resonance peaks are plotted in **Fig. 6** (Kazimierczuk *et al.*, 2014). One can clearly see that the excitation power, at which the reduction of the peak area sets in, is shifted to lower values with increasing n.

This behavior implies that the reduced absorption can be traced back to exciton interaction effects akin to the dipole blockade effect known from Rydberg atoms (Urban *et al.*, 2009; Gaetan *et al.*, 2009). The blockade arises due to dipole-dipole interactions between Rydberg excitons, which strongly depend on their separation and spatial extension. Upon creation of an exciton, the required energy to create another one will be shifted by the dipole interaction energy. This energy shift will depend on the distance to the first exciton and within a certain blockade volume V_b it will be larger than the linewidth of the laser used for excitation. Accordingly, it will not be possible to create a second Rydberg exciton within V_b. In most cases in a real crystal, the illuminated volume will be larger than V_b. The experimental signature of dipole blockade will thus be a reduction of the absorption α when Rydberg excitons with density ρ_X are present. For a fixed laser power P_L, the absorption of the crystal at a certain laser energy will be reduced from the empty crystal absorption $\alpha_0(E)$ by a factor of $(1-\rho_X V_b)$. ρ_X Will in turn depend on P_L, α and the exciton lifetime τ_n.

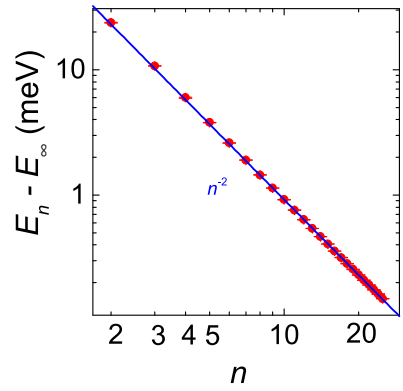

Fig. 2 Rydberg exciton binding energies. Symbols represent the measured energies. The solid line depicts the theoretically expected binding behavior assuming a binding energy of 92 meV. The experimental values show excellent agreement with the expected inverse square law.

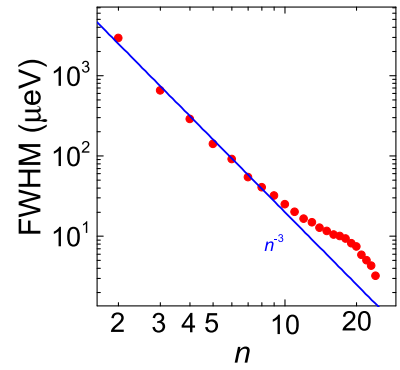

Fig. 3 Spectral width of the Rydberg exciton series. Symbols represent the experimentally determined values. The solid line represents the expected n^{-3}-behavior.

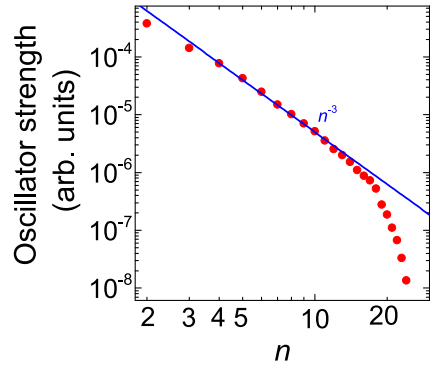

Fig. 4 Oscillator strength of Rydberg excitons. Symbols represent experimental data, while the solid line corresponds to the inverse cubic law expected in the absence of blockade effects.

Taking all of these effects into account, one can find the following simple scaling law for the absorption:

$$\alpha(P_L, E) = \frac{\alpha_0(E)}{1 + S_n P_L} \tag{7}$$

Here S_n denotes an effective efficiency with which the presence of excitons with principal quantum number n blocks further absorption of photons at the same energetic position. **Fig. 6** shows fits of Eq. (7) to the peak areas of the different exciton peaks seen in the experiment. The agreement is reasonable and S_n can be deduced from the fits. For large energies, it changes drastically with the principal quantum number and shows variations of several orders of magnitude. However, this strong effect helps one to

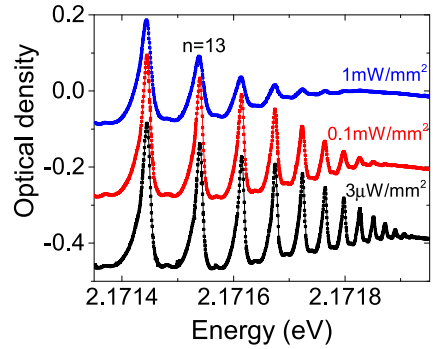

Fig. 5 Rydberg series absorption spectra measured with different laser intensities. The resonances at large n vanish at higher laser powers.

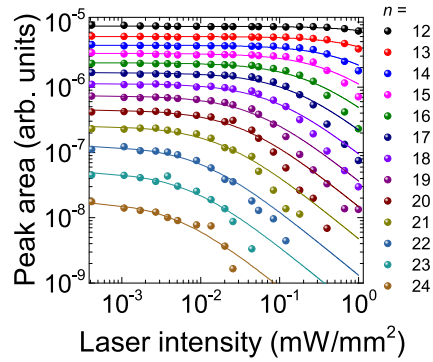

Fig. 6 Oscillator strength of Rydberg excitons for different laser powers. The solid lines represent fits according to Eq. (7). Reproduced from Feldmaier, M., Main, J., Schweiner, F., Cartarius, H., Wunner, G., 2016. Rydberg systems in parallel electric and magnetic fields: An improved method for finding exceptional points. Journal of Physics B: Atomic, Molecular and Optical Physics 49 (14), 144002.

identify the nature of the underlying effect causing the blockade. S_n shows a tenth power scaling law behavior with n. Possible dipole-type interaction mechanisms that could explain the blockade are either van-der-Waals interactions, where the interaction energy of two excitons separated by a distance R is given by

$$E_{V\,DW}(n) = -\frac{C_6}{R^6} \tag{8}$$

or Förster type interactions with a typical interaction energy of

$$E_F(n) = -\frac{C_3(n)}{R^3} \tag{9}$$

Theory shows that C_6 scales as n^{11}, while C_3 scales as n^4. In the limit of a spectrally narrow laser, the blockade radius R_b can be estimated by the dipole interaction becoming larger than the linewidth of the absorption:

$$R_B = (n) = \sqrt[q]{\frac{C_q(n)}{\Gamma_n}} \tag{10}$$

Assuming that the blockade efficiency is proportional to the product of $V_B = \frac{4}{3}\pi R_B^3$ and the exciton lifetime $\tau_n = \frac{\hbar}{\Gamma_n}$, one surprisingly finds that both types of interaction result in a blockade efficiency that scales as n^{10} in agreement with the experimental data, which is a good indicator that dipole blockade is indeed taking place in the exciton system (**Fig. 7**).

Symmetry Considerations

While it is instructive to study the similarities between hydrogen and excitons, especially for high-symmetry cases such as bulk semiconductors with cubic symmetry, also deviations from the hydrogen model, which arise due to reduced symmetry in other Rydberg systems, are highly interesting. For hydrogen, the spatial symmetry is governed by the continuous rotation group SO(3), which results in the square of the orbital momentum $L^2 = l(l+1)\hbar^2$ and its z-component $L_z = m\hbar$ being conserved. Also, this rotational symmetry assures that energy levels are degenerate with respect to the magnetic quantum number m. The hydrogen problem is also a Kepler-type inverse-square law central force problem. The bound Kepler problem may be mapped to a particle

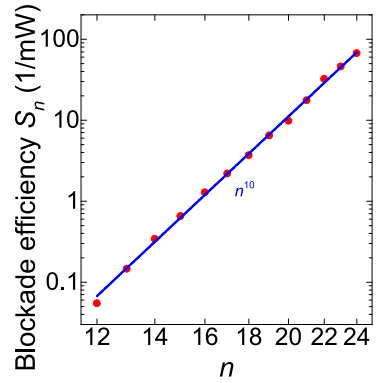

Fig. 7 Blockade efficiency for the different principal quantum numbers of the Rydberg exciton series. The solid line is a fit corresponding to the n^{10}-dependence.

confined on a three-dimensional sphere in four dimensions. Therefore, the problem is also symmetric with respect to rotations in four dimensions and the symmetry group of this problem is SO(4), which results in degeneracy of the energy levels with respect to the azimuthal quantum number l. One may now consider effects arising due to deviations of other Rydberg systems from the highly symmetric hydrogen problem. First, the SO(4) symmetry is broken if the central force shows deviations from the inverse-square law behavior. The most prominent cause for a modified central force is non-perfect screening, which is present is semiconductor Rydberg systems, but also in non-hydrogenic Rydberg atoms, where the inner electron shells only partially screen the nuclear Coulomb potential. For Rydberg atoms, this effect has been considered in terms of a quantum defect δ_l, which results in a modified binding energy. The probability to find an electron close to the nucleus varies for different l, which results in a pronounced dependence of the quantum defect on the angular momentum. However, screening by fully occupied electron shells leaves the SO(3) symmetry intact. The rotational symmetry is also lifted in semiconductor systems due to the presence of the periodic crystal lattice. This results in the reduction of the symmetry from a continuous rotation group towards discrete groups. For cubic bulk crystals these are O_h or T_d, depending on whether inversion symmetry is present or not. These still represent moderately high symmetry. Accordingly, it is in most cases sufficient to capture the deviations from full rotational symmetry in terms of the lattice-periodic Bloch functions of electrons and holes, while still treating the envelope wave function as hydrogen-like. The breakdown of this approximation results in the degeneracy of states with the same l being lifted and also in mixing of states with different l. In the following, we will discuss both symmetry reductions in terms of the equivalent of a quantum defect and the mixing of angular momentum states due to broken rotational symmetry for Rydberg excitons.

Mixing of Angular Momentum States

Due to its large Rydberg energy and small linewidths, Cu_2O is a promising candidate system for observing deviations from perfect rotational symmetry. Indeed these deviations have been observed in high resolution absorption studies. First, it is instructive to consider, which states one expects to see in experiments by means of group theory arguments. The valence and conduction bands associated with excitons of the yellow series correspond to the irreducible representations $\mathcal{D}_h = \Gamma_7^+$ and $\mathcal{D}_e = \Gamma_6^+$, respectively. The excitons may also be categorized in analogy to hydrogen in terms of the symmetry of the relative motion of electron and hole $\mathcal{D}_l$, which may or may not be irreducible. In order for a state to be optically active in one-photon transitions from the crystal ground state, the total symmetry representation of the exciton $\mathcal{D}_h \otimes \mathcal{D}_e \otimes \mathcal{D}_l$ must include the representation of the dipole operator corresponding to Γ_4^-. The representations of the most common exciton states (neglecting degeneracies) are listed in **Table 1**. As one can immediately see, P- and F-excitons are dipole-allowed, while S- and D-excitons are dipole-forbidden, but quadrupole-allowed. Also, one can readily see that only the representations of S- and P-excitons are irreducible, while the other angular momentum states consist of several representations and are thus reducible. This is already a signature of angular momentum not being a good quantum number anymore. Still, one can keep the classification of excitons in terms of angular momentum for simplicity, but should keep in mind that admixtures from other states are present (Schweiner *et al.*, 2016a). Therefore, the appearance of dipole-allowed F-exciton states in optical transmission experiments would be a good indicator for conservation of angular momentum being violated. A high-resolution transmission spectrum of a 30 μm crystal slab of Cu_2O held at 1.2 K in the region up to $n=9$ is shown in **Fig. 8** (Thewes *et al.*, 2015). While the P-excitons are most dominant in the spectrum, on their high energy side several weak signatures can be identified for excitons with a principal quantum number larger than 3. It should be noted that the highest possible angular momentum state that can occur for a fixed principal quantum number n is $n-1$. Accordingly, 4 is the lowest n, for which F-excitons can exist. Each of these structures consists of a triplet of lines, where the width of each line is in the low μeV range. Both the splitting between the lines of the triplet and the width of the individual lines decrease with increasing n. This triplet of states corresponds to the F-excitons. Starting from $n=6$ onwards another even weaker feature appears on the high energy side of

Table 1 Basic symmetries of Cu_2O

Object	*Symbol*	*Representation*
Electric dipole		Γ_4^-
Electric quadrupole		$\Gamma_3^+ \oplus \Gamma_5^+$
Magnetic dipole		Γ_4^+
Holes	$\mathcal{D}_h$	Γ_7^+
Electrons	$\mathcal{D}_e$	Γ_6^+
S-envelope	$\mathcal{D}_S$	Γ_1^+
P-envelope	$\mathcal{D}_P$	Γ_4^-
D-envelope	$\mathcal{D}_D$	$\Gamma_3^+ \oplus \Gamma_5^+$
F-envelope	$\mathcal{D}_F$	$\Gamma_2^- \oplus \Gamma_4^- \oplus \Gamma_5^-$
Excitons	$\mathcal{D}_h \otimes \mathcal{D}_e$	$\Gamma_2^+ \oplus \Gamma_5^+$
S-excitons	$\mathcal{D}_h \otimes \mathcal{D}_e \otimes \mathcal{D}_S$	$\Gamma_2^+ \oplus \Gamma_5^+$
P-excitons	$\mathcal{D}_h \otimes \mathcal{D}_e \otimes \mathcal{D}_P$	$\Gamma_2^- \oplus \Gamma_3^- \oplus \Gamma_4^- \oplus 2\Gamma_5^-$
D-excitons	$\mathcal{D}_h \otimes \mathcal{D}_e \otimes \mathcal{D}_D$	$\Gamma_1^+ \oplus 2\Gamma_3^+ \oplus 3\Gamma_4^+ \oplus 2\Gamma_5^+$
F-excitons	$\mathcal{D}_h \otimes \mathcal{D}_e \otimes \mathcal{D}_F$	$2\Gamma_1^- \oplus \Gamma_2^- \oplus 2\Gamma_3^- \oplus 4\Gamma_4^- \oplus 3\Gamma_5^-$

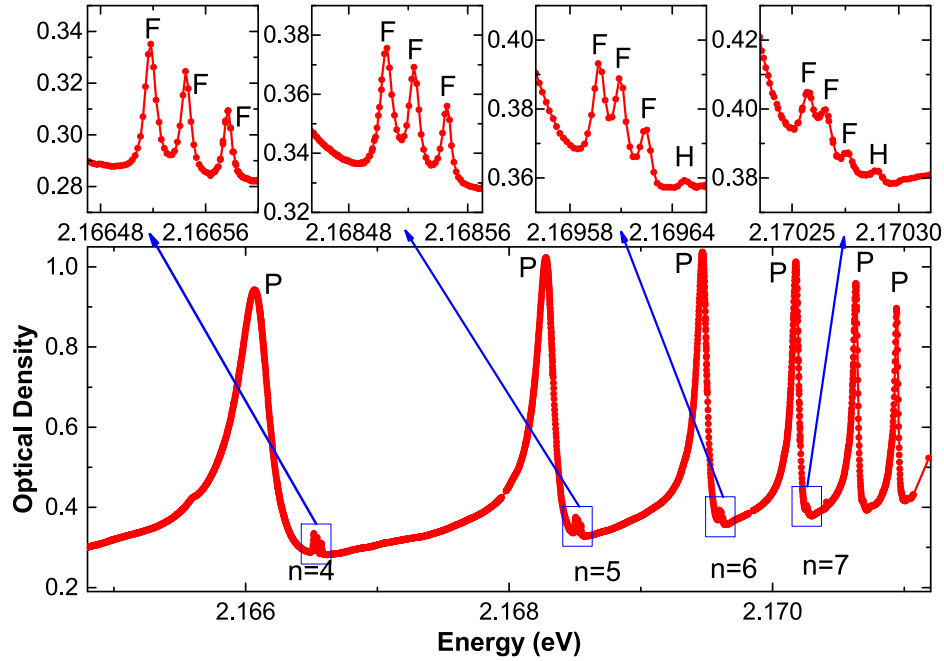

Fig. 8 Optical density of excitons in the spectral region between $n=4$ and $n=9$. On the high-energy side of the P-excitons additional resonances occur, which can be identified as F- and H-excitons. Insets show these resonances in more detail. The F-excitons show a threefold splitting, while a splitting is not resolvable for H-excitons.

the triplet. This state approaches the triplet with increasing n and can be identified as the H-exciton. It may consist of up to 5 lines, which are, however, not resolvable in the spectrum shown.

Quantum Defect and Non-Parabolic Bands

Historically, the first series of discrete emission lines in atomic spectra successfully classified, was the Balmer series of hydrogen (Jakob Balmer, 1885), which revealed the n^{-2}-scaling discussed earlier. However, already slightly more complicated atoms like alkali required modifications to the energy level scheme. The reason for that is that for all hydrogen-like atoms besides hydrogen the positive charges in the nucleus are not completely screened by the inner electrons. If the outer shell electron has a non-zero probability to be found close to the nucleus, the substructure of the nucleus will result in a slightly modified Coulomb potential. Accordingly, the largest deviations occur for s-states, which have a moderately large probability of being close to the nucleus. Rydberg (1890) showed that these deviations can be accounted for by adding an empirical correction to the binding energy term. He replaced

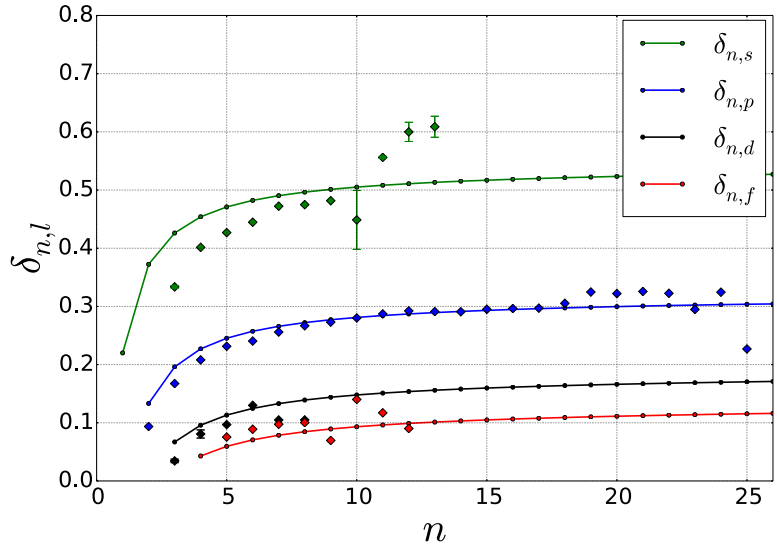

Fig. 9 Comparison of quantum defects determined experimentally (symbols) and via ab initio theory (lines) for the yellow exciton series. For S- and D-excitons, the experimental data has been obtained using an external electric field and interpolation to the zero-field limit. For lines showing a splitting, the center of gravity of the lines has been used. Adapted from Schöne, F., Krüger, S.-O., Grünwald, P., *et al.*, 2016a. Coupled valence band dispersions and the quantum defect of excitons in Cu_2O. Journal of Physics B: Atomic, Molecular and Optical Physics 49 (13), 134003.

the principal quantum number n by an effective non-integer quantum number $n^* = n - \delta_l$, where δ_l is the quantum defect mentioned before. Although this approach seems simplistic at first sight, it has been hugely successful because the quantum defect directly translates into a phase shift of the wave function in the region outside the core and thus simplifies calculations drastically.

Rydberg excitons in Cu_2O also show deviations from the ideal Rydberg series. These deviations can also be treated in terms of a quantum defect as the quantum defect determined by doing so converges to a constant value for fixed angular momentum and large principal quantum numbers (Schöne *et al.*, 2016b). Of course one has to keep the differences between atom and semiconductor physics in mind: Most calculations in atom physics are performed in real space, while Rydberg excitons are usually treated in momentum space and the physical origin of the deviations is vastly different. In fact, several effects could play a role. Central cell corrections (Washington *et al.*, 1977) corresponding to short range deviations of the electron-hole interaction potential from a pure Coulomb potential might arise due to coupling to LO-phonons (Schweiner *et al.*, 2016b), Coulomb screening, a strongly frequency dependent permittivity of the material or exchange interactions. However, it could be shown by fitting the results of spin-DFT calculations to the band Hamiltonian of Suzuki and Hensel (1974), that the non-parabolicity of the valence band is the dominant contribution to the quantum defect in Cu_2O (Schöne *et al.*, 2016a). A comparison of these results, which have been calculated without using experimentally determined fitting parameters, to the experimental data is shown in **Fig. 9**. The experimental data were obtained via absorption spectroscopy for P- and F-states. For S-and D-states an additional electrical field was applied in order to mix S-excitons with P-excitons and D-excitons with P- and F-excitons in first order. The results were then extrapolated to zero external field. The overall agreement between theory and experiment is convincing. Similar to the situation in atom physics, the quantum defect becomes largest for states of low angular momentum. The theory value saturates at a value slightly above 0.5 for S-excitons, while it amounts to about 0.3 for P-excitons. For D- and F-excitons the quantum defect is between 0.1 and 0.2, which is a rather large value compared to cold atoms, where the quantum defect approaches zero quickly for large angular momentum states. However, for the S-series one still observes systematic differences between experiment and theory for low n. In this spectral range, the background permittivity $\varepsilon(k,\omega)$ varies strongly (Dawson *et al.*, 1973) and the binding energy for $n=2$ is in the close vicinity of a phonon resonance. Most likely the deviations can be traced back to these effects.

Coherent Features

For principal quantum numbers larger than 12, the transmission spectrum of Rydberg excitons shown in **Fig. 1** shows some additional resonances between two P-exciton states. Even assuming a huge quantum defect, these additional states are too far away from the P-states to be excitons with larger angular momentum. They appear almost, but not exactly, in the middle between two P-states and are only visible in laser transmission experiments, but not using a white light sources and vanish for large laser powers. It was shown that these resonances can be traced back to coherent coupling between P-excitons of different n (Grünwald *et al.*, 2016) and therefore to off-diagonal elements of the system density matrix in the Fock basis. This coherent effect is a good testbed for estimating whether Rydberg excitons are a suitable candidate for investigating more complex quantum coherent effects. One may gain deeper insights into coherences and dephasing in the system. Especially the latter is usually very strong in bulk systems. A close-up view of some of these intermediate resonances is shown in **Fig. 10**. Their existence cannot be explained by a theory just

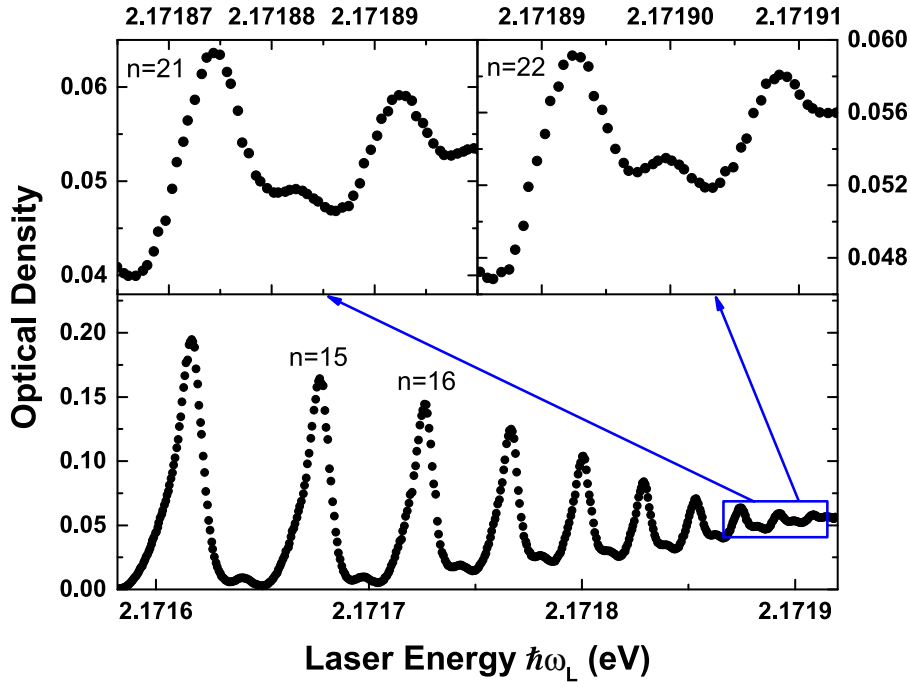

Fig. 10 Experimental absorption spectra of the Rydberg exciton series showing additional resonances between isolated resonances.

assuming the bare absorption spectrum of the system according to the Toyozawa line shape of resonances shown in Eq. (6). However, the presence of a strong light field driving a system may change its spectral properties by introducing incoherent and coherent scattering contributions. These result in resonance fluorescence contributions known, for example, from well-studied effects like the Mollow triplet (Mollow, 1969) and also in modified absorption properties (Mollow, 1972). Accordingly, the total spectral response of the system can be decomposed into a contribution due to the bare absorption $a(E)$ of the system and a function $I_X(E,E_L)$ that describes the incoherent redistribution towards other energies when driving the system at a well defined energy E_L. The total absorption of the system $A(E_L)$ will be given by a convolution of both:

$$A(E_L) = \int_{-\infty}^{\infty} dE\, a(E) I_X(E, E_L) \tag{11}$$

In the limit of small light-matter coupling strength, the so-called Heitler regime (Matthiesen *et al.*, 2012), I_X will correspond to a delta distribution. In that case, the basic Toyozawa lineshape is recovered. For larger light intensities, one can recover the full spectrum for an isolated resonance using a master equation approach that yields:

$$A(E_L) = g^2 \Omega_n^2 C_n \frac{\frac{\Gamma_n}{2} - 2q_n(E_n - E_L)}{\left[\left(\frac{\Gamma_n}{2}\right)^2 + 2\Omega_n^2 + (E_n - E_L)^2\right]^2} \tag{12}$$

where g denotes the exciton dipole transition moment and Ω_n represents the Rabi energy. While there are differences between this corrected lineshape and the Toyozawa lineshape, for isolated resonances these can only be clearly distinguished far away from the peak and the resonances are too close to allow for a trivial identification of the more appropriate lineshape. However, there are other effects like power broadening not covered by the Toyozawa lineshape, which can be used to identify the difference in a systematic measurement at different laser powers. Also, for the range of states where the correction matters, the spacing between states of consecutive principal quantum numbers is roughly of the same order of magnitude as the Rabi energy. The energy separation of Rydberg exciton states with $n=15$ and $n=16$ amounts to about 50 μeV. Accordingly, the resonances rather form a V-type system, where the states $|n\rangle$ and $|n+1\rangle$ are the excited states coupled to the excitonic vacuum $|0\rangle$ and a laser placed in the middle between the two resonances can drive both of them at a moderate detuning. The level scheme is sketched schematically in **Fig. 11**. The excited states are not coupled to each other and there is only one exciton present in the system. This system can be modeled using a master equation approach (Grünwald *et al.*, 2016). One finds that off-diagonal terms in the density matrix occur, which correspond to an effective quadrupolar coherent coupling of the two excited states via the common ground state. The transition rate directly depends on the two Rabi frequencies. The amplitude of this intermediate resonance will depend strongly on the incoherent redistribution spectrum $S_X = \frac{I_X}{g^2}$ of the excitons, which is shown in **Fig. 12** for pumping exactly in the middle between the resonances. When increasing the Rabi energy, the resonance at the laser position starts to build, which corresponds to a dressed state of the two quadrupole-coupled exciton states. The total absorption spectrum is now given by the convolution of I_X and the bare Toyozawa absorption lineshape. A comparison of experimental results and theoretical results obtained by

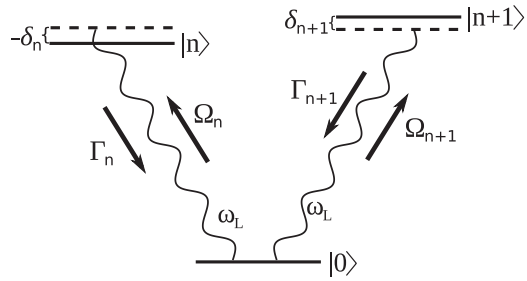

Fig. 11 Model of Rydberg excitons as a V-type three level system. Two Rydberg excitons with principal quantum numbers n and $n+1$ are coupled to the excitonic ground state via a single light field at different detunings.

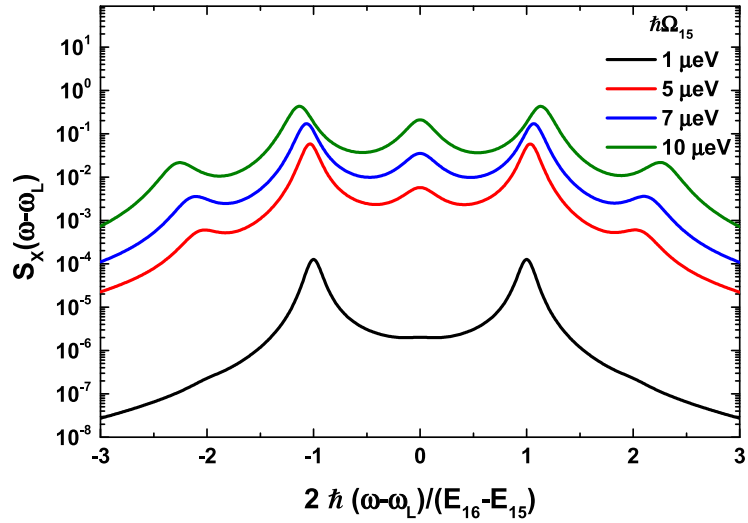

Fig. 12 Incoherent exciton spectral redistribution function for a laser placed exactly in the middle between resonances with principal quantum number $n=15$ and $n=16$.

incorporating the quadrupolar coupling is shown in **Fig. 13**. Both the asymmetry of the intermediate resonance and its appearance at moderate and disappearance at huge pump powers are reproduced qualitatively. The former is a consequence of the asymmetry of the underlying Toyozawa lineshape, while the latter is caused by the onset of saturation. The onset of Rydberg blockade may also add to this effect, but is not incorporated into the theoretical model. It should be noted that the presence of pure dephasing should destroy these coherences quickly, so the results imply that the pure dephasing rate for Rydberg excitons is surprisingly small, which makes them a good candidate to look for other coherent effects and implement coherent control in a bulk system.

Summary and Outlook

This article has placed a special emphasis on Rydberg states in Cu_2O, but this material system is of course not the only possible candidate to observe Rydberg physics in semiconductor systems. Any system with large exciton binding energy and moderately narrow linewidths might show similar effects. Low-lying levels of the Rydberg series have been observed, for example, in transition metal dichalcogenides (Chernikov *et al.*, 2014) or impurities in Si (Vinh *et al.*, 2008), but so far either the linewidths or the quality of the structures prohibit observation of states with large n. This article has placed some emphasis on differences to Rydberg atoms, which arise due to the semiconductor setting. Along these lines, it has further been suggested that the deviations between excitons and the case of hydrogen should also provide deeper insights into the classical-to-quantum transition in large external fields. When placed in moderate magnetic fields, it could already be shown that the Rydberg exciton system shows quantum chaos of a non-trivial kind (Aßmann *et al.*, 2016). In fact, the breaking of all anti-unitary symmetries results in a drastic change of the level spacing statistics, which in turn leads to a quadratic suppression of small level spacings. As the Rydberg energy of exciton systems is much smaller compared to atom physics, also the regime of high external fields is reached much easier. Typical field strengths available in a laboratory setting are sufficient. Therefore, one may expect to find further specific effects of quantum chaos, which are hard to observe in atomic physics like exceptional points (Feldmaier *et al.*, 2016), which have been observed so far only in non-Hermitian photonic billiards (Gao *et al.*, 2015).

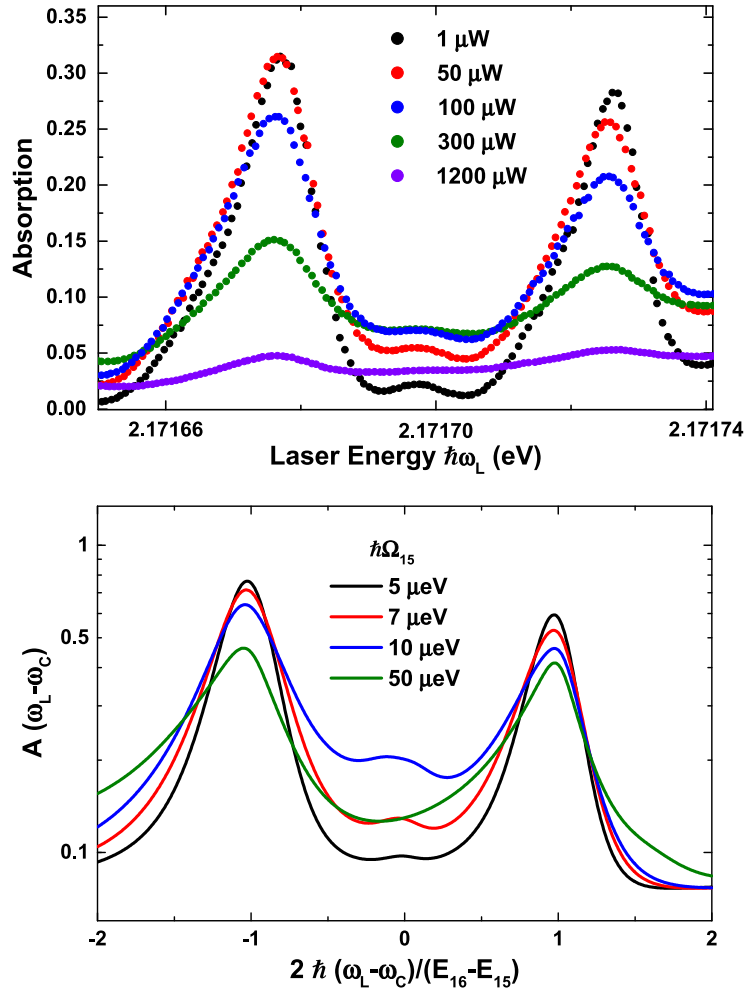

Fig. 13 Upper panel: Experimental absorption spectra between the resonances with $n=15$ and $n=16$ for the laser powers shown in the legend. Bottom: Theoretical calculation of the absorption spectrum in the same spectral region.

For Rydberg excitons in Cu_2O one may envisage that future research will also develop along the lines of what has already been done for cold atoms (Saffman *et al.*, 2010). Either by going to higher principal quantum numbers or thinner samples, it should become possible to create a blockade radius comparable to the thickness of the sample, resulting in an effective two-dimensional Rydberg system, where each blockade volume is occupied by a single exciton. Using tailored pulses matched to the dephasing time and the spectral width of the resonances, it may become possible to drive Rabi oscillations in such a system (Dudin *et al.*, 2012; Johnson *et al.*, 2008), which would pave the way towards transferring the broad range of quantum technologies already realized with Rydberg atoms to a semiconductor system.

Acknowledgements

This article owes much to our collaboration with active researchers in the field. It is our pleasure to express our gratitude to Dietmar Fröhlich, Tomasz Kazimierczuk, Stefan Scheel, Heinrich Stolz, Jörg Main, Harald Gießen and Mikhail Glazov. We also gratefully acknowledge the support by the Deutsche Forschungsgemeinschaft in the frame of SPP 1929 GiRyd and ICRC TRR 160, the Russian Foundation for Basic Research in the frame of ICRC TRR 160 and the support from the Russian Ministry of Science and Education (contract number 14. Z50.31.0021).

References

Aßmann, M., Thewes, J., Fröhlich, D., Bayer, M., 2016. Quantum chaos and breaking of all anti-unitary symmetries in Rydberg excitons. Nature Materials 15, 741.
Agekyan, V.T., 1977. Spectroscopic properties of semiconductor crystals with direct forbidden energy gap. Physica Status Solidi (a) 43 (1), 11–42.

Chernikov, A., Berkelbach, T.C., Hill, H.M., *et al.*, 2014. Exciton binding energy and non-hydrogenic Rydberg series in monolayer WS_2. Physical Review Letters 113 (7), 076802.
Dawson, P., Hargreave, M.M., Wilkinson, G.R., 1973. The dielectric and lattice vibrational spectrum of cuprous oxide. Journal of Physics and Chemistry of Solids 34 (12), 2201–2208.
Dudin, Y.O., Li, L., Bariani, F., Kuzmich, A., 2012. Observation of coherent many-body Rabi oscillations. Nature Physics 8 (11), 790–794.
Feldmaier, M., Main, J., Schweiner, F., Cartarius, H., Wunner, G., 2016. Rydberg systems in parallel electric and magnetic fields: An improved method for finding exceptional points. Journal of Physics B: Atomic, Molecular and Optical Physics 49 (14), 144002.
Gaetan, A., Miroshnychenko, Y., Wilk, T., *et al.*, 2009. Observation of collective excitation of two individual atoms in the Rydberg blockade regime. Nature Physics 5 (2), 115–118.
Gallagher, T.F., 2005. Rydberg Atoms, vol. 3. Cambridge University Press.
Gao, T., Estrecho, E., Bliokh, K.Y., *et al.*, 2015. Observation of non-hermitian degeneracies in a chaotic exciton-polariton billiard. Nature 526 (7574), 554–558.
Gross, E.F., 1956. Optical spectrum of excitons in the crystal lattice. Il Nuovo Cimento (1955–1965) 3, 672–701.
Gross, E.F., Karryjew, I.A., 1952. The optical spectrum of the exciton. Doklady Akademii Nauk SSSR 84, 471–474.
Grünwald, P., Aßmann, M., Heckötter, J., *et al.*, 2016. Signatures of quantum coherences in Rydberg excitons. Physical Review Letters 117, 133003.
Hayashi, M., Katsuki, K., 1952. Hydrogen-like absorption spectrum of cuprous oxide. Journal of the Physical Society of Japan 7, 589.
Jakob Balmer, J., 1885. Notiz über die spectrallinien des wasserstoffs. Annalen der Physik 261 (5), 80–87.
Johnson, T.A., Urban, E., Henage, T., *et al.*, 2008. Rabi oscillations between ground and Rydberg states with dipole–dipole atomic interactions. Physical Review Letters 100, 113003.
Kavoulakis, G.M., Chang, Y.-C., Baym, G., 1997. Fine structure of excitons in Cu_2O. Physical Review B 55 (12), 7593.
Kazimierczuk, T., Fröhlich, D., Scheel, S., Stolz, H., Bayer, M., 2014. Giant rydberg excitons in the copper oxide Cu_2O. Nature 514 (7522), 343–347.
Matsumoto, H., Saito, K., Hasuo, K., Kono, S., Nagasawa, N., 1996. Revived interest on yellow-exciton series in Cu_2O: An experimental aspect. Solid State Communications 97, 125–129.
Matthiesen, C., Vamivakas, A.N., Atatüre, M., 2012. Subnatural linewidth single photons from a quantum dot. Physical Review Letters 108, 093602.
Mollow, B.R., 1969. Power spectrum of light scattered by two-level systems. Physical Review 188, 1969–1975.
Mollow, B.R., 1972. Stimulated emission and absorption near resonance for driven systems. Physical Review A 5, 2217–2222.
Rydberg, J.R., 1890. XXXIV. On the structure of the line-spectra of the chemical elements. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 29 (179), 331–337.
Saffman, M., Walker, T.G., Mølmer, K., 2010. Quantum information with Rydberg atoms. Reviews of Modern Physics 82, 2313–2363.
Schöne, F., Krüger, S.-O., Grünwald, P., *et al.*, 2016a. Coupled valence band dispersions and the quantum defect of excitons in Cu_2O. Journal of Physics B: Atomic, Molecular and Optical Physics 49 (13), 134003.
Schöne, F., Krüger, S.-O., Grünwald, P., *et al.*, 2016b. Deviations of the exciton level spectrum in Cu_2O from the hydrogen series. Physical Review B 93, 075203.
Schweiner, F., Main, J., Feldmaier, M., Wunner, G., Uihlein, C., 2016a. Impact of the valence band structure of Cu_2O on excitonic spectra. Physical Review B 93, 195203.
Schweiner, F., Main, J., Wunner, G., 2016b. Linewidths in excitonic absorption spectra of cuprous oxide. Physical Review B 93, 085203.
Suzuki, K., Hensel, J.C., 1974. Quantum resonances in the valence bands of germanium. I. Theoretical considerations. Physical Review B 9 (10), 4184.
Thewes, J., Heckötter, J., Kazimierczuk, T., *et al.*, 2015. Observation of high angular momentum excitons in cuprous oxide. Physical Review Letters 115, 027402.
Toyozawa, Y., 1964. Interband effect of lattice vibrations in the exciton absorption spectra. Journal of Physics and Chemistry of Solids 25 (1), 59–71.
Urban, E., Johnson, T.A., Henage, T., *et al.*, 2009. Observation of Rydberg blockade between two atoms. Nature Physics 5 (2), 110–114.
Vinh, N.Q., Greenland, P.T., Litvinenko, K., *et al.*, 2008. Silicon as a model ion trap: Time domain measurements of donor Rydberg states. Proceedings of the National Academy of Sciences 105 (31), 10649–10653.
Washington, M.A., Genack, A.Z., Cummins, H.Z., *et al.*, 1977. Spectroscopy of excited yellow exciton states in Cu_2O by forbidden resonant Raman scattering. Physical Review B 15 (4), 2145.

Two-Dimensional Coherent Spectroscopy of Transition Metal Dichalcogenides

Galan Moody, National Institute of Standards & Technology, Boulder, CO, United States

Published by Elsevier Ltd.

Introduction

Coherent optical spectroscopy is a powerful technique for understanding light–matter interaction in condensed matter, atomic, and molecular systems. The simplest conceptual experiment one can perform is to excite a system with a weak optical field (compared to the internal fields of the system) and measure the optical absorption. The incident light generates a coherent superposition between the ground and excited states, which is described macroscopically as a polarization that is linearly proportional to the incident field (Fig. 1). The linear absorption spectrum, arising from interference between the incident light and re-radiated light from the polarization, provides information about the optical transition frequencies, linewidths, and dipole moments.

Despite its utility, linear absorption spectroscopy has several limitations, particularly when used to study heterogeneous ensembles. An individual emitter in the ensemble – a semiconductor nanostructure, atom, or molecule – experiences scattering processes that alter its phase, resulting in free induction decay, or dephasing, of the optical polarization (or equivalently, *homogeneous broadening* γ of the absorption linewidth). Each emitter is also coupled to the surrounding environment, which leads to stochastic fluctuations of its optical and electronic properties. For example, random disorder potentials from imperfections in semiconductor nanostructures introduce a distribution of transition frequencies to the system (*inhomogeneous broadening* σ), which results in polarization decay as individual emitters in the ensemble oscillate out of phase. In linear spectroscopy, inhomogeneous broadening masks the homogeneous optical response as shown in Fig. 1.

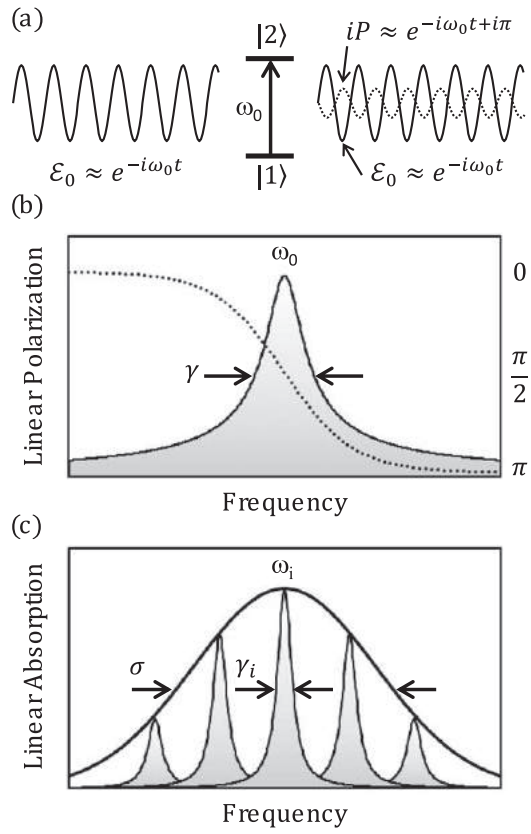

Fig. 1 (a) An incident optical field on resonance with a two-level transition drives a macroscopic polarization, which radiates out of phase with the incident field. (b) The full-width at half-maximum of the polarization frequency response is equal to the homogeneous dephasing rate $\gamma = \hbar/T_2$. On resonance, the polarization is $\pi/2$ out of phase with the incident field (dashed line, right axis). (c) The linear absorption spectrum of an ensemble of two-level systems, where the ith transition has frequency ω_i and homogeneous linewidth γ_i. The distribution of transition frequencies leads to inhomogeneous broadening σ of the absorption linewidth, masking the individual transitions.

 doi:10.1016/B978-0-12-803581-8.09581-3

This ambiguity can be circumvented with coherent nonlinear optical spectroscopy techniques, which are performed with sufficiently strong optical fields to induce a nonlinear polarization that is proportional to higher-order contributions of the electric field. If the incident field comprises a series of pulses with variable delays, the dynamical evolution of the system's optical response can be measured. One of the most widely-used nonlinear spectroscopy techniques is two- or three-pulse transient four-wave mixing (FWM). The FWM field – re-radiated light from the nonlinear polarization – contains rich information about the photo-excited states in the system. Examples include spectrally resolved transient absorption measurements that provide excited-state lifetimes and photon-echo experiments that reveal the average homogeneous dephasing rate of an ensemble even in the presence of disorder.

An enhanced version of three-pulse FWM, multi-dimensional coherent spectroscopy (MDCS) provides a more comprehensive picture of a system's optical and electronic properties. Whereas one-dimensional nonlinear spectroscopy yields the average system response, the appeal of MDCS lies in its ability to interrogate different emitters in an ensemble. In MDCS, the phase evolution of each oscillator is correlated across two (2DCS) or three (3DCS) time periods. Fourier transformation of time-domain measurements spreads the optical response into a multi-dimensional spectrum, elucidating both the individual and ensemble-averaged properties of the system. In principle, each emitter can be isolated and studied one-at-a-time with one-dimensional techniques; however, this approach does not provide details of interactions or coupling between them.

The Fourier-transform methods of optical MDCS were originally developed for nuclear magnetic resonance spectroscopy of spin systems. With the advent of femtosecond laser technologies, multi-dimensional Fourier-transform techniques emerged in the infrared and visible regimes to study vibrational and electronic coherences in molecules, energy transfer and many-body interactions in semiconductor nanostructures, and dipolar interactions in atomic vapors. More recently, MDCS experiments on a new class of atomically thin semiconductors – namely, monolayer transition metal dichalcogenides – have demonstrated the capabilities of the technique for disentangling complex one-dimensional spectra, elucidating the Coulomb interaction strengths between photo-excitations, and revealing new types of quasiparticles not easily accessible with other techniques.

In this article, the basic principles of multi-dimensional coherent spectroscopy are introduced with a focus on optical 2DCS. A typical experimental implementation and a theoretical description of light–matter interaction in semiconductors based on the optical Bloch equations are introduced. The capabilities of 2DCS in identifying and characterizing fundamental optical excitations in disordered media are illustrated with examples from recent experiments on monolayer semiconductors.

Optical Two-Dimensional Coherent Spectroscopy

Coherent Light–Matter Interaction

Nonlinear optical spectroscopy techniques rely on sufficiently strong optical fields to drive the polarization beyond the linear regime, i.e.

$$P = \chi^{(1)}\mathscr{E} + \chi^{(2)}\mathscr{E}^2 + \chi^{(3)}\mathscr{E}^3 + \ldots \tag{1}$$

where $\mathscr{E}$ is the incident electric field, $\chi^{(n)}$ is the $n^{(th)}$-order susceptibility, and P is the induced polarization (the vector and tensor notation has been omitted). For weak fields, only the linear response is measured, which contains information about the refractive index and linear absorption. Second-order $\chi^{(2)}$ nonlinearities, which encompass three-wave mixing effects including sum- and difference-frequency generation, are absent in materials with spatial inversion symmetry. Thus, the lowest-order nonlinear optical response that exists in all materials arises from $\chi^{(3)}$ processes.

The $\chi^{(3)}$ susceptibility gives rise to a host of nonlinear optical phenomena, including Kerr-lens mode-locking, self-phase modulation, and four-wave mixing (FWM), which is the basis for optical 2DCS techniques. In 2DCS experiments, each optical pulse from the output of a mode-locked oscillator or parametric amplifier is split into four phase-coherent replicas, three of which generate an optical polarization in the system. The incident field intensity is sufficiently strong to drive the system into the nonlinear regime but is kept weak enough to avoid generating higher-order nonlinearities beyond $\chi^{(3)}$, i.e. FWM spectroscopy is a perturbative technique. The FWM signal field is measured independently from other linear and nonlinear contributions to the polarization through spatial or temporal pulse modulation schemes, discussed in more detail in the following section.

The FWM excitation pulse sequence for optical 2DCS experiments is shown in **Fig. 2**. The electric field of each pulse can be written as $\mathscr{E}_i(\boldsymbol{r},t) = \hat{\mathscr{E}}_i e^{i(\boldsymbol{k}_i\boldsymbol{r}-\omega_i t)} + c.c.$, where $\boldsymbol{k}_i$, ω_i, and $\hat{\mathscr{E}}_i$ are the wavevector, angular frequency, and field envelope of the ith pulse. The total electric field is a sum of the three incident fields: $\mathscr{E}(\boldsymbol{r},t) = \sum_i \mathscr{E}_i(\boldsymbol{r}_i,\omega_i,t-t_i)$, where the ith pulse arrives at time t_i. Only considering contributions to the third-order polarization term, $P^{(3)} = \chi^{(3)}\mathscr{E}^3$, that are a product of all three fields, the polarization can be written as

$$P^{(3)}(\boldsymbol{r},t_1,t_2,t_3) = \int_0^\infty \mathscr{R}^{(3)}\left(t_1',t_2',t_3'\right)\mathscr{E}_1\left(\boldsymbol{r},t_1'-t_1\right)\mathscr{E}_2\left(\boldsymbol{r},t_2'-t_2\right)\mathscr{E}_3\left(\boldsymbol{r},t_3'-t_3\right)dt_1'dt_2'dt_3', \tag{2}$$

where $\mathscr{R}^{(3)}$ is the third-order time-dependent response function that contains details of the system's transition frequencies, dipole moments, excited-state dynamics, couplings, etc. The nth-order response function $\mathscr{R}^{(n)}$ can be calculated using time-dependent perturbation theory; in general, a sum-over-states expression for $\mathscr{R}$ can be derived for an any arbitrary system. $P^{(3)}$ obtained from Eq. (2) can be inserted into Maxwell's equations as a source term to describe radiation and propagation of the FWM signal field.

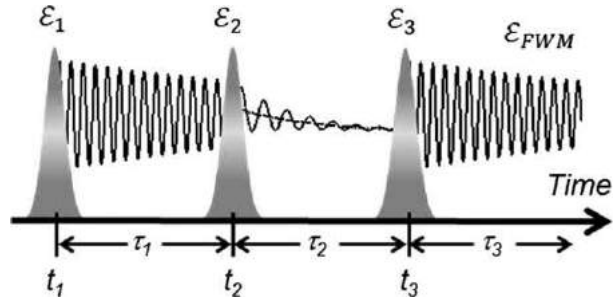

Fig. 2 Time-ordering for three-pulse four-wave mixing (FWM). The FWM signal field can be measured as a function of delays τ_1, τ_2, and τ_3. Fourier transformation with respect to any two delays generates a two-dimensional coherent spectrum.

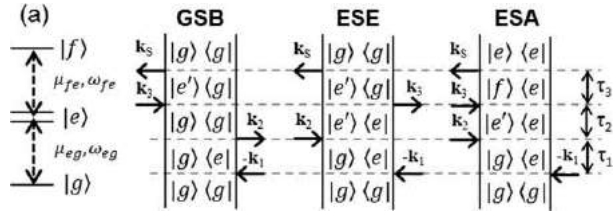

Fig. 3 Double-sided Feynman diagrams representing the coherent quantum pathways for a 2DCS experiment in the rephasing time-ordering (conjugate pulse arrives at the sample first). The singly-excited state manifold $|e\rangle$ with transition frequency ω_{eg} and dipole moment μ_{eg} is accessed through ground-state bleaching (GSB) and excited-stated emission (ESE) pathways; the doubly-excited state manifold $|f\rangle$ with transition frequency ω_{fe} and dipole moment μ_{fe} is accessed through excited-state absorption (ESA) pathways.

A convenient approach for generating $\mathscr{R}^{(3)}$ is through the density matrix formalism, which is useful for describing a statistical ensemble of quantum states. The equations of motion for the density operator $\rho = \sum_n p_n |\psi_n(t)\rangle\langle\psi_n(t)|$ are written as $\dot{\rho} = -i/\hbar[H,\rho]$, where $|\psi_n(t)\rangle$ is the wave function of state n with occupation probability p_n. The light–matter interaction with the system can be described by the Hamiltonian $H = H_0 + H_I$, where the free-particle eigenenergies are given by H_0 and the light–matter interaction takes on the form $H_I = -\boldsymbol{\mu}\cdot\mathscr{E}(\boldsymbol{r},t)$ in the dipole approximation. The equations of motion for ρ, known as the optical Bloch equations (OBEs), describe the time evolution of the density matrix elements, which as expressed do not contain any relaxation or dephasing parameters. These can be included phenomenologically, resulting in modified OBEs with matrix elements described by

$$\dot{\rho}_{ij} = \frac{-i}{\hbar}\sum\nolimits_k \left(H_{ik}\rho_{kj} - \rho_{ik}H_{kj}\right) - \gamma_{ij}\rho_{ij}, \tag{3}$$

where $\gamma_{ij} = \left(\Gamma_i + \Gamma_j\right)/2 + \gamma_{ij}^{ph}$, Γ_i $\left(\Gamma_j\right)$ is the decay rate of state i (j), and γ_{ij}^{ph} $(i \neq j)$ is the elastic pure dephasing rate corresponding to scattering processes that affect coherence without altering the population state of the system. The OBEs are useful for modeling strong light–matter interaction that gives rise to non-perturbative phenomena such as Rabi flopping and coherent control of quantum systems.

For FWM experiments that are performed in the weak-field limit, the light–matter interaction can be treated perturbatively by expanding the OBEs in terms of the Rabi frequency $\Omega \equiv \mu\hat{\mathscr{E}}/\hbar$. The nth-order equation of motion for the density matrix (in the rotating-wave approximation) is expressed as:

$$\dot{\rho}_{ij}^{(n)} = -i\Delta_{ij}\rho_{ij}^{(n)} + \frac{i\Omega}{2}\rho_{kl}^{(n-1)}, \tag{4}$$

where n is the perturbation order and $\Delta_{ij} \equiv \omega_i - \omega_j - i\gamma_{ij}$. From Eq. (4), it is evident that the light–matter interaction increases the perturbation order from n-1 to n. To first-order, the optical field drives the system from a ground state population $\rho_{11} = |1\rangle\langle 1|$ (diagonal element of the density matrix) to an optical coherence $\rho_{12} = |1\rangle\langle 2|$ or $\rho_{21} = |2\rangle\langle 1|$ (off-diagonal elements). The action of the second field drives the coherence to a second-order population (ρ_{11} or ρ_{22}). Thus, odd-orders of the field generate coherences while even-orders generate populations.

Four-wave mixing experiments can be modeled by expanding Eq. (4) up to $\rho^{(3)}$, the components of which can be conveniently represented through doubled-sided Feynman diagrams that depict the quantum pathways connecting the initial and final states. Representative Feynman diagrams for a three-level ladder system are shown in **Fig. 3**, where the singly excited-state manifold $|e\rangle$ is connected to the ground state and doubly excited-state manifold $|f\rangle$ through the electric dipole interaction. The Feynman diagrams are shown for a "rephasing", or "photon echo", time ordering of the pulse sequence in which the first pulse incident on the sample acts as a conjugate field. In an inhomogeneous ensemble, the field $\mathscr{E}_1^*$ drives a first-order coherence that evolves with opposite phase compared to the third-order coherence generated by subsequent excitation with fields $\mathscr{E}_2$ and $\mathscr{E}_3$. In an inhomogeneous

ensemble, polarization decay due to inhomogeneous dephasing of the different oscillators during τ_1 is reversed during the delay τ_3, resulting in constructive interference of the polarization and a photon echo signal.

The three Feynman diagrams in **Fig. 3** represent ground state bleaching (GSB), excited-state emission (ESE), and excited state absorption (ESA) quantum pathways. Feynman diagrams are extremely useful for developing an intuitive understanding of FWM spectroscopy experiments. For example, $\rho^{(3)}$ for the ESA pathway is given by

$$\rho_{32}^{(3)} = \frac{i\mu_{fe}\mu_{eg}\mu_{ge}}{8\hbar^3} e^{i\boldsymbol{k}_s\cdot\boldsymbol{r}} \hat{\mathscr{E}}_1^* \hat{\mathscr{E}}_2 \hat{\mathscr{E}}_3 \Theta(\tau_1)\Theta(\tau_2)\Theta(\tau_3) \times e^{-(\gamma_{ge}+i\omega_{ge})\tau_1} e^{-\gamma_{ee}\tau_2} e^{-(\gamma_{fe}+i\omega_{fe})\tau_3}, \tag{5}$$

for $|e\rangle = |e'\rangle$ and delta-function pulse envelopes. In Eq. (5), μ_{ij} and $\omega_{ij} = \omega_i - \omega_j$ are the transition dipole moment and frequency between states i and j, respectively, Θ is the Heaviside step function, $\boldsymbol{k}_s = -\boldsymbol{k}_1 + \boldsymbol{k}_2 + \boldsymbol{k}_3$ is the signal wavevector, and γ_{ij} is the relaxation rate defined in Eq. (3). This expression can be adapted for other quantum pathways by replacing the variables with the appropriate values in the corresponding Feyman diagram. The macroscopic third-order polarization, $P^{(3)}$, is obtained by taking the trace of $\rho^{(3)}$ and μ.

Time-domain expressions like Eq. (5) for all quantum pathways for a specific level system can be summed to obtain the total third-order optical response. A two-dimensional rephasing spectrum is then generated by taking a Fourier transform with respect to τ_1 and τ_3. For instance, **Fig. 4** illustrates how the different quantum pathways contribute to the 2D rephasing spectrum for different types of level systems. The vertical and horizontal axes are the frequency-domain representation of the first-order (excitation energy, $\hbar\omega_1$) and third-order (emission energy, $\hbar\omega_3$) coherences during the delays τ_1 and τ_3, respectively, while the delay τ_2 is fixed at zero. Because the first field $\mathscr{E}_1$ acts as a conjugate pulse for this sequence, the excitation energies are shown as negative.

For a three-level *V*-system, the 2D spectrum features four peaks. The two peaks on the diagonal dashed line arise from excitation and emission of the two transitions described by GSB and ESE pathways in **Fig. 3**, where $|e\rangle = |e'\rangle$. The off-diagonal peaks indicate quantum mechanical coupling between the transitions, which is expected because they share a common ground state. These peaks are described by GSB and ESE quantum pathways for which $|e\rangle \neq |e'\rangle$. The first two pulses drive the system into a non-radiative Raman-like coherence that evolves at the difference frequency between the transitions, which appears as peaks off of the diagonal line in the 2D spectrum; in the time-domain, these low-frequency oscillations are responsible for coherent quantum beats between the transitions.

For two-independent two-level systems, the absence of a shared state eliminates the off-diagonal coupling peaks (right panel of **Fig. 4**) and only two peaks appear on the diagonal; however, off-diagonal peaks can appear if incoherent population relaxation, or energy transfer, is present between the transitions. In the example in **Fig. 4**, energy relaxation from state *B* to state *A* appears as a below-diagonal peak. The dynamics of this process is described by $\rho_{BB} \rightarrow \rho_{AA}$, which can be mapped by generating 2D spectra for various delays τ_2.

Examples of simple, homogeneous systems have been presented to illustrate the ability of 2DCS to discern different types of coupling between transitions. Inhomogeneous broadening in disordered media can also be modeled by integrating expressions like Eq. (5) over a two-dimensional frequency distribution function with characteristic width σ (see **Fig. 5**). In a rephasing 2D spectrum, inhomogeneity appears as broadening of a peak along the diagonal dashed line. In the inhomogeneous limit ($\sigma \gg \gamma$), the diagonal and anti-diagonal lineshapes provide a measure of the inhomogeneous and homogeneous linewidths, respectively. Thus, the homogeneous linewidth for an ensemble can be obtained for all resonance energies by taking anti-diagonal slices across the inhomogeneous distribution. For an arbitrary amount of inhomogeneous broadening described by a Gaussian distribution

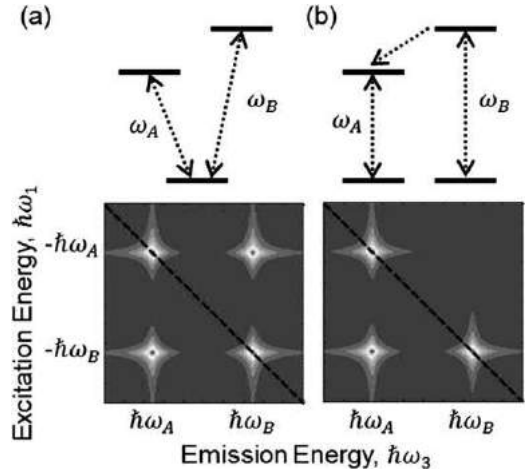

Fig. 4 (a) A three-level *V*-system and the corresponding rephasing 2D spectrum. (b) Two-independent two-level systems coupled by incoherent energy relaxation (population transfer) from state *B* to state *A* and the corresponding 2D spectrum.

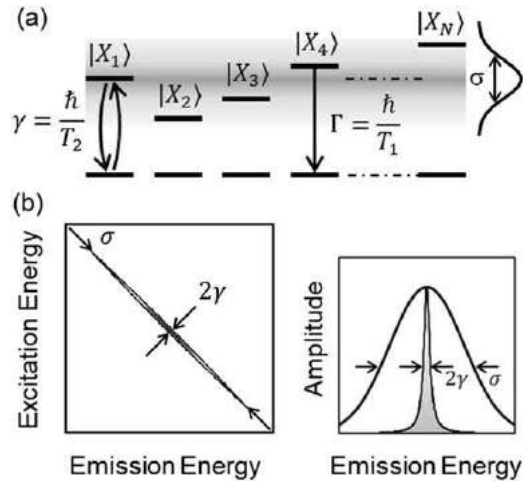

Fig. 5 (a) A heterogeneous ensemble of two-level systems with a distribution of transition energies given by σ; each transition $|X_i\rangle$ has a characteristic homogeneous linewidth $\gamma_i = \hbar/T_{2,i}$ and excited state lifetime $\Gamma_i = \hbar/T_{1,i}$. (b) Calculated 2D rephasing spectrum for a Gaussian inhomogeneous distribution function. In the strong inhomogeneous limit, the lineshapes along the anti-diagonal and diagonal directions, shown in the right panel, provide a measure of the homogeneous (γ) and inhomogeneous (σ) linewidths, respectively.

function, the projection-slice theorem of Fourier transforms enables the diagonal (S_D) and anti-diagonal (S_{AD}) lineshapes to be expressed as

$$S_D = \frac{\sqrt{2\pi}}{\gamma} Voigt(\gamma, \sigma), \tag{6}$$

$$S_{AD} = \frac{e^{-(\gamma - i\omega^2)^2/2\sigma^2} Erfc\left(\frac{\gamma - i\omega}{\sqrt{2}\sigma}\right)}{\sigma(\gamma - i\omega)} \tag{7}$$

where the *Voigt* profile is a convolution of Gaussian and Lorentzian functions and *Erfc* is the complementary error function. In the limit of strong inhomogeneous broadening, these expressions take on the form of Gaussian and square-root Lorentzian lineshapes, respectively.

The density matrix formalism presented in this section provides a background for intuitively interpreting and modeling a 2D spectrum in the rephasing pulse sequence. The rephasing spectrum reveals the individual and collective properties of disordered media generally not accessible with other techniques. For example, projecting the 2D spectrum onto the emission energy axis is analogous to acquiring an pump-probe spectrum; in this case, the homogeneous and inhomogeneous linewidths cannot be separated, and diagonal peaks spectrally overlap with off-diagonal peaks, masking evidence of coherent coupling in the system. Several different types of 2D spectra not discussed here are also accessible by employing different pulse time-ordering sequences and scanning different delays. These techniques can provide additional insight into the nature of many-body interactions, lifetime dynamics, and non-radiative coherences.

Experimental Implementation of 2DCS

2DCS techniques are implemented in several different configurations, several of which are depicted in **Fig. 6**. In one commonly used variant, the pulses propagate in a non-collinear geometry (the so-called "box" geometry). For three pulses $\mathscr{E}_1$, $\mathscr{E}_2$, and $\mathscr{E}_3$ incident on the system (in that order) with wavevectors $\boldsymbol{k}_1$, $\boldsymbol{k}_2$, and $\boldsymbol{k}_3$, respectively, the FWM signal field is radiated in the phase-matched direction determined by conservation of momentum, which separates the signal from other linear and nonlinear contributions to the polarization. Different types of 2D spectra can be generated by detecting the FWM signal along different directions: a rephasing or "photon echo" spectrum (S_I) along $\boldsymbol{k}_s = -\boldsymbol{k}_1 + \boldsymbol{k}_2 + \boldsymbol{k}_3$; a non-rephasing spectrum (S_{II}) along $\boldsymbol{k}_s = \boldsymbol{k}_1 - \boldsymbol{k}_2 + \boldsymbol{k}_3$; and a two-quantum spectrum (S_{III}) along $\boldsymbol{k}_s = \boldsymbol{k}_1 + \boldsymbol{k}_2 - \boldsymbol{k}_3$. In practice, the signal is usually detected along only one of these directions (see **Fig. 6**) and the different types of 2D signals are measured by changing the arrival time of the conjugate pulse $\mathscr{E}_1^*$. Other pulse geometries include a two-pulse pump-probe configuration and a collinear geometry, in which case the 2D signals are isolated through dynamic phase cycling techniques instead of wavevector selection.

An advantage of 2DCS compared to conventional pump-probe and linear spectroscopy techniques is its ability to measure both the FWM field and phase. Phase-sensitive measurements are performed through heterodyne interferometry with a fourth phase-stabilized local oscillator (LO) field $\mathscr{E}_{LO}$ with a known spectral phase. In a 2DCS experiment, the FWM/LO interference $\hat{\mathscr{E}}_s e^{i\varphi_s} \hat{\mathscr{E}}_{LO}^* e^{-i\varphi_{LO}} + c.c.$ is recorded as a function of two delays. For example, a 2D rephasing spectrum is obtained by recording the FWM signal while delays τ_1 and τ_3 are stepped by scanning pulses $\mathscr{E}_1^*$ and $\mathscr{E}_{LO}$. The FWM signal is then numerically Fourier

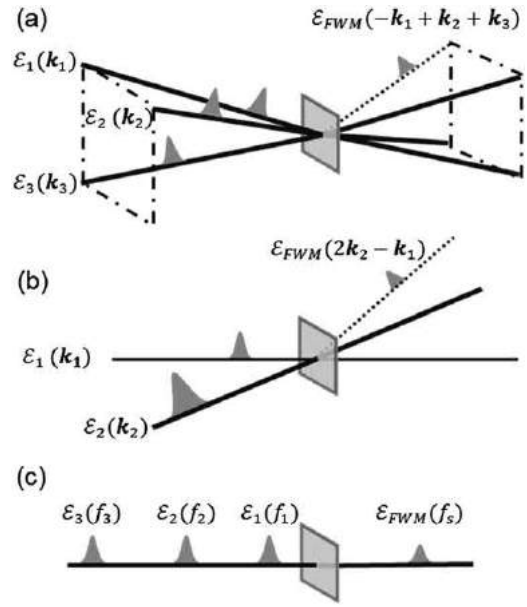

Fig. 6 Different implementations for MDCS experiments: (a) Three-pulse "box" geometry where the FWM signal is detected along $\boldsymbol{k}_s = -\boldsymbol{k}_1 + \boldsymbol{k}_2 + \boldsymbol{k}_3$; (b) Two-pulse pump–probe geometry, where the strong pump field $\mathscr{E}_1$ acts twice, resulting in $\boldsymbol{k}_s = 2\boldsymbol{k}_2 - \boldsymbol{k}_1$; (c) Collinear geometry in which the FWM signal is detected through phase-cycling techniques.

transformed with respect to these delays to generate the excitation and emission energies $\hbar\omega_1$ and $\hbar\omega_3$ of the 2D spectrum, respectively. Alternatively, the emission energies can be accessed by spectrally resolving the FWM/LO interference.

The FWM pulse sequence shown in Fig. 6 is obtained by splitting the output of a mode-locked laser into a series of variable-delay replicas using Michelson interferometers, Mach–Zehnder interferometers, or diffraction optics and pulse shapers. With any approach, the relative phase of the pulses must be precisely controlled with accuracy to much less than the wavelength to reliably perform a multi-dimensional Fourier transform of the FWM signal. Phase stability to better than $\lambda/100$ and timing-stepping resolution of $\sim$1 fs are possible using actively-stabilized interferometers or common path (passively-stabilized) optics. The advantage of the interferometer-based approach is the long maximum scan duration ($\sim$nanoseconds compared to $\sim$picoseconds for pulse shapers), but at the expense of greater experimental complexity.

In the rephasing time-ordering, 2DCS is an extremely powerful technique for characterizing disordered media owing to the separation and simultaneous determination of homogeneous and inhomogeneous dephasing. This capability makes 2DCS an ideal technique to investigate layered van der Waals semiconductors including transition metal dichalcogenides, which are sensitive to disorder and the surrounding environment. In the following section, examples of 2DCS experiments that have elucidated the intrinsic optical properties of 2D semiconductors are presented.

Optical Properties of Monolayer Transition Metal Dichalcogenides

Monolayer transition metal dichalcogenides (TMDs) are a recently discovered class of layered van der Waals semiconductors with the chemical formula MX_2, where M = Mo, W and X = S, Se, and Te. In the bulk form, TMDs are an indirect bandgap semiconductor with a honeycomb lattice. When reduced to a single monolayer (three atoms thick), TMDs become a direct bandgap semiconductor with conduction and valence band extrema at the K and K' points, or valleys, in the first Brillouin zone in momentum space (see Fig. 7). The combination of large spin-orbit splitting and time-reversal symmetry between the valleys introduces valley-contrasting properties; charges residing at the K and K' points possess opposite orbital magnetic moment, spin, and Berry curvature. This link between the electronic properties and the valley index leads to valley-dependent transition dipole selection rules. Specifically, photo-excitation of electron–hole pairs occurs in the K (K') valley for left (right) circularly polarized optical excitation.

One consequence of the atomic thickness of monolayer TMDs is a reduction in dielectric screening of the Coulomb interaction between band-edge electrons and holes. Combined with the large carrier effective masses, photo-excited electron–hole pairs form tightly bound hydrogenic states (excitons) with a binding energy on the order of 500 meV – nearly two orders of magnitude larger than conventional semiconductors such as GaAs. The exciton binding energy is significantly larger than k_bT at room temperature, which has implications for practical opto-electronic, photonic, and spin/valleytronic devices. In addition to excitons, higher-order few-body configurations exist, including charged excitons (trions) in doped TMDs, bound two-excitons (biexcitons), and charged biexcitons.

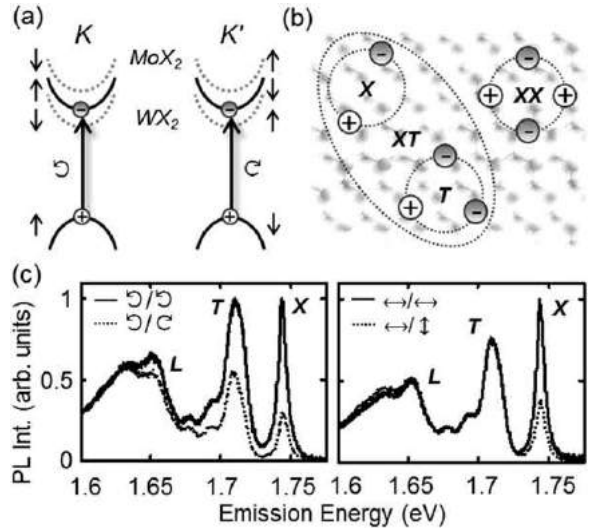

Fig. 7 (a) Optical transitions in monolayer TMDs at the *K* and *K*′ points at the edge of the first Brillouin zone. Electron–hole pairs are excited in the *K* (*K*′) valley for left (right) circularly polarized light. The optical transition to the lowest- (highest-) energy conduction band is dark in tungsten (molybdenum) based materials, indicated by the dashed bands. (b) Strong Coulomb interactions lead to the formation of tightly bound electron–hole pairs (excitons, *X*), charged excitons (trions, *T*), biexcitons (*XX*), and charged biexcitons *(XT*). (c) Photoluminescence spectra of monolayer WSe_2 at 10 K for circularly (linearly) polarized excitation and detection in the left (right) panel. The peaks are associated with excitons (*X*), trions (*T*), and localized (*L*) states. (Part (c) is adapted from Jones *et al.* 2013. Optical generation of excitonic valley coherence in monolayer WSe_2. *Nature Nanotechnology* 8: 634–638. Copyright (2013) by the Nature Publishing Group.)

Owing to their large binding energies, excitons and trions dominate the band-edge optical response of TMDs. Shown in Fig. 7, the low temperature photoluminescence spectrum of monolayer WSe_2 features several peaks that are attributed to excitons (X), trions (T), and localized states (L) that arise from impurities, defects, and many-body interactions with background charge carriers in doped samples. Several important aspects of TMDs can be gleaned from this spectrum alone: (1) the trion binding energy (difference between the exciton and trion resonance energies) is $\sim$ 30 meV; (2) the optical linewidths are on the order of $\sim$10 meV; and (3) after photo-excitation with left circularly polarized light, the exciton and trion emission is primarily left polarized. This last point reveals that photo-excitation and recombination of excitons and trions primarily occurs from the same valley (valley polarization). Furthermore, using linearly polarized excitation, the emission polarization axes for the exciton is parallel with the excitation, which is evidence of a coherent superposition of an exciton between the K and K' valleys (valley coherence).

The dynamics of valley depolarization and decoherence are sensitive to the type of TMD examined; for example, the lowest-energy transition in WX_2 (MoX_2) is optically dark (bright), which can affect the valley and lifetime dynamics. Despite the utility of photoluminescence (and other linear spectroscopies) in identifying excitonic states in TMDs, more advanced spectroscopy techniques are required to fully characterize the incoherent (lifetime) and coherent (dephasing time) dynamics.

2DCS of Monolayer TMDs

Homogeneous and Inhomogeneous Broadening

A representative 2D rephasing spectrum of monolayer WSe_2 on a transparent sapphire substrate is shown in Fig. 8 for co-circular excitation and detection of the signal. The spectrum features a single peak centered on the diagonal dashed line associated with the exciton. The peak lineshape asymmetry indicates that the exciton is inhomogeneously broadened, which is attributed to a spatially varying disorder potential from impurities and defects that weakly localize excitons in the transverse plane of the sample.

A square-root Lorentzian fit to the anti-diagonal lineshape yields a homogeneous linewidth $\gamma = 2.7$ meV corresponding to a dephasing time $T_2 = 250$ fs (in this case, $\sigma \gg \gamma$). Compared to T_1 measurements of the exciton lifetime, T_2 measurements are particularly sensitive to many-body interactions that perturb the phase evolution of optical coherences. Excitation-density dependent measurements of 2D spectra, obtained by increasing the average power of the excitation fields, reveal that the homogeneous linewidth increases by more than a factor of two for an order-of-magnitude increase in the exciton density (see Fig. 9). The density dependence of γ is a clear indication of excitation-induced dephasing (EID) due to exciton–exciton interactions. A comparison of the slope in Fig. 9 to other material systems, such as GaAs and CdTe bulk and quantum well nanostructures, reveals nearly an order of magnitude stronger exciton–exciton interaction. Pronounced EID effects in TMDs are consistent with strong Coulomb interactions due to reduced dielectric screening in two dimensions.

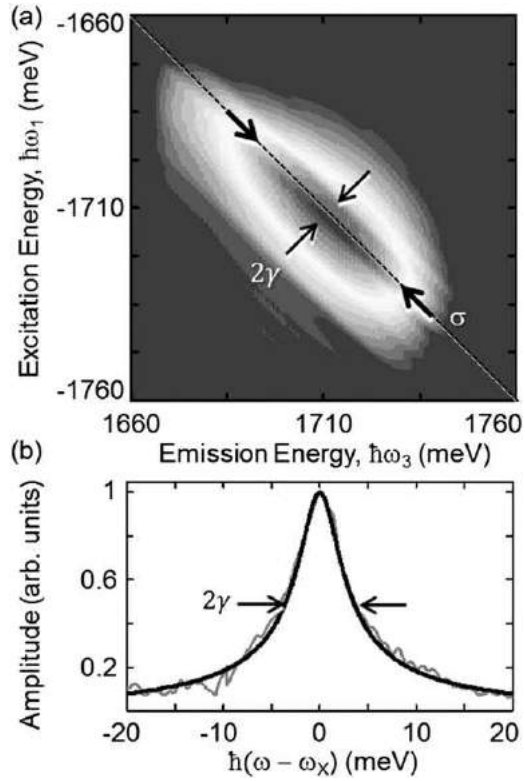

Fig. 8 (a) 2D rephasing spectrum amplitude of excitons in WSe_2 at 10 K. The linewidth along the diagonal dashed line provides a measure of the inhomogeneous broadening (σ) due to static disorder; the linewidth along the anti-diagonal direction provides a measure of the homogeneous linewidth γ. (b) The homogeneous lineshape is fit with a square-root Lorentzian function with a full-width at half-maximum equal to 2γ.

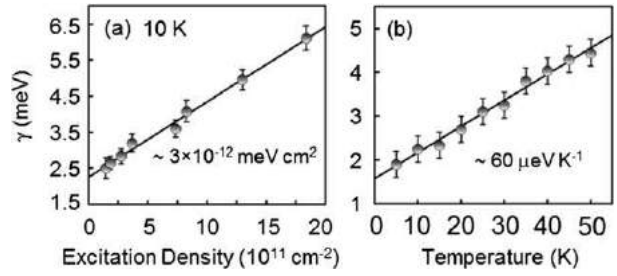

Fig. 9 The homogeneous linewidth of excitons in monolayer WSe_2 measured as a function of excitation density (a) and temperature (b). Extrapolating the linear fits (solid lines) reveals an intrinsic, lifetime-limited linewidth at zero temperature and density of $\gamma = 1.6$ meV ($T_2 = 0.4$ ps).

The role of exciton scattering with phonons is also apparent from the homogeneous broadening with increasing temperature (Fig. 9). The linear increase in γ with temperature T is indicative of exciton dephasing from acoustic phonons with energy much smaller than k_bT. Extrapolation of the linewidth to zero temperature and density reveals a residual dephasing rate $\gamma = 1.6$ meV ($T_2 = 400$ fs), which is more than an order of magnitude smaller than the inhomogeneous linewidth $\sigma \approx 50$ meV.

Additional 2DCS experiments performed by scanning the delay τ_2 instead of τ_1, which are sensitive to the exciton population lifetime (T_1) dynamics, reveal that $T_2 = 2T_1$, i.e. no additional pure dephasing exists in the material at low temperature and density ($\gamma_{ph} = 0$). Thus, exciton dephasing is limited only by the recombination lifetime (sub-picosecond). Short T_2 and T_1 times are consistent with theoretical models of the radiative lifetime, which is predicted to be on the order of ~ 100 fs due to the large exciton oscillator strength.

Coherent and Incoherent Energy Transfer

In doped TMD monolayers, the linear optical spectrum features a peak associated with the trion in addition to the exciton (Fig. 7); however, 2DCS reveals a more complex scenario than just two independent transitions. Fig. 10 shows a time series of the 2D

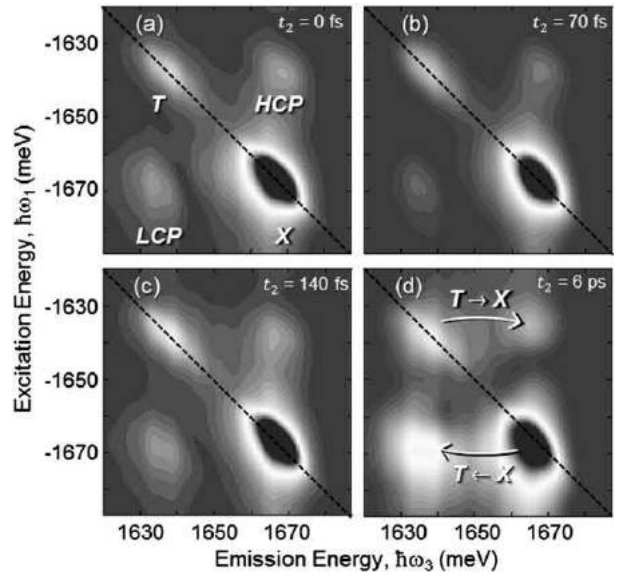

Fig. 10 Normalized 2D rephasing spectrum (amplitude) of monolayer $MoSe_2$ for co-circular polarization at 10 K *versus* delay τ_2 equal to (a) 0 fs, (b) 70 fs, (c) 140 fs, and (d) 6 ps. At short delays (τ_2 less than a few hundred femtoseconds), oscillations of the off-diagonal coupling peaks *HCP* and *LCP* indicate coherent coupling between excitons (*X*) and trions (*T*). At longer delays ($\tau_2 >$ 1ps), an increase in the relative magnitude of *LCP* and *HCP* compared to *X* and *T* indicate incoherent energy and charge transfer between excitons and trions due to trion formation (*LCP*) and dissociation (*HCP*).

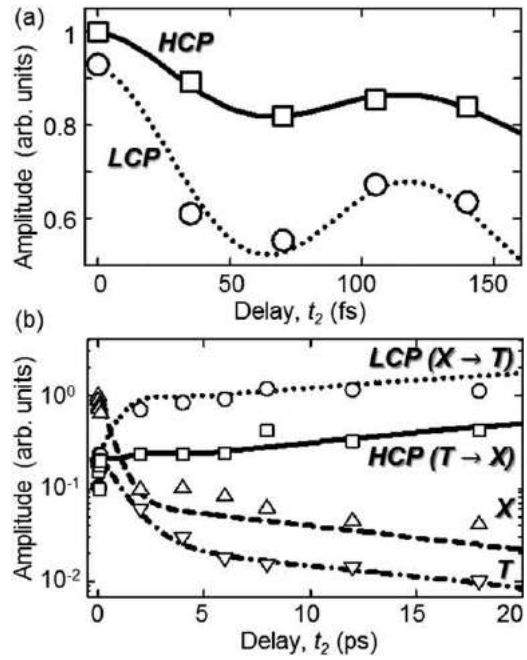

Fig. 11 (a) Oscillations of the cross peaks *HCP* and *LCP versus* delay τ_2 arise from quantum beats between the exciton and trion. The data are modeled with the optical Bloch equations for a four-level diamond system, revealing the dephasing time $\tau_c = 250$ fs and the oscillation period $\tau_{XT} = 130$ fs (corresponding to the exciton–trion energy difference of $\sim$31 meV). (b) On longer timescales, the exciton (*X*) and trion (*T*) peaks exhibit biexponential relaxation dynamics due to radiative and nonradiative recombination and exciton–trion energy transfer. The increase in relative magnitude of the *LCP* and *HCP* (compared to *X* and *T*) indicate exciton-to-trion and trion-to-exciton energy transfer, respectively.

rephasing spectrum from a doped (*n*-type) monolayer $MoSe_2$ at 10 K. At $\tau_2 = 0$ fs, the spectrum features two peaks on the diagonal dashed line that correspond to GSB and ESE of the exciton (*X*) and trion (*T*). The 2D spectrum reveals a new aspect of TMDs not apparent from the linear spectrum – the appearance of off-diagonal peaks (*LCP* and *HCP*), which are evidence of coherent coupling between excitons and trions. Furthermore, the off-diagonal peak amplitudes oscillate as the delay τ_2 is scanned up to a

few hundred femtoseconds. These oscillations can be understood from the fact that fields $\mathscr{E}_1^*$ and $\mathscr{E}_2$ drive the system into a second-order coherence described by $\rho_{XT}=|X\rangle\langle T|$ (*HCP*) and $\rho_{TX}=|T\rangle\langle X|$ (*LCP*). These non-radiative Raman-like coherences oscillate during the second delay τ_2 at the exciton–trion difference frequency (Fig. 11). Modeling the 2D spectra with the OBEs for a four-level diamond system, which enables exciton–trion interactions to be included phenomenologically, yields excellent agreement with the data (solid lines in Fig. 11). An exciton–trion coherence dephasing rate $\gamma_c=2.6$ meV ($\tau_c=250$ fs) and an oscillation period $\tau_{XT}=130$ fs (corresponding to the exciton–trion difference energy of $\sim$31 meV) are obtained from the fits.

On longer timescales after the coherences have decayed ($\tau_2 > 1$ ps), the off-diagonal peaks decay more slowly than the diagonal exciton and trion peaks. The persistence of the coupling peaks is a characteristic signature of incoherent energy transfer between the two states. For example, exciton-to-trion relaxation (equivalently, trion formation) can be described by quantum pathways in which the first two pulses $\mathscr{E}_1^*$ and $\mathscr{E}_2$ drive the system into an excited-state exciton population $\rho_{XX}=|X\rangle\langle X|$. During τ_2, relaxation occurs through $\rho_{XX}\rightarrow\rho_{TT}$, which is monitored by the interaction with the third pulse $\mathscr{E}_3$. Whereas the off-diagonal peaks exhibit a relative enhancement in their amplitudes, the diagonal peaks decay with a biexponential behavior (see Fig. 11).

A global fit to the exciton and trion population relaxation and transfer dynamics is performed using a rate equation model that considers radiative and non-radiative recombination of bright and dark excitons and trions and incoherent energy transfer between them. Results from the model (solid lines in Fig. 11) reveal the dominant population relaxation processes in this sample: (1) $\sim$picosecond radiative recombination and bright-to-dark state scattering for *X* and *T*; (2) exciton-to-trion down-conversion (trion formation) via the exciton binding to a residual electron in $\sim$2.5 ps; and (3) trion-to-exciton up-conversion (trion dissociation) in $\sim$8 ps.

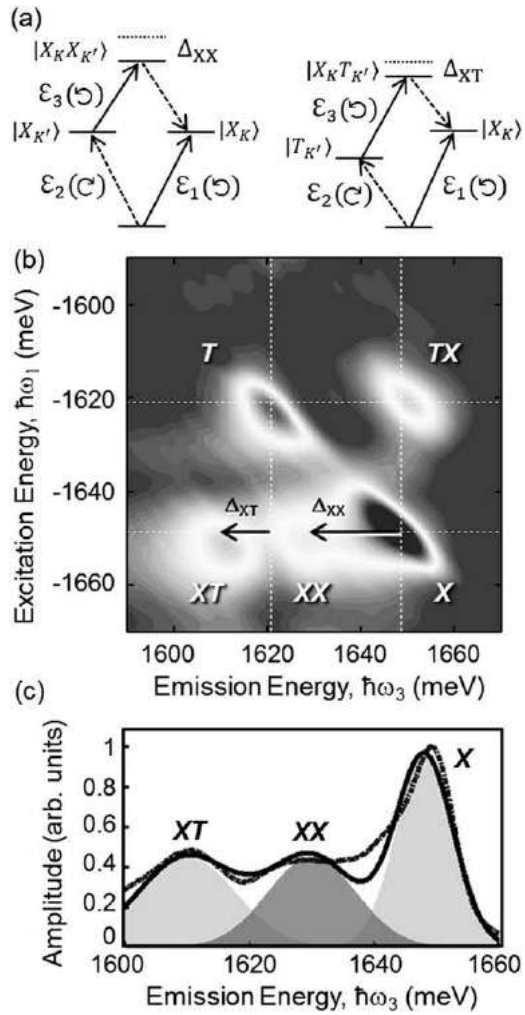

Fig. 12 Neutral and charged biexcitons appear in the 2D spectrum of monolayer $MoSe_2$ at 10 K when using a cross-circularly polarized excitation sequence. The quantum pathways giving rise to the neutral and charged biexciton resonances are shown in the left and right panels of (a), respectively. (b) A normalized 2D rephasing spectrum (magnitude) for the cross-circular polarization sequence. Neutral (*XX*) and charged (*XT*) biexcitons appear red-shifted along the emission energy axis from the exciton and trion emission energies by their corresponding binding energies Δ_{XX} and Δ_{XT}, respectively. (c). A slice along the emission energy axis at the excitation energy of the exciton showing the exciton, neutral biexciton, and charged biexciton. The biexciton binding energies are obtained from Gaussian fit functions.

Intervalley Biexcitons

The strong interactions responsible for tightly bound excitons lead to enhanced interactions between them. Higher-order correlated states, such as a bound biexciton resulting from four-particle correlations between two electrons and two holes, are expected to be stable in TMDs. These states are accessible with one-dimensional techniques such as photoluminescence and pump probe spectroscopy; however, signatures of biexcitons are difficult to distinguish from other phenomena, such as excitons localized by impurities or defects, exciton–electron interactions, and exciton renormalization, which can all spectrally overlap in one-dimensional spectra and exhibit a similar optical response.

Optical 2DCS is particularly sensitive to exciton many-body interactions, since congested one-dimensional spectra are unraveled onto two frequency axes. Signatures of biexcitons appear in the 2D rephasing spectrum of monolayer $MoSe_2$ (**Fig. 12**) when using a cross-circular polarization scheme, i.e. fields $\mathscr{E}_1^*$ and $\mathscr{E}_3$ ($\mathscr{E}_2$ and $\mathscr{E}_s$) are left (right) circularly polarized. Compared to the co-circular spectra shown in **Fig. 10**, the new peaks *XX* and *XT* are ascribed to the neutral and charged biexciton, respectively. For example, the quantum pathway associated with the biexciton peak *XX* can be interpreted from a four-level diamond system with ground state $|g\rangle$, singly excited states $|K\rangle$ and $|K'\rangle$, and doubly excited state $|XX\rangle$: field $\mathscr{E}_1^*$ drives a coherence $\rho_{gk}=|g\rangle\langle K|$ between the ground and *K*-valley exciton states during τ_1; field $\mathscr{E}_2$ drives the system into an excited-state $\rho_{K'K}=|K'\rangle\langle K|$; field $\mathscr{E}_3$ generates a third-order coherence $\rho_{XX,K}=|XX\rangle\langle K|$ during τ_3, which radiates as the FWM signal. The system evolves during τ_1 and τ_3 at energies $\hbar\omega_X$ and $\hbar\omega_X-\Delta_{XX}$, where Δ_{XX} is the biexciton binding energy. In the Fourier domain, this pathway appears as peak *XX*; a similar pathway describes the bound charged biexciton state with binding energy Δ_{XT}.

A horizontal slice from the spectrum along the emission energy axis at the exciton excitation energy is shown in **Fig. 12**. The data are fit with a triple Gaussian function, which provides approximate binding energies of $\Delta_{XX}=20$ meV and $\Delta_{XT}=5$ meV. These values are consistent with current predictions from microscopic calculations. It is worth noting that information about these many-body states are masked in a 1D pump-probe spectrum by the trion resonance (projection of the 2D spectrum onto the emission axis); thus, while all the details of incoherent interactions in the sample are accessible with pump-probe spectroscopy, numerous quantum pathways spectrally overlap making analysis and interpretation difficult.

Summary

This article provides an overview of the basic principles of two-dimensional coherent spectroscopy and its implementation to study monolayer TMDs. In just the last couple of years, 2DCS has enhanced our understanding of the optical and electronic properties of 2D semiconductors, including: (1) the intrinsic homogeneous linewidth is on the order of 1 meV and is lifetime-limited for the exciton, whereas significant pure dephasing exists for the trion; (2) the exciton, trion, and background free carriers in doped TMDs form a coherently coupled system; (3) energy transfer between states is an important non-radiative relaxation channel; (4) high-order correlated states are stable in the monolayer TMD $MoSe_2$ and only exist between excitons/trions in opposite valleys. Compared to other semiconductor systems, TMDs offer novel opportunities for exploring quantum phenomena including condensation of exciton–polaritons, ultrathin biexciton lasers, and polarization-entangled photon sources.

The field of two-dimensional materials is extremely active with new insights reported almost daily. The ability to isolate and stack different 2D materials with precision, transfer them to nanopatterned substrates, and incorporate them with photonic microcavities and plasmonic nanostructures provide unprecedented possibilities for tailoring their physical properties for novel opto-electronic, photonic, and coherent valley/spintronic applications. Optical 2DCS will likely play a role in understanding and characterizing TMD-based heterostructures and nanophotonics in the same way that it has impacted our understanding of optical phenomena in monolayers.

See also: Two-Dimensional Electronic Spectroscopy

Further Reading

Berkelbach, T.C., Hybertsen, H.S., Reichman, D.R., 2013. Theory of neutral and charged excitons in monolayer transition metal dichalcogenides. Physical Review B 88, 045318.
Cundiff, S.T., Mukamel, S., 2013. Optical multidimensional coherent spectroscopy. Physics Today 66, 44.
Czech, K.J., Thompson, B.J., Kain, S., *et al.*, 2015. Measurement of ultrafast excitonic dynamics of few-layer MoS_2 using state-selective coherent multidimensional spectroscopy. ACS Nano 9, 12146–12157.
Dey, P., Paul, J., Wang, Z., *et al.*, 2016. Optical coherence in atomic-monolayer transition-metal dichalcogenides limited by electron–phonon interactions. Physical Review Letters 116, 127402.
Ernst, R.R., Bodenhausen, G., Wokaun, A., 1987. Principles of Nuclear Magnetic Resonance in One and Two Dimensions. Oxford: Oxford University Press.
Hao, K., Xu, L., Nagler, P., *et al.*, 2016. Coherent and incoherent coupling dynamics between neutral and charged excitons in monolayer $MoSe_2$. Nano Letters 16, 5109–5113.
Hao, K., Xu, L., Specht, J.F., *et al.*, 2017. Neutral and charged intervalley-biexcitons in monolayer $MoSe_2$. Nature Communications 8, 15552.
Jakubczyk, T., Delmonte, V., Koperski, M., *et al.*, 2016. Radiatively limited dephasing and exciton dynamics in $MoSe_2$ monolayers revealed with four-wave mixing microscopy. Nano Letters 16, 5333–5339.
Moody, G., Dass, C.K., Hao, K., *et al.*, 2015. Intrinsic homogeneous linewidth and broadening mechanisms of excitons in monolayer transition metal dichalcogenides. Nature Communications 6, 8315.
Mukamel, S., 2000. Multidimensional femtosecond correlation spectroscopies of electronic and vibrational excitations. Annual Review of Physical Chemistry 51, 691–729.
Xu, X., Yao, W., Xiao, X., Heinz, T.F., 2014. Spins and pseudospins in layered transition metal dichalcogenides. Nature Physics 10, 343–350.

Excitons in Magnetic Fields

Kankan Cong, G Timothy Noe II, and Junichiro Kono, Rice University, Houston, TX, United States

Introduction

When a photon of energy greater than the band gap is absorbed by a semiconductor, a negatively charged electron is excited from the valence band into the conduction band, leaving behind a positively charged hole. The electron can be attracted to the hole via the Coulomb interaction, lowering the energy of the electron-hole (*e-h*) pair by a characteristic binding energy, E_b. The bound *e-h* pair is referred to as an exciton, and it is analogous to the hydrogen atom, but with a larger Bohr radius and a smaller binding energy, ranging from 1 to 100 meV, due to the small reduced mass of the exciton and screening of the Coulomb interaction by the dielectric environment.

Like the hydrogen atom, there exists a series of excitonic bound states, which modify the near-band-edge optical response of semiconductors, especially when the binding energy is greater than the thermal energy and any relevant scattering rates. When the *e-h* pair has an energy greater than the binding energy, the electron and hole are no longer bound to one another (ionized), although they are still correlated. The nature of the optical transitions for both excitons and unbound *e-h* pairs depends on the dimensionality of the *e-h* system. Furthermore, an exciton is a composite boson having integer spin that obeys Bose-Einstein statistics rather than fermions that obey Fermi-Dirac statistics as in the case of either the electrons or holes by themselves. One of the most interesting consequences of the bosonic nature of excitons is the possibility of having many that occupy the same quantum state as opposed to fermions, which obey the Pauli exclusion principle.

An applied magnetic field, $\vec{B}$, will quantize the energies and angular momenta of charged particles and thus change the electronic state of an exciton. In general, the excitonic wavefunction shrinks in the direction perpendicular to $\vec{B}$, and the exciton binding energy increases with $B = |\vec{B}|$. How the energy of each exciton state varies with B depends on basic exciton parameters, and, therefore, the study of excitons in B provides a tool to characterize various exciton properties, such as the reduced mass, binding energy, Bohr radius, and g-factor. Importantly, the nature of the dependence of exciton energy on B depends on the strength of B. For example, at low B, where the effect of the Coulomb interaction is larger than that of B and thus B is treated as a perturbation, excitons retain a hydrogenic nature: diamagnetic shifts and Zeeman splittings modify the energies of excitonic states. Alternatively, in the high B limit where the effect of B dominates and the Coulomb interaction is treated as a perturbation, the energies of exciton states show Landau-level-like, linear-B dependence, similar to free *e-h* pairs.

For the hydrogen atom, the binding energy, 13.6 eV, is much higher than those for excitons in semiconductors, and an extremely high B (over 2×10^5 T) is required to enter the high B regime. The much smaller binding energies of excitons in semiconductors allow one to enter the high-B regime readily, which, together with the bosonic nature of excitons, makes excitons in high B attractive for the investigation of many-body phenomena in condensed matter in a highly controllable manner. A large body of theoretical work has thus been devoted in the last several decades to *e-h* systems in strong B, and a fascinating array of possible physical phenomena/states has been predicted – the excitonic insulator phase, a gas-liquid-type phase transition, Bose-Einstein condensation of magnetoexcitons, and quantum chaos. Recently, attention has been paid to spatially separated two-dimensional (2D) electrons and holes in layered structures, since the pioneering work of Lozovik and co-workers, who considered the pairing of electrons and holes across an interface of two-media, and the superfluidity of such pairs. A very interesting but complicated many-particle situation arises, in the presence of strong perpendicular B. In such a situation, spatially separated 2D electrons and holes, both Landau quantized, are present, and there is a Coulomb attraction between them as well as electron-electron (*e-e*) and hole-hole (*h-h*) correlations. However, all these interactions are expected to cancel each other in the high-B limit, where the "hidden symmetry" of 2D magnetoexcitons protects them from ionization, leading to the absence of an excitonic Mott transition. Furthermore, excitonic gain enhancement occurs in a high-density *e-h* system in high B, which leads to cooperative and coherent light emission (known as superfluorescence) from the highest-occupied magnetoexciton state.

Here, we will review various observations related to excitons in magnetic fields. First, we present the theory of excitons in magnetic fields and show how the exciton energy's magnetic field dependence relates to fundamental exciton parameters. We will explore how different dimensionalities of excitons and the strength of magnetic fields affect the exciton properties. Then, interband and intraband magneto-optical absorption observations are discussed in different semiconductor systems. Finally, some exotic phenomena unique to high-density excitons in high magnetic fields are presented.

Exciton States in Magnetic Fields

In this section, we describe the states of excitons in semiconductors in different dimensions and B regimes. We introduce the basic Hamiltonians of excitons, discuss the eigenenergies in the low-B and high-B limits, and define some basic parameters of excitons and various quantities to be used throughout the article.

For an electron with effective mass m_e^* located at $\vec{r}_e$ and a hole with effective mass m_h^* located at $\vec{r}_h$, interacting through the Coulomb potential $U(\vec{r}) = \frac{-e^2}{4\pi\varepsilon r}$, the Hamiltonian at zero B is given by

$$\hat{H} = -\frac{\hbar^2}{2M^*}\nabla_R^2 - \frac{\hbar^2}{2\mu^*}\nabla_r^2 + U(\vec{r}) \tag{1}$$

where $M^* = m_e^* + m_h^*$ is the total mass, $\mu^* = \left((m_e^*)^{-1} + (m_h^*)^{-1}\right)^{-1}$ is the reduced mass, the center-of-mass coordinate $\vec{R} = (m_e^*\vec{r}_e + m_h^*\vec{r}_h)/M^*$, the relative coordinate $\vec{r} = \vec{r}_e - \vec{r}_h$, ∇_R^2 $\left(\nabla_r^2\right)$ is the Laplacian expressed in the center-of-mass (relative) coordinates, $\hbar$ is the reduced Planck constant, e is the electronic charge, $\varepsilon = \varepsilon_r\varepsilon_0$ is the permittivity, ε_r is the relative permittivity (dielectric constant), and ε_0 is the vacuum permittivity.

Three-Dimensional Case

In a three-dimensional (3D) system at zero B, the Schrödinger equation for the relative motion is given by

$$\left(-\frac{\hbar^2}{2\mu^*}\nabla_r^2 - \frac{e^2}{4\pi\varepsilon r}\right)\psi(\vec{r}) = E\psi(\vec{r}) \tag{2}$$

where $\Psi(\vec{r})$ is the exciton wavefunction and E is the energy of the exciton. In spherical coordinates, the wavefunction can be expressed as $\Psi(r,\theta,\phi) = R_{nl}(r)Y_{lm}(\theta,\phi)$, where $R_{nl}(r)$ and $Y_{lm}(\theta, \phi)$ are the radial and angular wavefunctions, respectively, n ($=1, 2, 3, \ldots$) is the principal quantum number, l ($=0, 1, 2, \ldots, n-1$) is the azimuthal quantum number, and m ($=0, \pm1, \ldots, \pm l$) is the magnetic quantum number. The eigenenergies are

$$E_n = -\frac{R_\gamma^*}{n^2},\ n = 1, 2, 3, \ldots \tag{3}$$

where R_γ^* is the effective Rydberg constant defined as

$$R_\gamma^* \doteq \frac{\mu^* e^4}{32\pi^2\varepsilon^2\hbar^2} = \frac{\mu^*/m_e}{\varepsilon_r^2} \times 13.6\ \text{eV} \tag{4}$$

where $m_e = 9.11 \times 10^{-31}$ kg is the mass of a free electron in vacuum. Therefore, the binding (ionization) energy of a 3D exciton is $E_{b,3D}^* = E_\infty - E_1 = R_\gamma^*$. Using the effective Bohr radius

$$a_{B,3D}^* \doteq \frac{4\pi\varepsilon\hbar^2}{\mu^* e^2} = \frac{\varepsilon_r}{\mu^*/m_e} \times 0.0529\ \text{nm} \tag{5}$$

the binding energy can be expressed as

$$E_{b,3D}^* = \frac{\hbar^2}{2\mu^*}\frac{1}{(a_{B,3D}^*)^2} \tag{6}$$

and the 1s state wavefunction is given by

$$\Psi_{1s,3D} = \Psi_{100} = \frac{1}{\sqrt{\pi}}\frac{1}{\left(a_{B,3D}^*\right)^{3/2}}\exp\left(\frac{-r}{a_{B,3D}^*}\right) \tag{7}$$

Now that we have defined the basic quantities for excitons, let us consider the effect of a uniform and constant magnetic field, $\vec{B} = (0, 0, B)$. The excitonic eigenenergies and wavefunctions are modified through the inclusion of vector potential $\vec{A}$ in the Hamiltonian, for which $\vec{B} = \nabla \times \vec{A}$. The Hamiltonian of the system becomes (neglecting electron spin)

$$\begin{aligned}\hat{H} &= \frac{1}{2m_e^*}\left(\vec{p}_e + e\vec{A}(\vec{r}_e)\right)^2 + \frac{1}{2m_h^*}\left(\vec{p}_h - e\vec{A}(\vec{r}_h)\right)^2 - \frac{e^2}{4\pi\varepsilon r} \\ &= -\frac{\hbar^2}{2m_e^*}\nabla_e^2 - \frac{\hbar^2}{2m_h^*}\nabla_h^2 - \frac{e^2}{4\pi\varepsilon r} - \frac{ie\hbar}{m_e^*}\vec{A}(\vec{r}_e)\cdot\vec{\nabla}_e \\ &\quad + \frac{ie\hbar}{m_h^*}\vec{A}(\vec{r}_h)\cdot\vec{\nabla}_h + \frac{e^2}{2m_e^*}A^2(\vec{r}_e) + \frac{e^2}{2m_h^*}A^2(\vec{r}_h)\end{aligned} \tag{8}$$

where $\vec{p}_e$ and $\vec{p}_h$ are the momentum operators of the electron and hole, respectively.

Following Gor'kov and Dzyaloshinskii, we use a canonical transformation to obtain a new wavefunction $\Psi'\left(\vec{R}, \vec{r}\right)$

$$\Psi'\left(\vec{R}, \vec{r}\right) = \exp\left(i\left[\vec{K} - \frac{e}{\hbar}\vec{A}(\vec{r})\right]\cdot\vec{R}\right)F(\vec{r}) \tag{9}$$

where $F(\vec{r})$ is an arbitrary function of $\vec{r}$, and $\vec{K} = \vec{k}_e - \vec{k}_h$. After the transformation, the Schrödinger equation becomes

$$\hat{H}'F(\vec{r}) = \left[E - \frac{\hbar^2K^2}{2M^*}\right]F(\vec{r}) \tag{10}$$

with the new Hamiltonian expressed as

$$\hat{H}' = \left(-\frac{\hbar^2}{2\mu^*}\nabla_r^2 - \frac{e^2}{4\pi\varepsilon r}\right) - ie\hbar\left(\frac{1}{m_e^*} - \frac{1}{m_h^*}\right)\vec{A}(\vec{r})\cdot\vec{\nabla}_r + \frac{e^2}{2\mu^*}A^2(\vec{r}) - \frac{2e\hbar}{M^*}\vec{A}(\vec{r})\cdot\vec{K}$$

$$= \hat{H}_{B=0} + \frac{\omega_{c,e} - \omega_{c,h}}{2}\hat{L}_z + \frac{e^2B^2}{8\mu^*}(x^2+y^2) \tag{11}$$

$$-\frac{eB\hbar}{M^*}(xK_y - yK_x) \tag{12}$$

where $\omega_{c,i} = eB/m_i^*$ $(i=e,h)$ is the cyclotron frequency, $\hat{L}_z = -i\hbar\left(x\frac{\partial}{\partial y} - y\frac{\partial}{\partial x}\right)$ is the z-component of the orbital angular momentum operator, and the symmetric gauge, $\vec{A} = (1/2)(-By, Bx, 0)$, was used in the final step. The first term of $\hat{H}'$ is the Hamiltonian at zero B. The second term is referred to as the (orbital) Zeeman term, which leads to shifts and splittings of exciton states, depending on the value and degeneracy of the orbital angular momentum of a given state. The third term is the Langevin diamagnetism term, which is proportional to the spatial extent of the exciton wavefunction in a plane perpendicular to $\vec{B}$ and increases quadratically with increasing B. The fourth term can be rewritten as $-e(\vec{V}_R \times \vec{B})\cdot\vec{r}$, where $\vec{V}_R$ is the center-of-mass velocity. Since $\vec{r}$ is the relative coordinate, the fourth term implies that the relative and center-of-mass degrees of freedom are coupled in the presence of a B.

It should be pointed out that one cannot obtain an exact solution to the Schrödinger equation corresponding to Eq. (11), which is related to quantum chaos. This is due to the fact that the number of conserved quantities (two), i.e., E and L_z, is smaller than the number of degrees of freedom (three), a situation known to be *non-integrable*, for which the classical trajectories of energy states in phase space are chaotic. Correspondingly, energy levels exhibit fluctuations; with increasing energy, the statistics of levels evolve from Poissonian to Gaussian. Approximate solutions can be obtained only for low-energy levels either in the low-B limit or the high-B limit, and a one-to-one correspondence of states between these two limits does not exist, in general, because of the different symmetries the magnetic field and the Coulomb potential possess: the cylindrical symmetry (for the former) and the spherical symmetry (for the latter).

To define the low-B and high-B limits, we introduce a dimensionless parameter

$$\gamma \doteq \frac{\hbar\omega_c}{2}/R_\gamma^* = \left(\frac{a_B^*}{l_B}\right)^2 = \frac{16\pi^2\hbar^3\varepsilon^2}{\mu^{*2}e^3}B \tag{13}$$

where $l_B = \sqrt{\hbar/eB}$ is the magnetic length and $\omega_c = eB/\mu^*$ is the reduced cyclotron frequency. The value of γ is a measure of whether the energy of the exciton is dominated by the influence of the magnetic field or the Coulomb interaction. In the low-B limit ($\gamma \ll 1$), the system is nearly hydrogenic, and the Zeeman and diamagnetic terms can be treated as a perturbation in Eq. (11). (Note that the fourth term in $\hat{H}'$ is usually negligible because of the smallness of K relevant to optical transitions.) For the 1s [or $(n,l,m)=(100)$] state, $\langle L_z\rangle = m\hbar = 0$, and therefore, the Zeeman term is zero, and the first-order correction is the diamagnetic term:

$$\Delta E_{1s,3D} = \left\langle \frac{e^2B^2}{8\mu^*}(x^2+y^2)\right\rangle_{1s,3D} = \frac{e^2B^2}{8\mu^*}\langle\Psi_{1s,3D}|(x^2+y^2)|\Psi_{1s,3D}\rangle = \frac{e^2B^2}{8\mu^*}\cdot 2\left(a_{B,3D}^*\right)^2 = \sigma_{3D}^* B^2 \tag{14}$$

$$\sigma_{3D}^* \doteq \frac{e^2}{4\mu^*}\left(a_{B,3D}^*\right)^2 = \frac{4\pi^2\varepsilon^2\hbar^4}{e^2(\mu^*)^3} \tag{15}$$

Here, σ_{3D}^* is the 3D diamagnetic-shift coefficient, which depends on μ^* and $a_{B,3D}^*$. This fact offers a straightforward experimental method for determining $a_{B,3D}^*$ (or, more generally, the spatial extent of the exciton wavefunction) once the reduced mass is known. In addition, measuring σ_{3D}^* as the direction of $\vec{B}$ is varied allows one to map out the wavefunction anisotropy. Similarly, the diamagnetic shift for the ns exciton state can be calculated to be

$$\Delta E_{ns,3D} = \frac{n^2(5n^2+1)}{6}\sigma_{3D}^* B^2 \tag{16}$$

For calculating the energy for a p-like state, one must include the Zeeman term. For example, the 2p state splits into three states in a B with $\langle L_z\rangle = 0, \pm 1$, so the orbital Zeeman energy is

$$\Delta E_{\text{Zeeman}} = 0, \pm\frac{\hbar\omega_c}{2} \tag{17}$$

In the high-B limit ($\gamma \gg 1$), where the effect of the magnetic field is much larger than the Coulomb interaction, the Coulomb interaction is treated as a small perturbation. Under this condition, the energies and wavefunctions of excitons can be separated into different components for the x-y plane and the z-direction (known as the 'adiabatic' approximation). In the x-y plane, only the effect of the magnetic field has to be considered, while in the z-direction, only the Coulomb interaction contributes, because a magnetic field applied in the z-direction exerts a Lorentz force only in the x-y plane. The wavefunction can then be expressed as

$$\Psi_{NMi} = \Phi_{NM}(x,y) f_{NMi}(z) \tag{18}$$

where $N = 0, 1, 2, \ldots$ is the Landau quantum number, $M = N, N-1, \ldots, -\infty$ is the azimuthal quantum number, and $i = 0, 1, 2, \ldots$ is the quantum number associated with the motion in the z-direction. The wavefunction $\Phi_{NM}(x,y)$ describes the in-plane motion and is proportional to the associated Laguerre function $L_{N+|M|}^{|M|}$.

The total energy is characterized by three quantum numbers (N,M,v^+) for $i = 2v (v = 0, 1, 2, \ldots)$, and (N,M,v^-) for $i = 2v-1$ $(v = 1, 2, \ldots)$:

$$E_{NM(v)} = \left(N + \frac{1}{2}\right)\hbar\omega_c + \Delta E_{NMv}(z) \tag{19}$$

Hence, in the high-B limit, the states become more Landau-level-like, exhibiting a linear B dependence (the first term) with the Coulombic correction, $\Delta E_{NMv}(z)$, which is a red-shift for each eigenenergy at a fixed value of z. The wavefunction $f_{NMv^+}(z)$ is an even function, while $f_{NMv^-}(z)$ is an odd function with respect to z.

Because of the reason stated earlier, there is no general correspondence between the low-field notation (n,l,m) and high-field notation (N,M,v) for a 3D exciton. However, using the principle of conservation of the number of nodal surfaces, Shinada *et al.* made the following correspondence for the lowest-lying exciton states:

$$\begin{aligned} &1s \leftrightarrow (0,0,0^+)\\ &2s \leftrightarrow (0,0,1^+)\\ &2p_0 \leftrightarrow (0,0,1^-)\\ &2p_+ \leftrightarrow (1,1,0^+)\\ &2p_- \leftrightarrow (0,-1,0^+)\\ &3d_0 \leftrightarrow (1,0,0^+) \end{aligned} \tag{20}$$

Two-Dimensional Case

For 2D excitons, we use a polar coordinate system,

$$\overrightarrow{\rho} = (x,y) = (\rho\cos\phi, \rho\sin\phi) \tag{21}$$

At zero B, the Hamiltonian for the relative motion for a 2D e-h pair is written as

$$\begin{aligned} \hat{H} &= -\frac{\hbar^2}{2\mu^*}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) - \frac{e^2}{4\pi\varepsilon\sqrt{x^2+y^2}}\\ &= -\frac{\hbar^2}{2\mu^*}\left(\frac{\partial^2}{\partial\rho^2} + \frac{1}{\rho}\frac{\partial}{\partial\rho} + \frac{1}{\rho^2}\frac{\partial^2}{\partial\phi^2}\right) - \frac{e^2}{4\pi\varepsilon\rho} \end{aligned} \tag{22}$$

By solving the Schrödinger equation

$$\hat{H}\Psi(\rho,\phi) = E\Psi(\rho,\phi) \tag{23}$$

we obtain the eigenenergies as

$$E_n = \frac{\mu^* e^4}{32\pi^2\varepsilon^2\hbar^2\left(n-\frac{1}{2}\right)^2} = -\frac{R_y^*}{\left(n-\frac{1}{2}\right)^2},\ n = 1,2,3,\ldots \tag{24}$$

The binding energy is then calculated to be

$$E_{b,2D}^* = E_\infty - E_1 = 4R_y^* = 4E_{b,3D}^* \tag{25}$$

which is four times larger than that of the 3D case for the same semiconductor. The eigenstates have the form

$$\Psi_{nm} = R_{nm}(\rho)e^{im\phi}$$

where n $(= 1, 2, 3, \ldots)$ is the principal quantum number, and m $(= -n+1, \ldots, -2, -1, 0, 1, 2, \ldots, n-1 = \ldots, d_-, p_-, s, p_+, d_+, \ldots)$ is the angular momentum quantum number. The $1s$ exciton wavefunction is written as

$$\Psi_{1s,2D} = \Psi_{10} = R_{10} = \frac{1}{\sqrt{2\pi}}\frac{2}{a_{B,2D}^*}\exp\left(\frac{-\rho}{a_{B,2D}^*}\right) \tag{26}$$

with the effective Bohr radius,

$$a^*_{B,2D} = \frac{2\pi\varepsilon\hbar^2}{\mu^* e^2} = \frac{a^*_{B,3D}}{2} \tag{27}$$

which is only half of that of a 3D exciton in the same semiconductor material. From Eqs. (4), (25) and (27), we have

$$E^*_{b,2D} = \frac{\hbar^2}{2\mu^*}\frac{1}{(a^*_{B,2D})^2} \tag{28}$$

In a similar manner to the 3D case, the Hamiltonian of the 2D system at a finite B after the Gor'kov-Dzyaloshinskii transformation can be written as

$$\begin{aligned}\hat{H}' &= \left(-\frac{\hbar^2}{2\mu^*}\nabla^2_\rho - \frac{e^2}{4\pi\varepsilon\rho}\right) - ie\hbar\left(\frac{1}{m^*_e} - \frac{1}{m^*_h}\right)\vec{A}(\vec{\rho})\cdot\vec{\nabla}_\rho \\ &\quad + \frac{e^2}{2\mu^*}A^2(\vec{\rho}) - \frac{2e\hbar}{M}\vec{A}(\vec{\rho})\cdot\vec{K}\end{aligned} \tag{29}$$

$$\begin{aligned}&= \hat{H}_{B=0} + \frac{\omega_{c,e} - \omega_{c,h}}{2}\hat{L}_z + \frac{e^2B^2}{8\mu^*}(x^2 + y^2) \\ &\quad - \frac{eB\hbar}{M}(xK_y - yK_x)\end{aligned} \tag{30}$$

where the symmetric gauge is used. Also, similar to the 3D case, the diamagnetic shift for the 1s state can be calculated to be

$$\begin{aligned}\Delta E_{1s,2D} &= \frac{e^2B^2}{8\mu^*}\langle\rho^2\rangle_{1s,2D} = \frac{e^2B^2}{8\mu^*}\frac{3}{8}\left(a^*_{B,3D}\right)^2 \\ &= \sigma^*_{2D}B^2\end{aligned} \tag{31}$$

$$\begin{aligned}\sigma^*_{2D} &= \frac{3e^2}{64\mu^*}\left(a^*_{B,3D}\right)^2 = \frac{3e^2}{16\mu^*}\left(a^*_{B,2D}\right)^2 \\ &= \frac{3\pi^2\varepsilon^3\hbar^4}{4e^2(\mu^*)^3} = \frac{3}{16}\sigma^*_{3D}\end{aligned} \tag{32}$$

$$a^*_{B,2D} = \sqrt{\langle\rho^2\rangle_{1s,2D}} = \sqrt{8\mu^*\sigma^*_{2D}}/e \tag{33}$$

Here, σ^*_{2D} is the 2D diamagnetic-shift coefficient, and $a^*_{B,2D}$ is equal to the root mean squared (r.m.s.) radius of the 1s exciton. So, the diamagnetic-shift coefficient of a 2D exciton is much smaller than the corresponding 3D excitons in the same semiconductor material, i.e., a larger Coulomb interaction makes the wavefunction more compact, leading to a smaller diamagnetic shift.

In a 2D system, both the B and the Coulomb energy have circular symmetry, and thus, there is a straightforward, one-to-one correspondence between the low-B notation (n,m) and the high-B notation (N,M), as summarized in **Table 1**. The correspondence rules are:

$$m = N - M \tag{34}$$

$$n = N + 1\,(\text{for}\,N \geq M) \tag{35}$$

$$n = M + 1\,(\text{for}\,M \geq N) \tag{36}$$

Here, n and m are the principle and angular momentum quantum numbers, respectively, in the low-B hydrogenic notation, while M and N are the Landau indices of the electron and hole, respectively, in the high-B Landau-level notation. For example, if $N=0$ and $M=0$, we get $n=1$ and $m=0$, i.e., $(N,M)=(0,0)$ corresponds to the 1s state. Also, if $N=3$ and $M=2$, we obtain $n=4$ and $m=+1$, and so $(N,M)=(3,2)$ corresponds to the $4p^+$ state.

Table 1 Correspondence between the low- and high-B labels $(n,m)\leftrightarrow(N,M)$ for 2D magnetoexcitons, where $m=N-M$ and $n=N+1$ for $N\geq M$ and $n=M+1$ for $M\geq N$

M	*0*	*1*	*2*	*3*
N				
0	1s	$2p^-$	$3d^-$	$4f^-$
1	$2p^+$	2s	$3p^-$	$4d^-$
2	$3d^+$	$3p^+$	3s	$4p^-$
3	$4f^+$	$4d^+$	$4p^+$	4s

Source: Reproduced (adapted) with permission from MacDonald, A.H., Ritchie, D.S., 1986. Hydrogenic energy levels in two dimensions at arbitrary magnetic fields. Physical Review B 33, 8336–8344. Copyright 1986 by the American Physical Society.

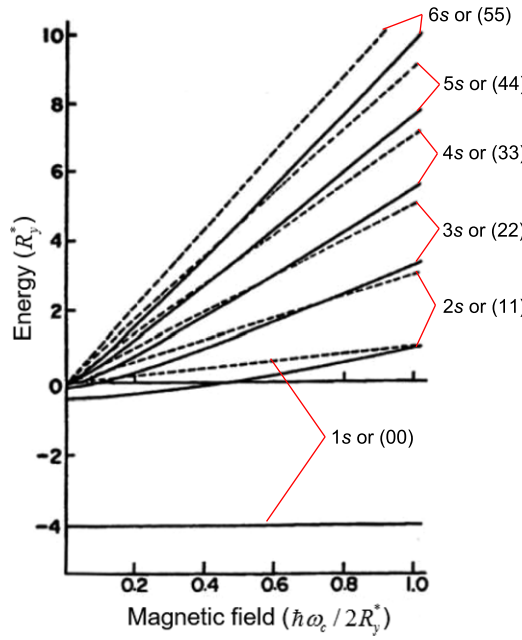

Fig. 1 (color online). Calculated eigenenergies of 2D excitons as a function of dimensionless magnetic field, $\gamma = \hbar\omega_c/2Ry^*$, with (solid lines) and without (dashed lines) the Coulomb interaction. Reproduced (adapted) with permission from Akimoto, O., Hasegawa, H., 1967. Interband optical transitions in extremely anisotropic semiconductors. II . Coexistence of exciton and the Landau levels. Journal of the Physical Society of Japan 22, 181–191. Copyright 1967 by the Physical Society of Japan.

Fig. 1 plots the eigenenergies of the lowest six 2D exciton states as a function of dimensionless magnetic field γ, calculated by Akimoto and Hasegawa. The solid and dashed lines represent the energies of the exciton states with and without including the Coulomb interaction in the Hamiltonian, respectively. At low B, the excitonic feature dominates the low-index energy states. The high-index energy states and the states at high fields show a more Landau-level-like behavior, more appropriate for free electrons and holes. The lowest-energy state or the 1s [or the (00)] state remains constant with increasing γ up to $\sim$1, but in the higher γ range it curves upward to cross the $E=0$ line and eventually runs parallel to the linear Landau level line at $\gamma \sim 12.5$.

Interband Magneto-optical Absorption

Interband magneto-optical absorption can provide direct information on the energy levels of optically accessible (or 'bright') exciton states: ns states (1s, 2s, 3s, …) in the low-B notation, or $(N,M)=(00)$, (11), (22), … states in the high-B notation. To be clear, these are simply notations, and there is a one-to-one correspondence between the two (see Table 1). In the low-B limit, the 1s energy state shows a hydrogenic character with the energy quadratically dependent on B. The effective diamagnetic-shift coefficient, σ^*, can be determined from the B^2 fitting of the exciton energy. If the reduced mass of excitons is known, the Bohr radius can be obtained from Eq. (33). In addition, the binding energy can be determined using Eq. (28), assuming that a hydrogenic model applies to the semiconductor under study. The reduced effective mass can be obtained from the cyclotron frequency $\omega_c = eB/\mu^*$, which is deduced from the slope versus B of exciton energies in the high-B limit. Furthermore, by extrapolating the B-dependent energy lines of excited exciton states in the high-B limit to the zero-B value, it is found that they nearly converge to a single point, which gives the band gap. The difference between the band gap and the 1s exciton peak at zero B then gives the exciton binding energy.

Quantum Wells

The first magneto-optical absorption experiment in a semiconductor quantum well system was performed by Tarucha *et al.* in 1984. They used undoped GaAs/AlAs multiple-quantum wells grown on a (100) GaAs substrate by molecular beam epitaxy. Optical absorption spectra were taken at 4.2 K in pulsed B up to 30 T applied perpendicular to the sample. Fig. 2(a) shows absorption spectra at different B. At 0 T, the absorption spectrum shows two distinct peaks, corresponding to the 1s heavy-hole ($E_{hh}(1s)$) and 1s light-hole ($E_{lh}(1s)$) excitons, respectively. With increasing B, the 1s exciton peaks were observed to blue-shift, and additional peaks emerged at higher energies, corresponding to the ns states (2s, 3s, …) in the low-B notation, or $(N,M)=(11)$, (22), … states in the high-B notation. These higher-energy peaks blue-shifted more rapidly with increasing B.

Fig. 2(b) shows the photon energies of the absorption peaks as a function of B. The 1s heavy-hole and light-hole excitons are represented by solid and open circles, respectively, and the peaks of higher energy states are plotted with crosses. In the low-B

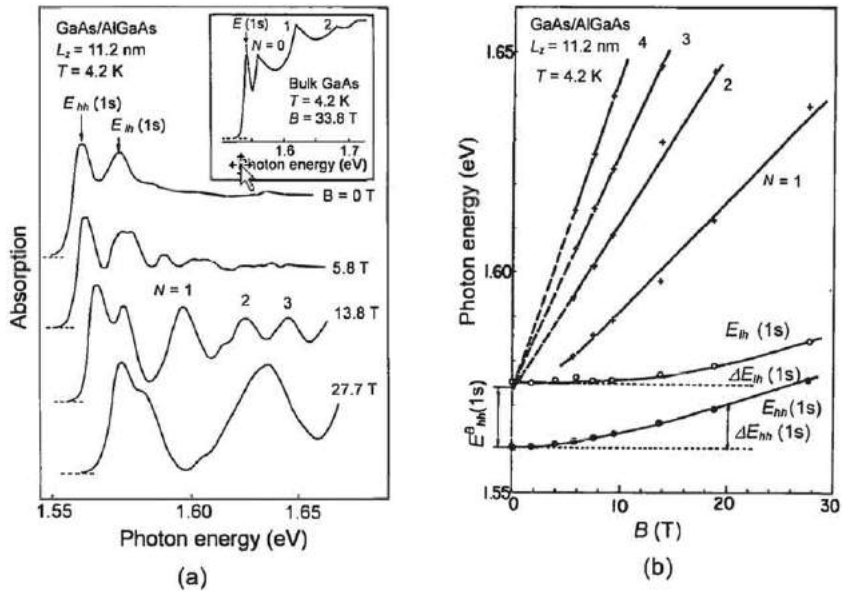

Fig. 2 (color online). (a) Optical absorption spectra at different magnetic fields for a GaAs/AlAs multiple-quantum-well sample. (b) Magnetic field dependence of the photon energies of absorption peaks. Reproduced (adapted) with permission from Tarucha, S., Okamoto, H., Iwasa, Y., Miura, N., 1984. Exciton binding energy in GaAs quantum wells deduced from magneto-optical absorption measurement. Solid State Communications 52, 815–819. Copyright 1984 by Elsevier.

region, the 1*s* exciton energy shows a quadratic field dependence, in agreement with the expected diamagnetic shift in energy for excitons. In addition, the change of the heavy-hole transition energy at a certain field $\Delta E_{\mathrm{hh}}(1s)$ is observed to be larger than the change of the light-hole transition energy $\Delta E_{\mathrm{lh}}(1s)$ at the same field, indicating a smaller reduced mass for the 1*s* heavy-hole excitons than that of the 1*s* light-hole excitons, since the diamagnetic shift is proportional to $1/\mu^*$; see Eq. (31). The difference in the reduced masses of excitons can be explained by the anisotropic nature of the kinetic energy expression in the diagonal term of the Luttinger-Kohn Hamiltonian. The diagonal term gives the larger reduced mass for the light-hole excitons associated with $J_z(\pm 3/2)$ bands than that for the heavy-hole excitons associated with $J_z(\pm 1/2)$ bands.

In contrast, the absorption peaks for higher-energy heavy-hole states show a linear B dependence. The effect of the Coulomb interaction can only be seen on the 2*s* state, since its energy shift shows a small curvature in the low-B region. By extrapolating the 3*s*, 4*s*, and 5*s* absorption peaks from the high-B region to the low-B region with straight lines, it was found that the lines converge at a photon energy within an error of less than 0.5 meV. The difference between the energy at the convergent point and the zero field value of $E_{\mathrm{hh}}(1s)$ gives the binding energy of the 1*s* heavy-hole exciton, which was $\sim$15 meV in this case. The exciton binding energy of the 1*s* heavy-hole exciton increases with decreasing well size, since the quasi-2D nature is a closer approximation for smaller quantum well sizes. On the contrary, the diamagnetic shift increases with increasing well size, indicating a larger Bohr radius in a higher-dimensionality system.

Transition Metal Dichalcogenide

Atomically thin transition-metal dichalcogenides (TMDs), a new class of 2D materials, have recently attracted much attention for magneto-optical studies due to their extremely large exciton binding energies. For example, Stier *et al.* performed low-temperature polarized reflection measurements on atomically thin WS_2 and MoS_2 in high magnetic fields up to 65 T, observing a strong Zeeman splitting and a small diamagnetic shift for the 2D excitons. Based on their measurements, some basic parameters of excitons in these TMDs, such as the g-factor, exciton size, and binding energy, were determined.

As shown in **Fig. 3(a)**, 2D TMDs have a direct band gap located at the degenerate K and K' valleys of their hexagonal Brillouin zones. The strong spin-orbit coupling of the valence band lifts the degeneracy of the spin-up and spin-down components, leading to well-separated A and B excitons. Due to the valley-specific optical selection rules, the interband transitions in the K valley only couple to right circularly polarized light, σ^+, for both the A and B excitons, but left circularly polarized light, σ^-, for the K' valley.

Fig. 3(b) plots the B-dependent energy shifts of the conduction and valence bands in the K valley (σ^+ polarized light) with $B \| \pm z$, and demonstrates different contributions to the Zeeman splitting. For a specific valley in a B, the Zeeman shift is $\Delta E_Z = -(\mu_c - \mu_v) \cdot B$. In addition, the time-reversal symmetry in TMDs leads to an equal-but-opposite total magnetic moment $\left(\mu_K^{c,v} = -\mu_{K'}^{c,v}\right)$ in the K and K' valleys. However, this time-reversal symmetry will be broken by B, and the time-reversed pairs in K and K' valleys will shift to opposite directions, i.e., the Zeeman splitting for excitons is doubled considering the broken time-reversal symmetry. Generally, the total magnetic moment μ comes from three sources: spin (μ_s), atomic orbital (μ_l), and valley

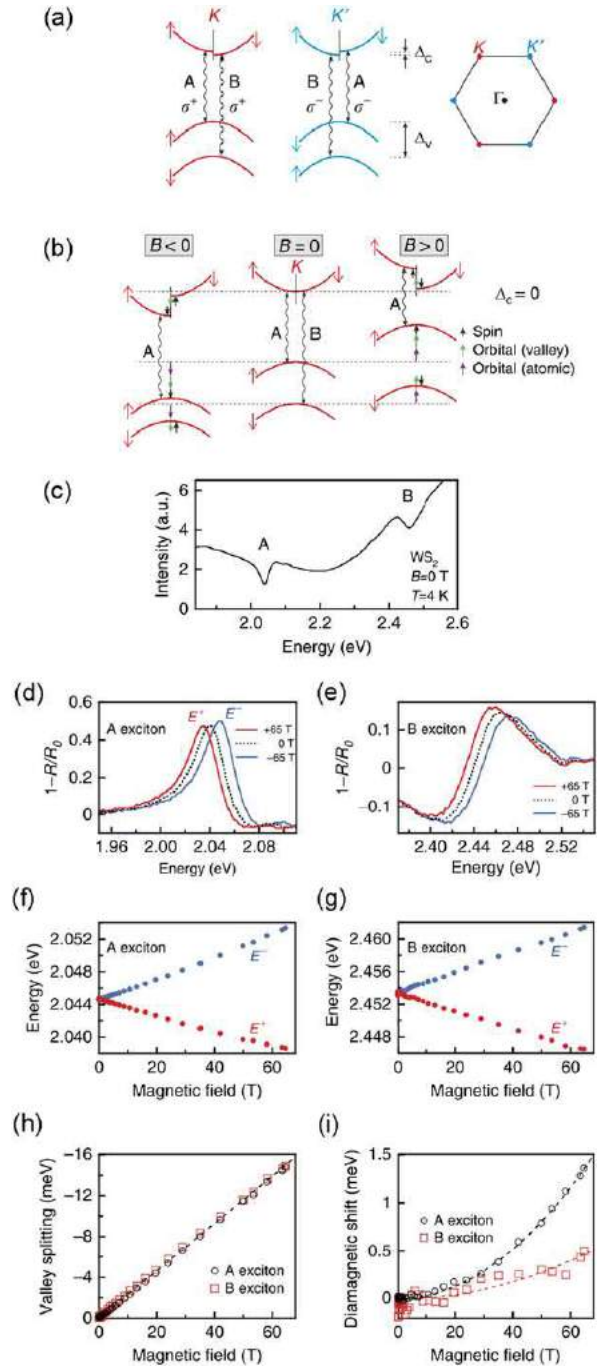

Fig. 3 (color online). (a) Excitonic transitions in monolayer transition metal dichalcogenides. (b) Zeeman shifts in different B fields. (c) Reflection spectra of monolayer WS_2 at $B=0$ T and $T=4$ K. (d) Normalized reflection spectra of the A exciton resonance in different B. (e) Same, but for the B exciton. (f) Energy of the field-split A exciton versus magnetic field. (g) Same, but for the B exciton. (h) The measured valley Zeeman splitting versus B, for both the A and B excitons. (i) The diamagnetic shifts for both the A and B excitons. Reproduced (adapted) with permission from Stier, A.V., McCreary, K.M., Jonker, B.T., Kono, J., Crooker, S., 2016. Exciton diamagnetic shifts and valley Zeeman effects in monolayer WS_2 and MoS_2 to 65 Tesla. Nature Communications 7, 10643. Copyright 2016 by Nature Publishing Group.

orbital (μ_k) (Berry curvature). Here, μ_s is zero, since the optically allowed transitions couple the conduction and valence bands with the same spin. μ_I is non-zero, since the conduction and valence bands have different atomic orbitals. The d_{z^2} orbitals of the conduction band have $L_z=0$ $\left(\mu_I^c=0\right)$, while the hybridized $d_{x^2-y^2}\pm id_{xy}$ orbitals for the valence bands have $L_z=\pm 2\hbar$ $\left(\mu_I^v=\pm 2\mu_B\right)$ in the K and K' valleys, generating a Zeeman shift of $\mp 2\mu_B B$ for the K and K' excitons, and leading to a total exciton splitting of $-4\mu_B B$. For the valley orbital contribution to the conduction and valence bands, $\mu_k^c=\pm\left(m_e/m_e^*\right)\mu_B$ and $\mu_k^v=\pm\left(m_e/m_h^*\right)\mu_B$ in the K and K' valleys. Assuming a simple two-band tight-binding model where $m_e^*=m_h^*$, it is easy to see that $\mu_k^c=\mu_k^v$, i.e., the

contribution of the valley orbital to the Zeeman splitting is zero. In total, the Zeeman splitting for excitons is expected to be $\Delta E_Z = -4\mu_B B$ assuming $m_e^* = m_h^*$.

Fig. 3(c) shows the reflection spectrum of monolayer WS_2 at 4 K and 0 T, from which the A and B excitonic transitions can be seen. The normalized reflection spectra at 0 T and ± 65 T of the A exciton are shown in Fig. 3(d), indicating a clear Zeeman splitting of ~ 15 meV. The valley splitting of the B exciton can also be observed, as shown in Fig. 3(e). For both excitons, the exciton transition energy in a positive B (called E^+) shifts to a lower energy, while a shift to a higher energy is observed for the exciton transition energy in a negative B (called E^-). Fig. 3(f) and (g) show the transition energies, E^+ and E^-, as a function of B, for the A and B excitons, respectively. The splittings between the two valleys, $E^+ - E^-$, are shown in Fig. 3(h) for both the A and B excitons. The measured Zeeman splitting energies are negative but increase linearly with B with almost identical rates of -228 ± 2 μeVT^{-1} for the A exciton and -231 ± 2 μeVT^{-1} for the B exciton. Such rates correspond to g-factors of -3.94 ± 0.04 and -3.99 ± 0.04, respectively, very close to the expected value (-4).

Fig. 3(i) shows the diamagnetic shifts for the A and B excitons in monolayer WS_2. As we saw in Eq. (33), $a^*_{B,2D}$ can be calculated if μ^* and σ^*_{2D} are known.

Through fits to the data, $\sigma^*_A = 0.32 \pm 0.02$ μeVT^{-2} and $\sigma^*_B = 0.11 \pm 0.02$ μeVT^{-2} were obtained for the A and B excitons, respectively. Using the theoretically estimated reduced mass of the A exciton, ranging from 0.15 to $0.22 m_e$, $a^*_{B,A} = 1.48 - 1.79$ nm is obtained.

Furthermore, the exciton binding energy and wavefunction can be determined by numerically solving the 2D Schrödinger equation with the values σ and a_B determined in these measurements. For 2D TMDs, a modified Coulomb potential $U(\rho)$ was used in the calculation,

$$U(\rho) = -\frac{e^2}{8\varepsilon_0 \rho_0}\left[H_0\left(\frac{\rho}{\rho_0}\right) - Y_0\left(\frac{\rho}{\rho_0}\right)\right] \tag{37}$$

where H_0 and Y_0 are the Struve function and Bessel function of the second kind, respectively, and the characteristic screening length $\rho_0 = 2\pi\chi_{2D}$, where χ_{2D} is the 2D polarizability of the monolayer material. Such a potential follows a $1/\rho$ Coulomb-like potential for large electron-hole separations $\rho \gg \rho_0$, but diverges weakly as $\log(\rho)$ for small separations $\rho \ll \rho_0$, leading to a different Rydberg series of exciton states with modified wavefunctions and binding energies that cannot be described within a hydrogen-like model. By doing this, some characteristic parameters are obtained for the A and B excitons in monolayer WS_2. For example, with $\mu^*_A = 0.16 m_0$ and $a^*_{B,A} = 1.53$ nm, the binding energy of the A exciton is calculated as $E^*_{b,A} = 410$ meV. For the B exciton with $\mu^*_B = 0.27 m_0$ and $a^*_{B,B} = 1.16$ nm, $E^*_{b,B} = 470$ meV.

Perovskites

Solar cells based on organic-inorganic tri-halide perovskites show significantly improved performance, and therefore, it is very important to know their fundamental optical properties, such as the exciton binding energy and the reduced mass, in order to better understand the functions of solar cells. In addition, these materials have structural phase transitions from cubic ($T > 350$ K) to tetragonal ($T > 145$ K) to orthorhombic (low T), which leads to different band structures and band gaps in the different phases. By performing transmission measurements on a ~ 300-nm-thick polycrystalline film of $CH_3NH_3PbI_3$ in a high B up to 150 T and at different temperatures, Miyata *et al.* obtained accurate values for the exciton binding energy and reduced mass.

As shown in Fig. 4, at 2 K, the transition energies are observed to depend quadratically on B for the lowest two states (labeled "$N=0$"), as well as the "$N=1$" transition at $B < 50$ T, indicating an excitonic feature for these low-energy states. On the other hand, the transition energies depend linearly on B for the $N=1$ transition at $B > 50$ T, and the higher energy states. The authors used full numerical calculations to fit the B-dependent transition energy data and obtained an accurate value for the reduced mass $\mu^* = 0.104 \pm 0.003 m_e$. They also obtained the exciton binding energy $E^*_b = 16 \pm 2$ meV, which is over three times smaller than previously assumed. At room temperature, in the tetragonal structural phase, the exciton binding energy decreases to only a few meV, due to a frequency dependent dielectric constant, which reveals the free-carrier like nature of the photovoltaic device performance at room temperature.

Quantum Wires and Dots

Similar to 2D excitons in quantum wells, 1D excitons are observed in quantum wires, showing combined effects of magnetic fields and quantum potential in high magnetic fields. Nagamune *et al.* measured photoluminescence (PL) spectra of a GaAs quantum wire in fields up to 40 T, with three orthogonal B configurations. Fig. 5(a) displays the PL peak position as a function of B in different configurations, for the GaAs quantum wire and a bulk GaAs sample. The bulk PL peak shifts are almost independent of the magnetic field configuration, in contrast with a clear magnetic anisotropy observed in the quantum wire sample.

In the low-B limit, the PL peaks for the quantum wire show a quadratic increase with B, from which the diamagnetic-shift coefficients were determined as 16.3 μeVT^{-2}, 7.0 μeVT^{-2}, and 4.5 μeVT^{-2} for the $\vec{W} \perp \vec{B} \parallel \vec{k}$, $\vec{W} \perp \vec{B} \perp \vec{k}$ and $\vec{W} \parallel \vec{B} \perp \vec{k}$ configurations, respectively, where $\vec{W}$ is the direction along the quantum wire, and $\vec{k}$ the wavevector of the excitation beam. These values are smaller than 103.9 μeVT^{-2} for the bulk sample, indicating increased confinement in a reduced dimensionality.

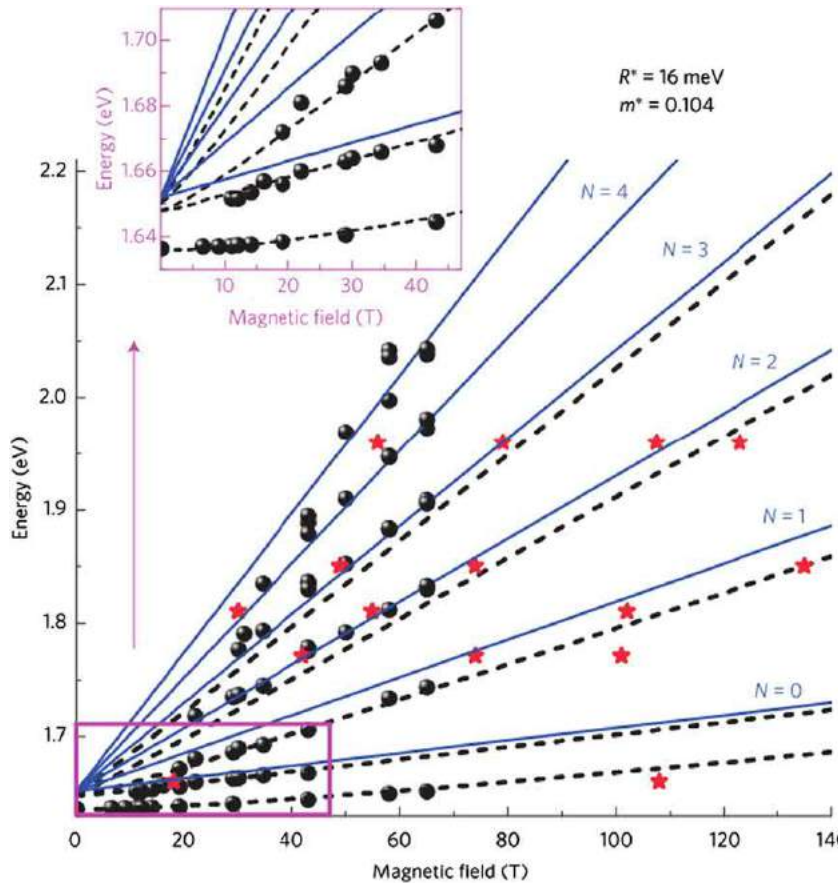

Fig. 4 (color online). Magnetic-field-dependent transition peaks of the perovskite $CH_3NH_3PbI_3$ at 2 K. The calculated transition energies are shown for the free-electron and hole levels (solid lines) and the excitonic transitions (dashed lines). Inset: measured and calculated transition energies at low fields. Reproduced (adapted) with permission from Miyata, A., Mitioglu, A., Plochocka, P., *et al.*, 2015. Direct measurement of the exciton binding energy and effective masses for charge carriers in organic-inorganic tri-halide perovskites. Nature Physics 11, 582–588. Copyright 2015 by Nature Publishing Group.

In the high-B limit, the PL peak shifts increase roughly linearly with increasing B. In particular, the PL energy shift under $\overrightarrow{W} \perp \overrightarrow{B} \parallel \overrightarrow{k}$ is almost parallel to that of a bulk sample, which is ascribed to the exciton radius becoming much smaller than the width of the confinement potential in a high B.

In quantum dots, electrons and holes are confined in all three directions, and quasi-zero-dimensional (0D) excitons are observed. Hayden *et al.* performed magneto-PL measurements on self-assembled InGaAs quantum dots in B up to 42 T. The PL peak shifts as a function of B are shown in **Fig. 5(b)** for different samples and B directions. Similar to the case of quantum wires, the low-B data is characterized by a diamagnetic shift, and the high-B data indicates a Landau-level like behavior. In addition, the diamagnetic shift is always smaller when B is applied perpendicular to the growth direction rather than parallel to it, revealing that the 0D excitons are better confined in the plane perpendicular to B.

Carbon Nanotubes

Single-wall carbon nanotubes (SWCNTs) are another representative 1D semiconductor, which are tubular crystals composed of sp^2-bonded carbon atoms. Unlike quantum wires, the basic properties of SWCNTs are determined by a pair of integers, or chirality, (n,m). For tubes with $n=m$, they are metals, called "armchair" tubes; for tubes with $n-m=3j$ (j is nonzero integer), they are small-gap semiconductors; and all others are medium-gap semiconductors. In an applied B where the magnetic flux ϕ passes through the axial direction of the SWCNTs (a tube threading flux), the quantum states of electrons in SWCNTs acquire an Aharonov-Bohm phase $2\pi\phi_0/\phi_0$ ($\phi_0=h/e$ is the magnetic flux quantum), causing the band structure of SWCNTs to change with ϕ/ϕ_0. For $0<\phi/\phi_0<1/6$, the Aharonov-Bohm phase can cause a band gap change of $6E_g\phi/\phi_0$ (E_g is the band gap) in a semiconducting SWCNT, which can be observed in interband absorption and PL measurements.

In a semiconducting SWCNT with time-reversal symmetry, at $B=0$ T, as shown in **Fig. 6(a)**, the bonding-like and anti-bonding-like linear combinations of two equivalent valleys, K and K', result in two lowest energy exciton states: optically active (or "bright") and inactive (or "dark") states. The energies of the two states are split by Δ_X, which is determined by the Coulomb interaction, the

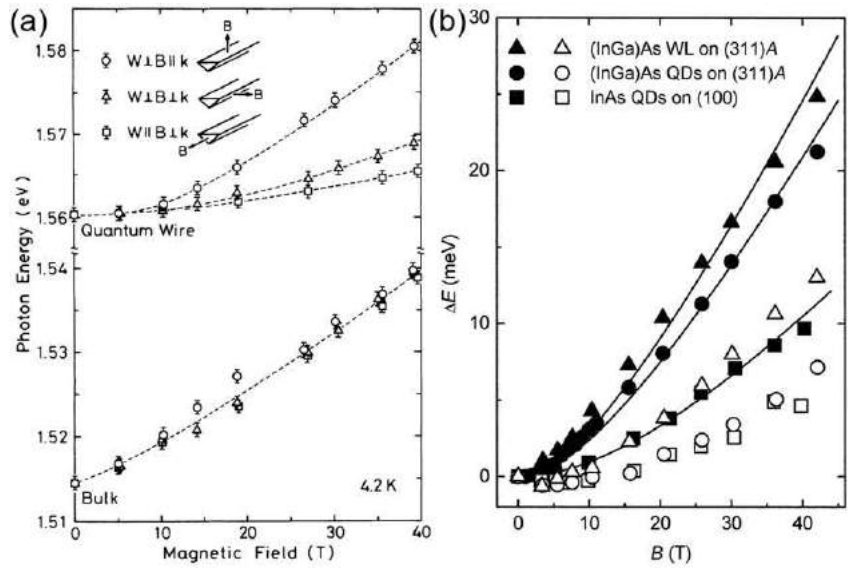

Fig. 5 (color online). (a) PL peak positions for a GaAs quantum wire and bulk GaAs samples as a function of B under various B configurations. Reproduced (adapted) with permission from Nagamune, Y., Arakawa, Y., Tsukamoto, S., Nishioka, M., 1992. Photoluminescence spectra and anisotropic energy shift of GaAs quantum wires in high magnetic fields. Physical Review Letters 69, 2963–2966. Copyright 1992 by the American Physical Society. (b) PL peak shifts as a function of B applied parallel to the growth axis (filled symbols) and perpendicular to the growth axis (open symbols), for quantum dots and wetting layers on different substrates. Reproduced (adapted) with permission from Hayden, R.K., Uchida, K., Miura, N., *et al.*, 1998. High field magnetoluminescence spectroscopy of self-assembled (InGa)As quantum dots on high index planes. Physica B: Condensed Matter 246, 93–96. Copyright 1998 by Elsevier.

tube diameter, and the dielectric constant of the environment. However, an applied B will break the time-reversal symmetry, and lift the $K-K'$ degeneracy, so that the dark state becomes optically active when the Aharonov-Bohm-induced splitting is larger than the Coulomb-induced splitting ($\Delta_{AB}>\Delta_X$). As a result, magnetic brightening of the dark exciton state can be observed in interband absorption or PL.

Fig. 6(b) shows absorption spectra of semiconducting SWCNTs at different B up to 71 T in a Voigt geometry, i.e., the light propagation vector is perpendicular to $\vec{B}$. The absorption increases with B, since the suspended carbon nanotubes progressively aligned with B more strongly due to the anisotropy of the magnetic susceptibility. The three main absorption peaks at $B=0$ T are labeled as A, B, and C, and each absorption peak splits into two well-resolved peaks at $B>55$ T, due to the Aharonov-Bohm-induced splitting. This phenomenon can only be observed when the light polarization is parallel with B, as shown in Fig. 6(c), as a result of magnetic anisotropy in SWCNTs.

Fig. 6(d) shows a PL contour map of a partially aligned carbon nanotube film in pulsed fields up to 56 T at 5 K. The PL intensity increases with B due to magnetic brightening. Furthermore, the Aharonov-Bohm effect causes a red-shift of each PL emission peak, since only the formerly dark state is populated due to a splitting much larger than the thermal energy. At low temperatures, the broken time-reversal symmetry increases the PL quantum yield by as much as a factor of 6, as shown in Fig. 6(e). The magnetic brightening becomes weaker with increasing temperature; for instance, at 200 K the integrated PL intensity only increases by $\sim$2 times.

Magnetic brightening has also been observed in single-nanotube PL. In this case, inhomogeneous broadening due to environmental effects are negligible, especially at low temperatures, and the dark-bright splitting, Δ_X, can be larger than the linewidth. Then, the dark state can become bright with an applied magnetic field. Fig. 6(f) shows B-dependent single-tube PL spectra at 5 K with B applied along the tube axis. At $B=0$ T, only a single, sharp PL peak can be observed, due to the bright exciton state. With increasing B, a second PL peak appears on the low energy side, grows rapidly in intensity and finally dominates the emission spectra at $B>3$ T. The splitting between the two PL peaks increases with B. These behaviors were universally observed for more than 50 tubes of different chiralities, and the dark-bright splitting was 1–4 meV for tube diameters 1.0–1.3 nm.

Intraexciton Magneto-optical Absorption

Photoluminescence and absorption measurements can reveal information on interband transitions of s-like states of excitons. However, in order to explore the internal structure of excitons, intraexciton transitions such as $1s\rightarrow np$ ($n=2,3,\ldots$) transitions have to be studied using far-infrared (FIR) or terahertz (THz) radiation, whose photon energy is comparable to or smaller than the binding energy of excitons.

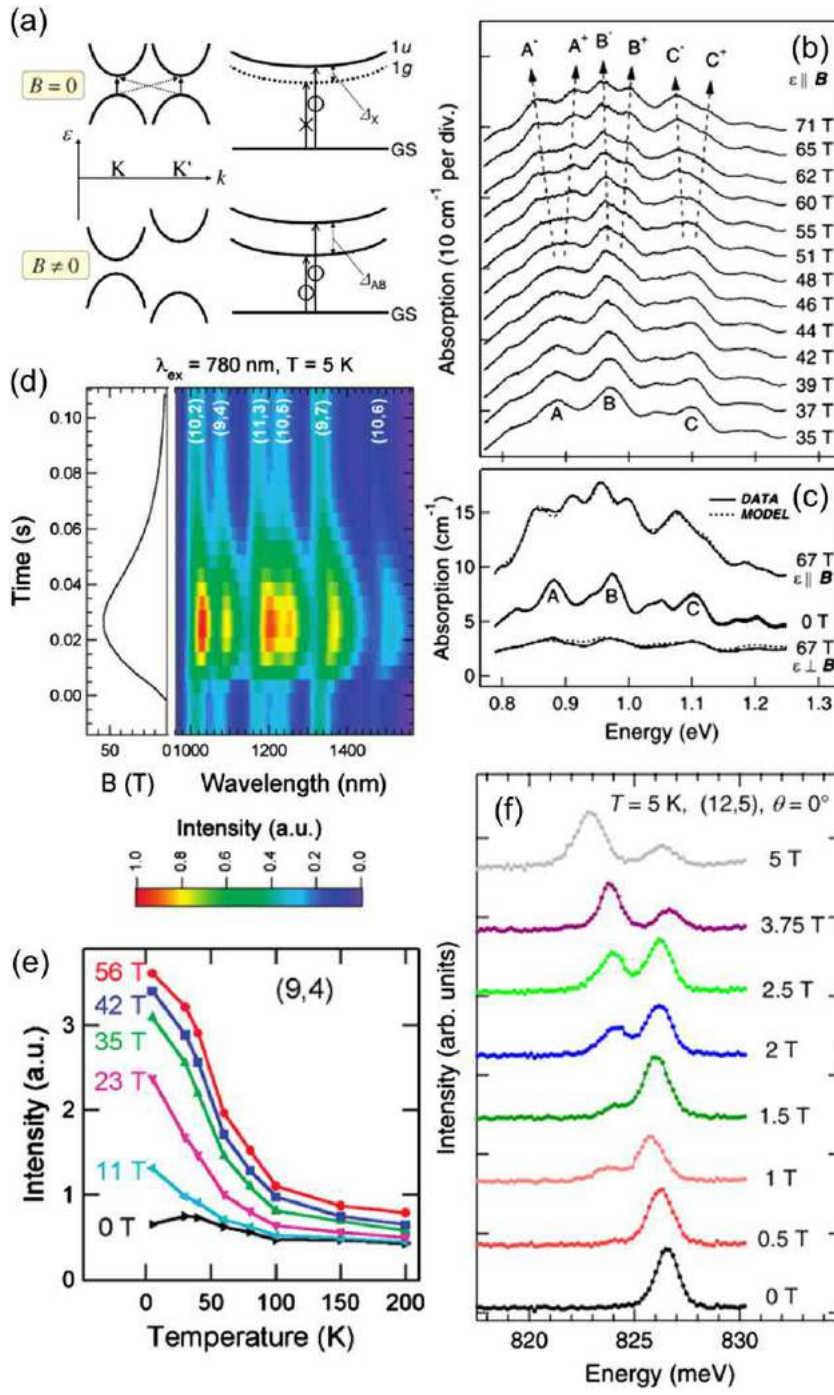

Fig. 6 (color online). (a) The expected B evolution of $K-K'$ intervalley mixing and splitting in a single-particle picture (left) and an excitonic picture (right). The solid (dashed) line represents a bright (dark) exciton state. Δ_X: Coulomb-induced splitting; Δ_{AB}: Aharonov-Bohm-induced splitting. Near band-edge absorption in semiconducting SWCNTs in high B for polarization (b) parallel to B (traces are offset) and (c) both polarizations (no intentional offset). Reproduced (adapted) with permission from Zaric, S., Ostojic, G. N., Shaver, J., *et al.,* 2006. Excitons in carbon nanotubes with broken time-reversal symmetry. Physical Review Letters 96, 016406. Copyright 2006 by the American Physical Society. (d) Magnetic brightening in SWCNTs in pulsed B. (e) Temperature dependence of magnetic brightening for (9,4) SWCNTs. Reproduced (adapted) with permission from Shaver, J., Kono, J., Portugall, O., *et al.,* 2007. Magnetic brightening of carbon nanotube photoluminescence through symmetry breaking. *Nano Letters* 7, 1851–1855. Copyright 2007 by the American Chemical Society. (f) B-dependent PL spectra for a single tube, showing the appearance of a dark exciton peak at a lower energy with respect to the main bright emission peak when B is applied parallel to the tube axis. Reproduced (adapted) with permission from Srivastava, A., Htoon, H., Klimov, V.I., Kono, J., 2008. Direct observation of dark excitons in individual carbon nanotubes: inhomogeneity in the exchange splitting. Physical Review Letters 101, 087402. Copyright 2008 by the American Physical Society.

FIR Magnetoabsorption in Photoexcited Bulk Semiconductors

Observation of intraexciton magneto-absorption in bulk semiconductors was initiated by Gershenzon *et al.* on the indirect bandgap semiconductor germanium in the microwave frequency range. Subsequently, Muro *et al.* and Ohyama observed FIR magnetoabsorption of indirect excitons in germanium and silicon, respectively.

Fig. 7(a) shows FIR magnetoabsorption spectra for undoped germanium induced by interband photoexcitation at 4.2 K for various wavelengths. The observed spectra consist of exciton lines (E_X) and cyclotron resonance (CR) lines (C_e and C_h). Since the valence band is four-fold degenerate at the Γ point in germanium, the indirect free exciton states are described by four coupled effective mass equations, leading to the observation of four types of excitonic peaks in the spectra, E'_{X1}, E'_{X2}, E'_{X3}, and E'_{X4}. Up to the highest photoexcitation intensity used, these exciton peaks were stable against the formation of *e-h* droplets. The splitting of C'_e comes from the slight tilt of B from the [100]-crystal axis. C'_{h1}, C'_{h2}, C'_{h3}, and C'_{h4} correspond to different cyclotron transitions related to the light-hole and heavy-hole bands.

Fig. 7(b) shows the transition energies of intraexciton lines and CR lines as a function of B. An indirect exciton in germanium is composed of an electron at the L-point and a hole at the Γ-point bound by the Coulomb interaction, giving a hydrogenic structure. At zero B, the exciton transition lines converge to a point, indicating a finite transition energy of indirect excitons, around 3 meV. At high B, where the cyclotron energy is much larger than the Coulomb interaction, the intraexciton energies are expected to increase linearly with B, although it is not observed in this experiment since the highest B (7 T) was still not large enough to satisfy the adiabatic condition (required for the decoupling of the x-y and z degrees of freedom). These results demonstrate that FIR magnetoabsorption measurements can not only reveal the CR of electrons and holes but also provide detailed information on the intraexciton transitions in indirect semiconductors with complicated band structures.

Optically Detected FIR Resonance Spectroscopy of Semiconductor Quantum Wells

In optically detected resonance (ODR) measurements, two lasers illuminate the sample at the same time. One laser with near-infrared (NIR) or visible radiation is used to create excitons, while the other with FIR (or THz) radiation manipulates them, with the frequency tuned to the resonance at a given B. By monitoring the band-edge PL, information about intraexcitonic transition energy and strength can be obtained. Fig. 8(a) shows the ratio of PL amplitudes with and without FIR irradiation as a function of B for a GaAs/AlGaAs quantum well sample, measured by Černe *et al.* Two types of transitions are observed: electron CR and the $1s \rightarrow 2p^+$ excitonic transition. At a low FIR frequency, such as 20 cm^{-1}, only electron CR can be observed at a B. With increasing FIR frequency, the electron CR shifts to a higher B, and at the same time the $1s \rightarrow 2p^+$ excitonic transition appears at a relatively low B, along with a possible $1s \rightarrow 3p^+$ excitonic transition. At a high FIR frequency, such as 130 cm^{-1}, the electron CR energy does not become equal to the FIR photon energy within the measured B range, and only excitonic transitions can be observed.

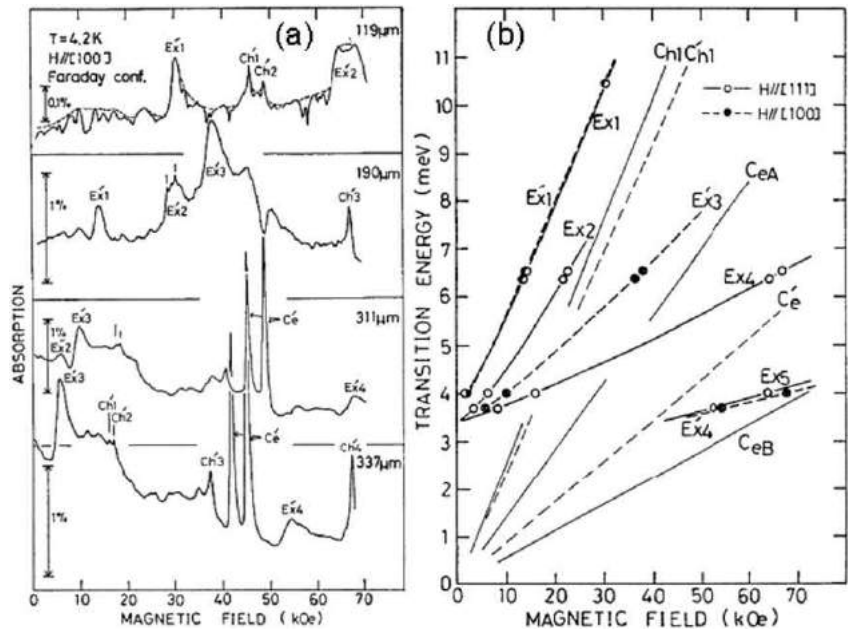

Fig. 7 (color online). (a) Magneto-absorption spectra for undoped germanium at 4.2 K with *H*∥[100] under optical excitation at varied wavelengths. (b) Transition energies of main exciton Zeeman lines and cyclotron resonance lines as a function of magnetic field. Reproduced (adapted) with permission from Muro, K., Nisida, Y., 1976. Far-infrared magneto-absorptions in photo-excited germanium. Journal of the Physical Society of Japan 40, 1069–1077. Copyright 1976 by the Physical Society of Japan.

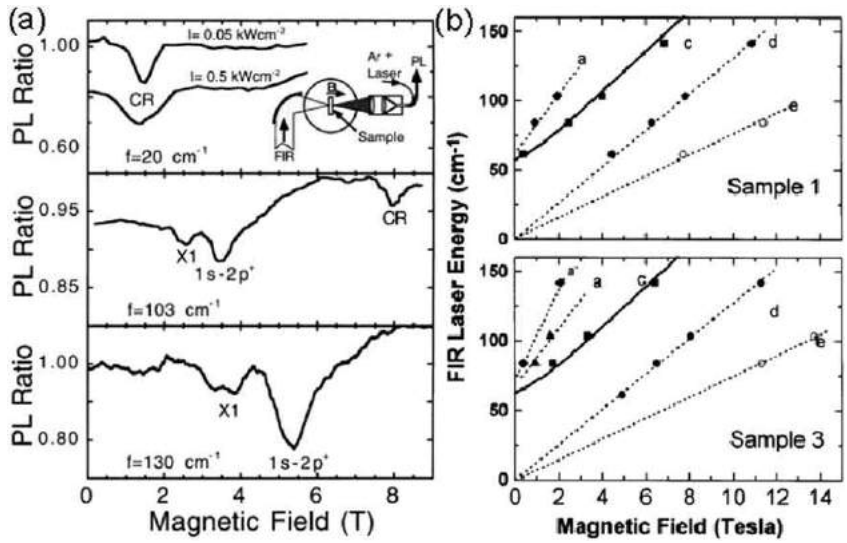

Fig. 8 (color online). Optically detected resonance spectroscopy. (a) The ratio of the PL amplitude with and without FIR irradiation as a function of B at three FIR frequencies. The inset shows the experimental setup. Reproduced (adapted) with permission from Černe, J., Kono, J., Sherwin, M. S., *et al.*, 1996. Terahertz dynamics of excitons in GaAs/AlGaAs quantum wells. Physical Review Letters 77, 1131–1134. Copyright 1996 by the American Physical Society. (b) Energies of the optically detected resonances as a function of B at 4.2 K. Sample 1: an undoped (12.5 nm well/15nm barriers)$_{30}$ structure. Sample 3: a (8 nm well/15 nm barrier)$_{45}$ structure, doped with Si donors ($n=1\times10^{16}$ cm^{-3}) throughout the central one-third of the wells. Reproduced (adapted) with permission from Salib, M.S., Nickel, H.A., Herold, G.S., Petrou, A., McComber, B.D., 1996. Observation of internal transitions of confined excitons in GaAs/AlGaAs quantum wells. Physical Review Letters 77, 1135–1138. Copyright 1996 by the American Physical Society.

Salib *et al.* also performed similar ODR measurements on GaAs/AlGaAs quantum wells to observe intraexciton transitions. Fig. 8(b) plots the energies of the ODR features for different quantum wells as a function of B. In both samples, electron and hole CR lines (Features d and e) are observed, with their frequencies linearly dependent on B. In addition, intraexcitonic transitions, $1s\rightarrow2p^+$ (Feature c) and $1s\rightarrow3p^+$ (Feature a), can also be measured. In Sample 3, even the $1s\rightarrow4p^+$ transition (Feature a′) can be well resolved. These ODR measurements provide a method to explore the internal $1s\rightarrow np$ ($n=2, 3, \ldots$) transitions of photoexcited excitons, directly determining the energies of p-like exciton states.

In the nonlinear optics version of ODR, the sample is irradiated simultaneously by a NIR beam and an *intense* THz beam in the presence of B. Then very strong NIR emission lines, or sidebands, appear at frequencies $\omega_{NIR}\pm2n\omega_{THz}$, where ω_{NIR} (ω_{THz}) is the frequency of the NIR (THz) beam and n is an integer. This process is highly resonant, and thus, changes sensitively with B, providing a novel method for intraexciton spectroscopy.

Fig. 9(a) and (b) show typical sideband spectra at 10 T, with the THz frequency $\omega_{THz}=115$ cm^{-1}. Fig. 9(a) shows the down-conversion effect: a narrow emission line, resulting from the creation of 2s heavy-hole excitons, is observed at 12745 cm^{-1}, and the -2ω sideband appears at exactly 12515 cm^{-1} [$=12745-(2\times115)$ cm^{-1}] only under the THz irradiation. Fig. 9(b) shows the up-conversion behavior, with $+2\omega$ and $+4\omega$ sidebands in the higher energy region, through the resonant creation of 1s excitons at 12531 cm^{-1}. The typical intensities of the -2ω, $+2\omega$ and $+4\omega$ sidebands are 0.05%, 0.15% and 0.0015%, respectively, of the incident intensity of the NIR excitation. The intensity of the $+2\omega$ sideband increases quadratically with the THz power under a constant NIR intensity, and increases linearly with the NIR power for a constant THz intensity. Therefore, the 2ω THz sideband generation process is a third-order nonlinear optical process, involving one NIR photon and two THz photons.

Fig. 9(c) shows the $+2\omega$ sideband intensity as a function of B with $\omega_{THz}=115$ cm^{-1} and $\omega_{NIR}=\omega_{1s}$, where ω_{1s} is the frequency for 1s heavy-hole excitons. Three distinct resonances can be observed, located at 4.5 T, 9.5 T and 11.5 T. These resonances occur when $\omega_{THz}=\omega_{2p^+}-\omega_{1s}$, $2\omega_{THz}=\omega_{2s}-\omega_{1s}$, and $\omega_{THz}=\omega_{2p^-}-\omega_{1s}$, as shown in Fig. 9(d). According to calculations using perturbation theory, the third-order nonlinear optical susceptibility $\chi^{(3)}$ is resonantly enhanced, when (A) $\omega_{NIR}=\omega_{ns}$, (B) $\omega_{NIR}+\omega_{THz}=\omega_{n'pm}(m=\pm1)$, and (C) $\omega_{NIR}+2\omega_{THz}=\omega_{n''s}$. In all three cases in Fig. 9(d), two of the three conditions (A)-(C) are satisfied: (left) $\omega_{NIR}=\omega_{1s}$, $\omega_{NIR}+\omega_{THz}=\omega_{2p^+}$, (middle) $\omega_{NIR}=\omega_{1s}$, $\omega_{NIR}+2\omega_{THz}=\omega_{2s}$, and (right) $\omega_{NIR}=\omega_{1s}$, $\omega_{NIR}+\omega_{THz}=\omega_{2p^-}$. In other words, when ω_{THz} coincides with magnetically tuned transitions of excitons, the intensity of sidebands is dramatically enhanced, which provides a highly sensitive way to explore the internal structure of excitons.

Optical-Pump/THz-Probe Magnetospectroscopy of Semiconductor Quantum Wells

In order to obtain information on the dynamics of intraexcitonic transitions, time-resolved THz spectroscopy can be a powerful technique. Zhang *et al.* performed time-resolved THz absorption measurements on photoexcited *e-h* pairs in undoped GaAs quantum wells at various magnetic fields, temperatures and pump intensities, investigating both CR and intraexciton resonance features through resonant and nonresonant excitations of the heavy-hole 1s excitons.

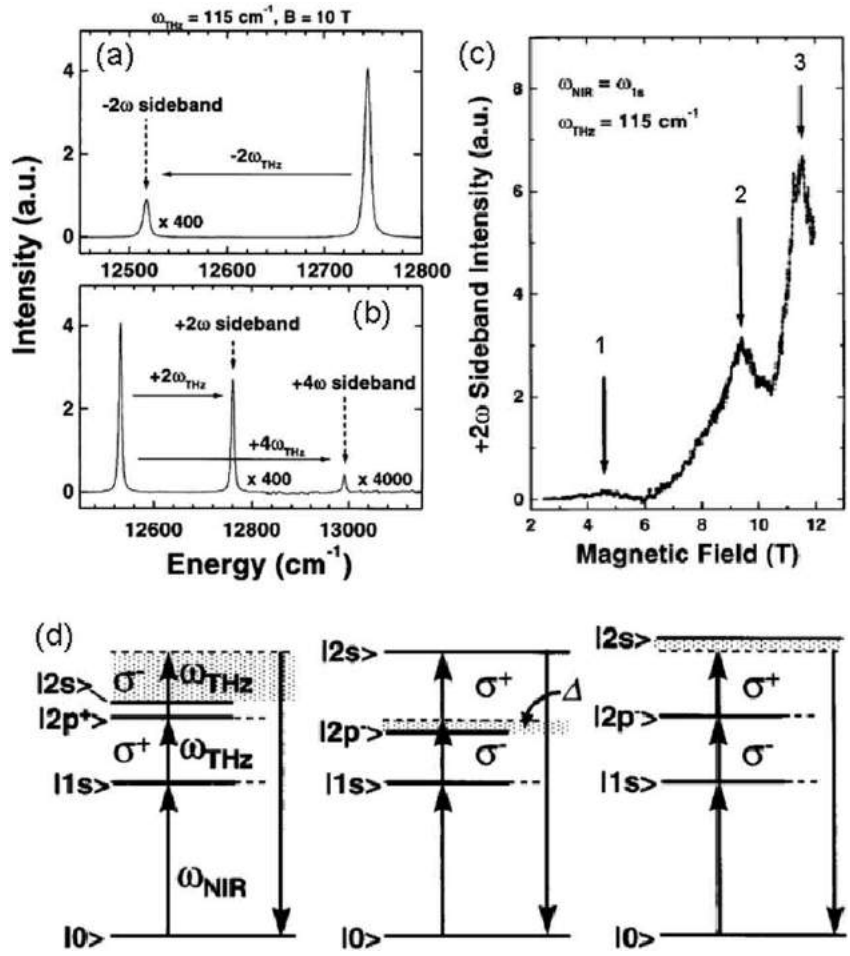

Fig. 9 (color online). Nonlinear THz ODR with ω_{THz}=115 cm^{-1}. (a) Down conversion: ω_{NIR}=ω_{2s}=12745 cm^{-1} and $\omega_{sideband}$=$\omega_{NIR}-2\omega_{THz}$=12515 cm^{-1}. (b) Up conversion: ω_{NIR}=ω_{1s}=12515 cm^{-1}, $\omega_{sideband}$=$\omega_{NIR}+2\omega_{THz}$=12745 cm^{-1}, and $\omega_{NIR}+4\omega_{THz}$=12975 cm^{-1}. (c) The B dependence of the $+2\omega$ sideband intensity for ω_{NIR}=ω_{1s} at all B. The data demonstrate nonlinear THz ODR. Pronounced resonances occur when 1: $\omega_{THz}=\omega_{2p^+}-\omega_{1s}$, 2: $2\omega_{THz}$=$\omega_{2s}-\omega_{1s}$, and 3: $\omega_{THz}=\omega_{2p^-}-\omega_{1s}$. (d) Diagrammatic representation of the three resonances in (c). The dashed lines represent virtual levels, whereas the solid lines represent real magnetoexcitonic levels. Resonances occur when real levels coincide with virtual levels. The shaded area represents the magnitude of the detuning, Δ. Reproduced (adapted) with permission from Kono, J., Su, M.Y., Inoshita, T., *et al.*, 1997. Resonant terahertz optical sideband generation from confined magnetoexcitons. Physical Review Letters 79, 1758–1761. Copyright 1997 by the American Physical Society.

Fig. 10(a) and **(b)** show the photoinduced change in conductivity, Re$\Delta\sigma(\omega)$, as a function of B and energy, at different time delays after nonresonant excitation at 5 K. At 15 ps, two distinct CR features due to unbound electrons and holes are observed, from which the effective masses can be determined: $m_e^*=0.070m_e$ and $m_h^*=0.54m_e$. At 1 ns, the intraexcitonic transitions, $1s\rightarrow 2p^+$ and $1s\rightarrow 2p^-$, are also observed, indicating the formation of excitons.

Fig. 10(c) and **(d)** present the time evolution of the change in conductivity at 9 T and 6 K with a pump fluence of 200 nJcm^{-2} for nonresonant and resonant excitation, respectively. Under nonresonant excitation, hole CR is observed at early time delays, such as 5 ps and 100 ps, but then decays with time and eventually the $1s\rightarrow 2p^-$ intraexciton transition appears after 100 ps. However, in the case of resonant excitation, only the $1s\rightarrow 2p^-$ transition can be observed throughout the entire time delay range, due to the direct creation of excitons. The change in conductivity gradually decreases with time through interband recombination. Through optical pump/THz-probe spectroscopy of magneto-excitons, in addition to exploring the internal structure of excitons, we are able to explore the dynamics of their formation and transitions from bound excitons to unbound electron-hole pairs.

Many-Body Effects of High-Density Excitons in High Magnetic Fields

With increasing *e-h* pair density, excitons can ionize to form an *e-h* plasma, where the electrons and holes are no longer bound to one another. This transition from excitons to an *e-h* plasma is referred to as the excitonic Mott transition. Here, we will focus on this high-density regime but in high B, where many-body effects can lead to many exotic behaviors, which are not observed in

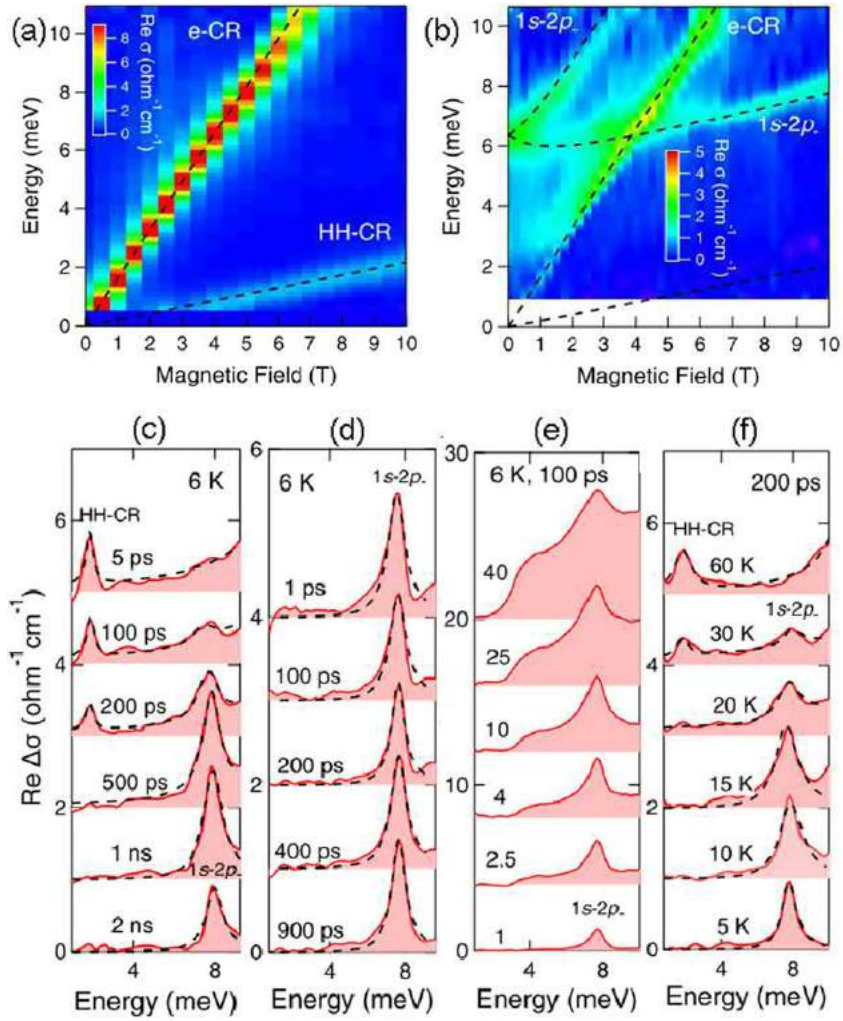

Fig. 10 (color online). Photoinduced change in conductivity, $\mathrm{Re}\Delta\sigma(\omega)$, as a function of B, at a time delay of (a) 15 ps and (b) 1 ns after nonresonant excitation of undoped GaAs quantum wells. Conductivity spectral evolution after (c) nonresonant and (d) resonant excitations with a fluence of 200 nJcm^{-2} at 9 T. (e) Pump fluence dependent conductivity spectra at 9 T, 5 K, and 100 ps after resonant excitation. (f) Temperature-dependent conductivity spectra at 9 T and 200 ps after resonant excitation. Reproduced (adapted) with permission from Zhang, Q., Wang, Y., Gao, W., *et al.*, 2016. Stability of high-density two-dimensional excitons against a Mott transition in high magnetic fields probed by coherent terahertz spectroscopy. Physical Review Letters 117, 207402. Copyright 2016 by the American Physical Society.

low-density interband and intraband absorption measurements. For example, in the high-density regime of a magnetoplasma, a macroscopic coherent dipole polarization can form initially from quantum fluctuations and finally release coherent, intense superfluorescent pulses of radiation. Another interesting phenomenon that occurs in the high-B limit of 2D magnetoexcitons is that they become ultrastable against the Mott transition due to the presence of a "hidden symmetry" that cancels the interactions between excitons.

Hidden Symmetry of 2D Magnetoexcitons

The hidden symmetry is unique to 2D magnetoexcitons and can be observed under the following conditions: the Coulomb interaction is charge-symmetric, and mixing of Landau levels can be neglected. Experimentally, the hidden symmetry can be observed when the filling factors of electrons and holes satisfy $0 \leq \nu_e, \nu_h \leq 2$, i.e., only the lowest Landau level is occupied. Under these conditions, the Coulomb interaction between excitons vanishes, and the 2D magnetoexcitons become ultrastable.

One example for the existence of hidden symmetry in 2D magnetoexcitons is shown in Fig. 10(e). Under 9 T, 5 K, and 100 ps after a resonant excitation of undoped GaAs quantum wells, both the intensity and linewidth of the $1s \rightarrow 2p^-$ peak increase with the pump fluence. Under the highest pump fluence (40×200 nJcm^{-2}), corresponding to a pair density of 10^{11} cm^{-2} per quantum well and a filling factor of $\nu_e < 2$, the intraexcitonic feature still remains and no CR is observed. Such a pair density is almost comparable to the Mott density at 0 T, so the absence of the Mott transition does confirm that the 2D magnetoexcitons are very

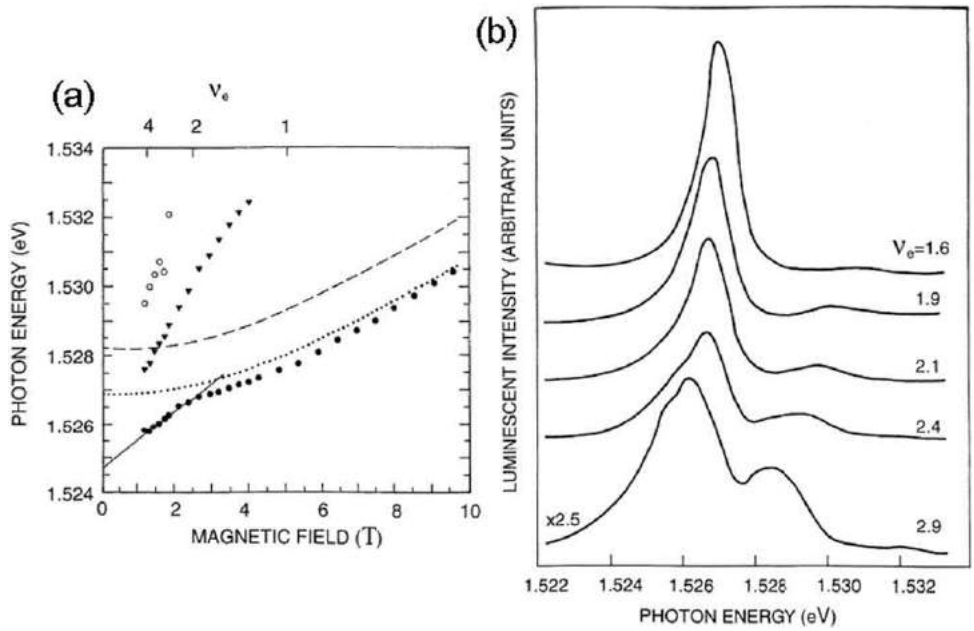

Fig. 11 (color online). (a) Photon energies of PL peak as a function of B at 4.2 K of a symmetric 20 nm quantum well with an electron density of 1.2×10^{11} cm^{-2}. The solid line is the best fit to the linear shift of the lowest Landau level at low fields. The dashed line is the B dependence of the exciton at a low electron density, and the dotted line is that of the trion. (b) Photoluminescence spectra in the vicinity of $\nu_e=2$. Reproduced (adapted) with permission from Rashba, E.I., Sturge, M.D., Yoon, H.W., Pferffer, L.N., 2000. Hidden symmetry and the magnetically induced "Mott transition" in quantum wells containing an electron gas. Solid State Communications 114, 593–596. Copyright (2000) by Elsevier.

stable against the density-driven dissociation. However, the 2D magnetoexcitons dissociate under thermal excitation, as shown in Fig. 10(f). At a low temperature, the $1s \rightarrow 2p^-$ peak dominates the spectrum, and disappears with increasing temperature. At a high temperature, only the CR feature can be observed.

The hidden symmetry was also observed in quantum wells containing an electron gas. As shown in Fig. 11(a), at a certain B, the photon energy of the lowest transition changes from a linear dependence to a quadratic dependence on B. Above this critical B, the emission energy is very close to that of the singlet trion transition observed in the same quantum well at a low electron density. Different from a Mott transition, which is induced by the changes in screening and band-filling, this magnetically induced transition at a certain field results from the changes of the dynamic symmetry of the system, or hidden symmetry. Therefore, it is called a symmetry-driven transition.

Fig. 11(b) shows the effect of hidden symmetry in the vicinity of $\nu_e=2$. At a relatively large filling factor for electrons, such as $\nu_e=2.9$, the PL spectrum consists of three peaks, from the lowest Landau level transition, the blue-shifted cyclotron resonance, and a red-shifted satellite, possibly due to a phonon-assisted transition. With decreasing ν_e, the shape of the spectrum changes and the linewidth becomes sharp. That is because, under a critical B where $\nu_e<2$, the frequency of the optical transition is not affected by the presence of electrons or holes, and the transition remains sharp.

Superfluorescence from High-Density 2D Magnetoexcitons

The concept of superradiance (SR) was proposed by Dicke in 1954, where he studied the radiative decay of an ensemble of N_{atom} incoherently excited two-level atoms, as shown in Fig. 12(a). If all of the atoms are confined in a region smaller than λ, where λ is the wavelength corresponding to the level separation, the emission behavior dramatically depends on the density of atoms. At a low density, the atoms do not interact with each other, and they radiate spontaneously with an intensity $\propto N_{\text{atom}}$ and a decay rate T_1^{-1}, where T_1 is the spontaneous radiative decay time of an isolated atom. However, at high density, all of the atoms lock in phase by photon exchange, and a giant dipole $P \sim N_{\text{atom}}d$ (d: individual atomic dipole) develops during a decay time τd. The macroscopic dipole leads to an intense coherent burst of radiation, with a decay rate $\Gamma \sim N_{\text{atom}}T_1^{-1}$, N_{atom} times faster than that of an individual atom, and a short pulse duration, $\tau_P \propto 1/N_{\text{atom}}$. Therefore, the intensity scales as $\propto N_{\text{atom}}/\tau_P \propto N_{\text{atom}}^2$, a hallmark of coherent emission.

Different from SR, where the polarization is generated by an external laser field, superfluorescence (SF) is a process in which a macroscopic polarization spontaneously develops from quantum fluctuations, which has been widely observed in atomic, molecular and solid systems. Very recently, SF was observed in InGaAs/GaAs quantum wells in quantizing B, in a setup schematically shown in Fig. 12(b). Interband transitions were studied through time-resolved PL, time-integrated PL and pump-probe spectroscopy. Since optical gain exists only for electromagnetic waves propagating along the quantum well plane, SF can be observed in the edge collection while ordinary spontaneous emission (SE) is collected in the center fiber.

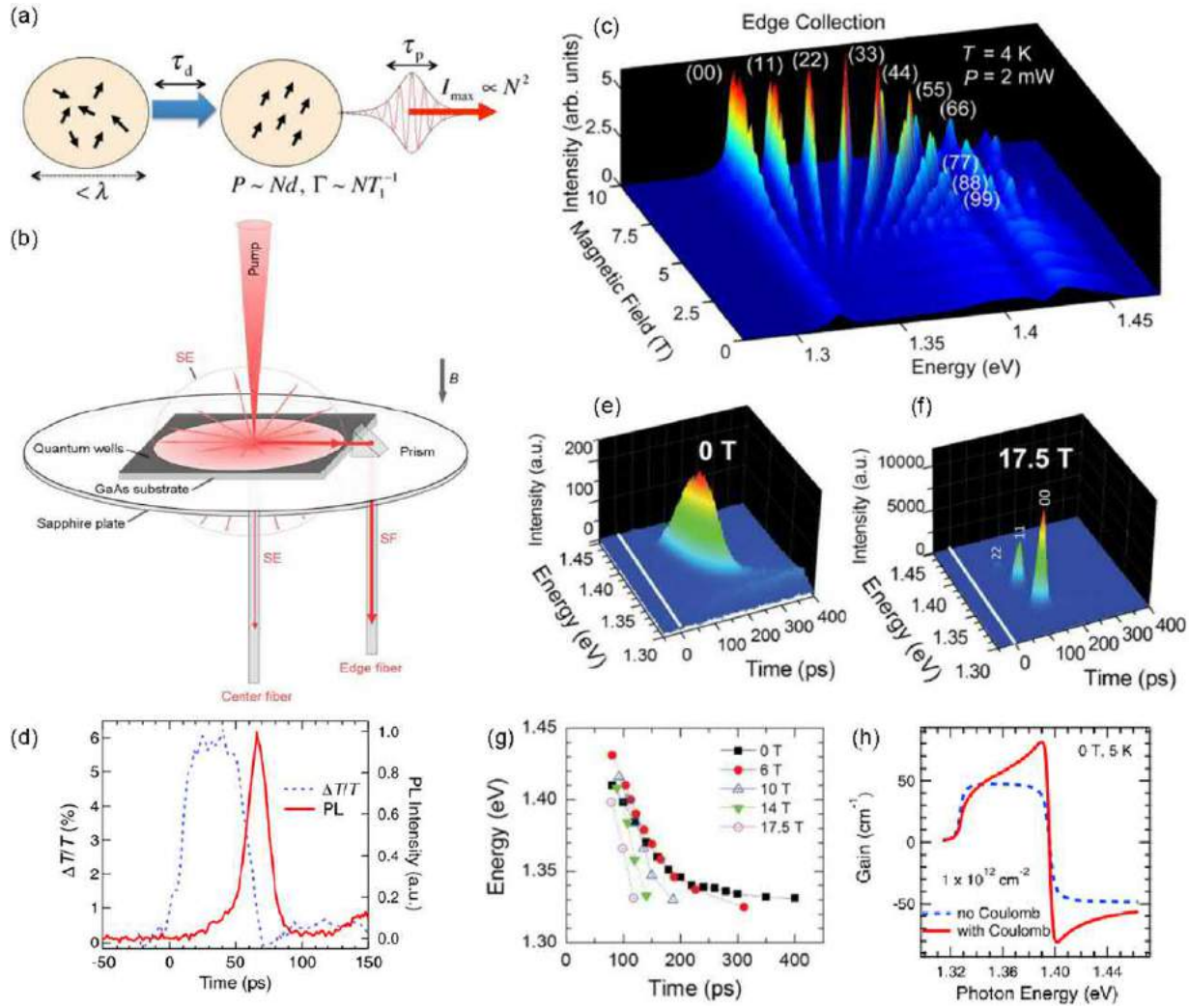

Fig. 12 (color online). (a) Basic processes and characteristics of SF. (b) Schematic diagram of the experimental geometry for SF observation from photoexcited semiconductor quantum wells in a *B*. (c) *B* dependence of time-integrated PL collected with the edge fiber at 4 K with an average excitation laser power of 2 mW. Reproduced (adapted) with permission from Cong, K., Wang, Y., Kim, J.H., *et al.*, 2015. Superfluorescence from photoexcited semiconductor quantum wells: Magnetic field, temperature, and excitation power dependence. Physical Review B 91, 235448. Copyright 2015 by the American Physical Society. (d) Simultaneously taken pump-probe and time-resolved PL data for the (22) transition at 17 T and 5 K, showing direct evidence of SF in the GaAs/InGaAs quantum well system. (a) and (d) reproduced (adapted) with permission from Cong, K., Zhang, Q., Wang, Y., *et al.*, 2016. Dicke superradiance in solids [Invited]. Journal of the Optical Society of America B 33, C80-C101. Copyright (2016) by the Optical Society of America. Time-resolved PL spectra at (e) 0 T and (f) 17.5 T with an excitation power of 2 mW at 5 K. (g) Peak shift of emission as a function of time at different magnetic fields. (h) Theoretical calculations of Coulomb-induced many-body enhancement of gain at the Fermi energy at 0 T. Reproduced (adapted) with permission from Kim, J.H., Noe II, G.T., McGill, S.A., *et al.*, 2013. Fermi-edge superfluorescence from a quantum-degenerate electron-hole gas. Scientific Reports 3, 3283. Copyright (2013) by Nature Publishing Group.

Fig. 12(c) shows *B*-dependent time-integrated SF emission at a pump intensity of 2 mW and a temperature of 4 K. At a low *B* (<4 T), SF is characterized by two peaks at around 1.32 eV and 1.43 eV, corresponding to the heavy-hole and light-hole transitions, respectively. With increasing *B*, emission peaks due to the $(N,M)=(00)$, (11), …, heavy-hole transitions, are observed, with stronger intensities and narrower widths.

In order to obtain direct evidence for SF emission, time-resolved pump-probe and PL measurements were performed. **Fig. 12(d)** shows pump-probe differential transmission and time-resolved PL data for the (22) transition at 17 T and 5 K. The pump creates the population inversion, as seen from the differential transmission, but at a time delay around 70 ps, it suddenly drops to zero, and at the same time a strong pulse of emission appears, as indicated by the time-resolved PL data.

The SF emission at 0 T is characterized by a continuous burst after a certain time delay, as shown in **Fig. 12(e)**, while the effects of quantization are obvious at a high *B*, as seen in **Fig. 12(f)**. By analyzing the emission time for different energies at various magnetic fields, as shown in **Fig. 12(g)**, a sequential manner of SF emission is found: SF from the highest occupied states is emitted

first, which is followed by emission from lower energy states. In addition, it can be seen that the SF emission occurs at shorter delay times with increasing B.

These results can be explained by the excitonic enhancement of gain near the Fermi energy in a high-density electron-hole system, as shown in **Fig. 12(h)**. After relaxation and thermalization, degenerate Fermi gases form inside the conduction and valence bands each with quasi-Fermi energies. The recombination gain just below the quasi-Fermi energy is enhanced due to Coulomb interactions among carriers, which causes a SF burst to form at the Fermi edge. After the SF emission, the population is depleted, leading to a decreased Fermi energy. Thus, as time progresses, the Fermi level moves toward the band edge continuously. As a result, a continuous SF emission is observed at zero field and a series of sequential SF bursts appear in a magnetic field.

Summary

In this review, we have shown that the presence of excitons significantly modifies the optical response of semiconductor materials. Magneto-optical spectroscopy provides a powerful experimental means to determine the basic parameters of excitons such as the reduced mass, binding energy, Bohr radius, g-factor, internal structure, and formation dynamics. Depending on the dimensionality of excitons, these parameters can vary greatly, and we have presented experimental observations in magnetic fields for varied systems, including bulk semiconductors, 2D semiconductor quantum wells, monolayer TMDs, 1D quantum wires, SWCNTs, and semiconductor quantum dots. In addition to characterizing exciton parameters, magneto-optical spectroscopy of semiconductors can reveal exotic phenomena such as the "hidden symmetry" of 2D excitons in high magnetic fields and superfluorescence from a semiconductor magnetoplasma. The powerful experimental tools we have discussed will surely continue to reveal new phenomena related to excitons in semiconductors, including intriguing features related to the bosonic nature of excitons.

Further Reading

Akimoto, O., Hasegawa, H., 1967. Interband optical transitions in extremely anisotropic semiconductors. II. Coexistence of exciton and the Landau levels. Journal of the Physical Society of Japan 22, 181–191.

Černe, J., Kono, J., Sherwin, M.S., *et al.*, 1996. Terahertz dynamics of excitons in GaAs/AlGaAs quantum wells. Physical Review Letters 77, 1131–1134.

Cong, K., Wang, Y., Kim, J.H., *et al.*, 2015. Superfluorescence from photoexcited semiconductor quantum wells: magnetic field, temperature, and excitation power dependence. Physical Review B 91, 235448.

Cong, K., Zhang, Q., Wang, Y., *et al.*, 2016. Dicke superradiance in solids [Invited]. Journal of the Optical Society of America B 33, C80–C101.

Hayden, R.K., Uchida, K., Miura, N., *et al.*, 1998. High field magnetoluminescence spectroscopy of self-assembled (InGa)As quantum dots on high index planes. Physica B: Condensed Matter 246, 93–96.

Kim, J.H., Noe II, G.T., McGill, S.A., *et al.*, 2013. Fermi-edge superfluorescence from a quantum-degenerate electron-hole gas. Scientific Reports 3, 3283.

Kono, J., Su, M.Y., Inoshita, T., *et al.*, 1997. Resonant terahertz optical sideband generation from confined magnetoexcitons. Physical Review Letters 79, 1758–1761.

MacDonald, A.H., Ritchie, D.S., 1986. Hydrogenic energy levels in two dimensions at arbitrary magnetic fields. Physical Review B 33, 8336–8344.

Miyata, A., Mitioglu, A., Plochocka, P., *et al.*, 2015. Direct measurement of the exciton binding energy and effective masses for charge carriers in organic-inorganic tri-halide perovskites. Nature Physics 11, 582–588.

Muro, K., Nisida, Y., 1976. Far-infrared magneto-absorptions in photo-excited germanium. Journal of the Physical Society of Japan 40, 1069–1077.

Nagamune, Y., Arakawa, Y., Tsukamoto, S., Nishioka, M., 1992. Photoluminescence spectra and anisotropic energy shift of GaAs quantum wires in high magnetic fields. Physical Review Letters 69, 2963–2966.

Rashba, E.I., Sturge, M.D., Yoon, H.W., Pferffer, L.N., 2000. Hidden symmetry and the magnetically induced "Mott transition" in quantum wells containing an electron gas. Solid State Communications 114, 593–596.

Salib, M.S., Nickel, H.A., Herold, G.S., Petrou, A., McComber, B.D., 1996. Observation of internal transitions of confined excitons in GaAs/AlGaAs quantum wells. Physical Review Letters 77, 1135–1138.

Shaver, J., Kono, J., Portugall, O., *et al.*, 2007. Magnetic brightening of carbon nanotube photoluminescence through symmetry breaking. Nano Letters 7, 1851–1855.

Srivastava, A., Htoon, H., Klimov, V.I., Kono, J., 2008. Direct observation of dark excitons in individual carbon nanotubes: inhomogeneity in the exchange splitting. Physical Review Letters 101, 087402.

Stier, A.V., McCreary, K.M., Jonker, B.T., Kono, J., Crooker, S., 2016. Exciton diamagnetic shifts and valley Zeeman effects in monolayer WS_2 and MoS_2 to 65 Tesla. Nature Communications 7, 10643.

Tarucha, S., Okamoto, H., Iwasa, Y., Miura, N., 1984. Exciton binding energy in GaAs quantum wells deduced from magneto-optical absorption measurement. Solid State Communications 52, 815–819.

Zaric, S., Ostojic, G.N., Shaver, J., *et al.*, 2006. Excitons in carbon nanotubes with broken time-reversal symmetry. Physical Review Letters 96, 016406.

Zhang, Q., Wang, Y., Gao, W., *et al.*, 2016. Stability of high-density two-dimensional excitons against a Mott transition in high magnetic fields probed by coherent terahertz spectroscopy. Physical Review Letters 117, 207402.

Ultrafast Studies of Semiconductors

J Shah, Arlington, VA, USA

Introduction

Linear optical spectroscopy, using the techniques of absorption, transmission, reflection and light scattering, has provided invaluable information about the electronic and vibrational properties of atoms, molecules, and solids. Optical techniques also possess some additional unique strengths: the ability to

1. generate nonequilibrium distributions functions of electrons, holes, excitons, phonons, etc. in solids;
2. determine the distribution functions by optical spectroscopy;
3. determine the nonequilibrium distribution functions on femtosecond timescales;
4. determine femtosecond dynamics of carrier and exciton transport and tunneling; and
5. investigate interactions between various elementary excitations as well as many-body processes in semiconductors.

Researchers have exploited these unique strengths to gain fundamental new insights into nonequilibrium, nonlinear, and transport physics of semiconductors and their nanostructures over the past four decades.

A major focus of these efforts has been devoted to understanding how a semiconductor in thermodynamic equilibrium, excited by an ultrashort optical pulse, returns to the state of thermodynamic equilibrium. Four distinct regimes can be identified:

(a) the coherent regime in which a well-defined phase relationship exists between the excitation created by the optical pulse and the electromagnetic (optical) field creating it;
(b) the nonthermal regime in which the coherence (well-defined phase relationships) has been destroyed by various collision and interference processes but the distribution function of excitations is nonthermal (i.e., cannot be described by a Maxwell–Boltzmann distribution or its quantum (degenerate) counterparts);
(c) the hot carrier regime in which the distributions for various excitations are thermal, but with different characteristic temperatures for different excitations and thermal bath; and
(d) the isothermal regime in which the excitations and the thermal bath are at the same temperature but there is an excess of excitation (e.g., electron–hole pairs) compared to thermodynamic equilibrium.

Various physical processes take the semiconductor from one regime to the next, and provide information not only about the fundamental physics of semiconductors but also about the physics and ultimate performance limits of electronic, optoelectronic, and photonic devices.

In addition to coherent and relaxation dynamics, ultrafast studies of semiconductors provide new insights into tunneling and transport of carriers, and demonstrate novel quantum mechanical phenomena and coherent control in semiconductors.

The techniques used for such studies have also advanced considerably over the last four decades. Ultrafast lasers with pulse widths corresponding to only a few optical cycles (1 optical cycle ≈ 2.7 fs at 800 nm) have been developed. The response of the semiconductor to an ultrafast pulse, or a multiple of phase-locked pulses, can be investigated by measuring the dynamics of light emitted by the semiconductor using a streak camera (sub-ps time resolution) or a nonlinear technique such as luminescence up-conversion (time resolution determined by the laser pulse width for an appropriate nonlinear crystal). The response of the semiconductor can also be investigated using a number of two or three beam pump-probe techniques such as transmission, reflection, light scattering or four-wave mixing (FWM). Both the amplitude and the phase of the emitted radiation or the probe can be measured by phase-sensitive techniques such as spectral interferometry. The lateral transport of excitation can be investigated by using time-resolved spatial imaging and the vertical transport and tunneling of carriers can be investigated using the technique of 'optical markers'. Electro-optic sampling can be used for transmission/reflection studies or for measuring THz response of semiconductors. More details on these techniques, and indeed many topics discussed in this brief article, can be found in the Further Reading.

Coherent Dynamics

The coherent response of atoms and molecules is generally analyzed for an ensemble of independent (noninteracting) two-level systems. The statistical properties of the ensemble are described in terms of the density matrix operator whose diagonal components relate to population of the eigenstates and off-diagonal components relate to coherence of the superposition state. The time evolution of the density matrix is governed by the Liouville variant of the Schrödinger equation $i\hbar\dot{\rho} = [H, \rho]$ where the system Hamiltonian $H = H_0 + H_{\text{int}} + H_R$ is the sum of the unperturbed, interaction (between the radiation field and the two-level system), and relaxation Hamiltonians, respectively. Using a number of different assumptions and approximations, this equation of motion is transformed into the optical Bloch equations (OBE), which are then used to predict the coherent response of the system, often using iterative procedures, and analyze experimental results.

 doi:10.1016/B978-0-12-809283-5.00916-2

Semiconductors are considerably more complex. Coulomb interaction profoundly modifies the response of semiconductors, not only in the linear regime (e.g., strong exciton resonance in the absorption spectrum) but also in the coherent and nonlinear regime. The influence of Coulomb interaction may be introduced by renormalizing the electron and hole energies (i.e., introducing excitons), and by renormalizing the field–matter interaction strength by introducing a renormalized Rabi frequency. These changes and the relaxation time approximation for the relaxation Hamiltonian lead to an analog of the optical Bloch equations, the semiconductor Bloch equations which have been very successful in analyzing the coherent response of semiconductors.

In spite of this success, it must be stressed that the relaxation time approximation, and the Boltzmann kinetic approach on which it is based, are not valid under all conditions. The Boltzmann kinetic approach is based on the assumption that the duration of the collision is much shorter than the interval between the collisions; i.e., the collisions are instantaneous. In the non-Markovian regime where these assumptions are not valid, each collision does not strictly conserve the energy and momentum and the quantum kinetic approach becomes more appropriate. This is true for photo-excited semiconductors as well as for the quantum transport regime in semiconductors. Also, semiconductor Bloch equations have been further extended by including four-particle correlations. Finally, the most general approach to the description of the nonequilibrium and coherent response of semiconductors following excitation by an ultrashort laser pulse is based on nonequilibrium Green's functions.

Experimental studies on the coherent response of semiconductors can be broadly divided into three categories:

1. investigation of novel aspects of coherent response not found in other simpler systems such as atoms;
2. investigation of how the initial coherence is lost, to gain insights into dephasing and decoherence processes; and
3. coherent control of processes in semiconductors using phase-locked pulses.

Numerous elegant studies have been reported. This section presents some observations on the trends in this field.

Historically, the initial experiments focused on the decay of the coherent response (time-integrated FWM as a function of delays between the pulses), analyzed them in terms of the independent two-level model, and obtained very useful information about various collision processes and rates (exciton–exciton, exciton–carrier, carrier–carrier, exciton–phonon, etc.). It was soon realized, however, that the noninteracting two-level model is not appropriate for semiconductors because of the strong influence of Coulomb interactions. Elegant techniques were then developed to explore the nature of coherent response of semiconductors. These included investigation of time- and spectrally resolved coherent response, and of both the amplitude and phase of the response to complement intensity measurements. These studies provided fundamental new insights into the nature of semiconductors and many-body processes and interactions in semiconductors. Many of these observations were well explained by the semiconductor Bloch equations. When lasers with <10 fs pulse widths became laboratory tools, the dynamics was explored on a time-scale much shorter than the characteristic phonon oscillation period (≈ 115 fs for GaAs longitudinal optical phonons), the plasma frequency (≈ 150 fs for a carrier density of 5×10^{17} cm^{-3}), and the typical electron–phonon collision interval (≈ 200 fs in GaAs). Some remarkable features of quantum kinetics, such as reversal of the electron–phonon collision, memory effects, and energy nonconservation, were demonstrated. More recent experiments have demonstrated the influence of four-particle correlations on coherent nonlinear response of semiconductors such as GaAs in the presence of a strong magnetic field.

The ability to generate phase-locked pulses with controllably variable separation between them provides an exciting opportunity to manipulate a variety of excitations and processes within a semiconductor. If the separation between the two phase-locked pulses is adjusted to be less than the dephasing time of the system under investigation, then the amplitudes of the excitations produced by the two phase-locked pulses can interfere either constructively or destructively depending on the relative phase of the carrier waves in the two pulses. The nature of interference changes as the second pulse is delayed over an optical period. A number of elegant experiments have demonstrated coherent control of exciton population, spin orientation, resonant emission, and electrical current. Phase-sensitive detection of the linear or nonlinear emission provides additional insights into different aspects of semiconductor physics.

These experimental and theoretical studies of ultrafast coherent response of semiconductors have led to fundamental new insights into the physics of semiconductors and their coherence properties. Discussion of novel coherent phenomena is given in a subsequent section. Coherent spectroscopy of semiconductors continues to be a vibrant research field.

Incoherent Relaxation Dynamics

The qualitative picture that emerges from investigation of coherent dynamics can be summarized as follows. Consider a direct gap semiconductor excited by an ultrashort laser pulse of duration τ_L (with a spectral width $h\Delta\nu_L$) centered at photon energy $h\nu_L$ larger than the semiconductor bandgap E_g. At the beginning of the pulse the semiconductor does not know the pulse duration so that coherent polarization is created over a spectral region much larger than $h\Delta\nu_L$. If there are no phase-destroying events during the pulse (dephasingrate$1/\tau_D \ll 1/\tau_L$), then a destructive interference destroys the coherent polarization away from $h\nu_L$ with increasing time during the excitation pulse. Thus, the coherent polarization exists only over the spectral region $h\Delta\nu_L$ at the end of the pulse. This coherence will be eventually destroyed by collisions and the semiconductor will be left in an incoherent (off-diagonal elements of the density matrix are 0) nonequilibrium state with peaked electron and hole population distributions whose central energies and energy widths are determined by $h\nu_L$, $h\Delta\nu_L$ and E_g, if the energy and momentum relaxation rates $(1/\tau_E, 1/\tau_M)$ are much smaller than the dephasing rates.

The simplifying assumptions that neatly separate the excitation, dephasing and relaxation regimes are obviously not realistic. A combination of all these (and some other) physical parameters will determine the state of a realistic system at the end of the exciting pulse. However, after times $\sim\tau_D$ following the laser pulse, the semiconductor can be described in terms of electron and hole populations whose distribution functions are not Maxwell–Boltzmann or Fermi–Direct type, but nonthermal in a large majority of cases in which the energy and momentum relaxation rates are much smaller than the dephasing rates. This section discusses the dynamics of this incoherent state.

Extensive theoretical work has been devoted to quantitatively understand these initial athermal distributions and their further temporal evolution. These include Monte Carlo simulations based on the Boltzmann kinetic equation approach (typically appropriate for time-scales longer than $\sim$100 fs, the carrier–phonon interaction time) as well as the quantum kinetic approach (for times typically less than $\sim$100 fs) discussed above. We present below some observations on this vast field of research.

Nonthermal Regime

A large number of physical processes determines the evolution of the nonthermal electron and hole populations generated by the optical pulse. These include electron–electron, hole–hole, and electron–hole collisions, including plasma effects if the density is high enough, intervalley scattering in the conduction band and intervalence band scattering in valence bands, intersub-band scattering in quantum wells, electron–phonon, and hole–phonon scattering processes. The complexity of the problem is evident when one considers that many of these processes occur on the same time-scale. Of these myriad processes, only carrier–phonon interactions and electron–hole recombinations (generally much slower) can alter the total energy in electronic systems. Under typical experimental conditions, the redistribution of energy takes place before substantial transfer of energy to the phonon system. Thus, the nonthermal distributions first become thermal distributions with the same total energy. Investigation of the dynamics of how the nonthermal distribution evolves into a thermal distribution provides valuable information about the nature of various scattering processes (other than carrier–phonon scattering) described above.

This discussion of separating the processes that conserve energy in the electronic system and those that transfer energy to other systems is obviously too simplistic and depends strongly on the nature of the problem one is considering. The challenge for the experimentalist is to devise experiments that can isolate these phenomena so each can be studied separately. If the laser energy is such that $h\nu_L - E_g < \hbar\omega_{LO}$, the optical phonon energy, then majority of the photo-excited electrons and holes do not have sufficient energy to emit an optical phonon, the fastest of the carrier–phonon interaction processes. The initial nonthermal distribution is then modified primarily by processes other than phonon scattering and can be studied experimentally without the strong influence of the carrier–phonon interactions. Such experiments have indeed been performed both in bulk and quantum well semiconductors. These experiments have exhibited spectral hole burning in the pump-probe transmission spectra and thus demonstrated that the initial carrier distributions are indeed nonthermal and evolve to thermal distributions. Such experiments have provided quantitative information about various carrier–carrier scattering rates as a function of carrier density. In addition, experiments with $h\nu_L - E_g > \hbar\omega_{LO}$ and $h\nu_L - E_g >$ intervalley separation have provided valuable information about intravalley as well as intervalley electron–phonon scattering rates. Experiments have been performed in a variety of different semiconductors, including bulk semiconductors and undoped and modulation-doped quantum wells. The latter provide insights into intersub-band scattering processes and quantitative information about the rates.

These measurements have shown that under typical experimental conditions, the initial nonthermal distribution evolves into a thermal distribution in times $<$ or $\sim$ 1 ps. However, the characteristic temperatures of the thermal distributions can be different for electrons and holes. Furthermore, different phonons may also have different characteristic temperatures, and may even be nonthermal even when the electronic distributions are thermal. This is the hot carrier regime that is discussed in the next subsection.

Hot Carrier Regime

The hot carriers, with characteristic temperatures T_c (T_e and T_h for electrons and holes, respectively), have high energy tails in the distribution functions that extend several kT_c higher than the respective Fermi energies. Some of these carriers have sufficient energy to emit an optical phonon. Since the carrier–optical phonon interaction rates and the phonon energies are rather large, this may be the most effective energy loss mechanism in many cases even though the number of such carriers is rather small. The emitted optical phonons have relatively small wavevectors and nonequilibrium populations larger than expected for the lattice temperature. These optical phonons are often referred to as hot phonons although nonequilibrium phonons may be a better term since they probably have nonthermal distributions. The dynamics of hot phonons will be discussed in the next subsection, but we mention here that in case of a large population of hot phonons, one must consider not only phonon emission but also phonon absorption. Acoustic phonons come into the picture at lower temperatures when the fraction of carriers that can emit optical phonons is extremely small or negligible, and also in the case of quantum wells with sub-band energy separation smaller than optical phonons.

The dynamics of hot carriers have been studied extensively to obtain a deeper understanding of carrier–phonon interactions and to obtain quantitative measures of the carrier–phonon interaction rates. Time resolved luminescence and pump-probe transmission spectroscopy are the primary tools used for such studies. For semiconductors like GaAs, the results indicate that polar optical phonon scattering (longitudinal optical phonons) dominates for electrons whereas polar and nonpolar optical phonon

scattering contribute for holes. Electron–hole scattering is sufficiently strong in most cases to maintain a common electron and hole temperature for times $> \sim 1$ ps. Since many of these experiments are performed at relatively high densities, they provide important information about many-body effects such as screening. Comparison of bulk and quasi-two-dimensional semiconductors (quantum wells) has been a subject of considerable interest. The consensus appears to be that, in spite of significant differences in the nature of electronic states and phonons, similar processes with similar scattering rates dominate both types of semiconductors. These studies reveal that the energy loss rates are influenced by many factors such as the Pauli exclusion principle, degenerate (Fermi–Dirac) statistics, hot phonons, and screening and many-body aspects.

Hot carrier distribution can be generated not only by photo-excitation, but also by applying an electric field to a semiconductor. Although the process of creating the hot distributions is different, the processes by which such hot carriers cool to the lattice temperature are the same in two cases. Therefore, understanding obtained through one technique can be applied to the other case. In particular, the physical insights obtained from the optical excitation case can be extremely valuable for many electronic devices that operate at high electric fields, thus in the regime of hot carriers. Another important aspect is that electrons and holes can be investigated separately by using different doping. Furthermore, the energy loss rates can be determined quantitatively because the energy transferred from the electric field to the electrons or holes can be accurately determined by electrical measurements.

Isothermal Regime

Following the processes discussed above, the excitations in the semiconductor reach equilibrium with each other and the thermal bath. Recombination processes then return the semiconductor to its thermodynamic equilibrium. Radiative recombination typically occurs over a longer time-scale but there are some notable exceptions such as radiative recombination of excitons in quantum wells occurring on picosecond time-scales.

Hot Phonons

As discussed above, a large population of nonequilibrium optical phonons is created if a large density of hot carriers has sufficient energy to emit optical phonons. Understanding the dynamics of these nonequilibrium optical phonons is of intrinsic interest.

At low lattice temperatures, the equilibrium population of optical phonons is insignificant. The optical phonons created as a result of photo-excitation occupy a relatively narrow region of wavevector (k) space near $k=0$ ($\sim 1\%$ of the Brillouin zone). This nonequilibrium phonon distribution can spread over a larger k-space by various processes, or anharmonically decay into multiple large-wavevector acoustic phonons. These acoustic phonons then scatter or decay into small-wavevector, low-energy acoustic phonons that are eventually thermalized.

Pump-probe Raman scattering provides the best means of investigating the dynamics of nonequilibrium optical phonons. Such studies, for bulk and quantum well semiconductors, have provided invaluable information about femtosecond dynamics of phonons. In particular, they have provided the rate at which the nonequilibrium phonons decay. However, this information is for one very small value of the phonon wavevector so the phonon distribution function within the optical phonon branch is not investigated. The large-wavevector acoustic phonons are even less accessible to experimental techniques. Measurements of thermal phonon propagation provide some information on this subject. Finally, the nature of phonons is considerably more complex in quantum wells, and some interesting results have been obtained on the dynamics of these phonons.

Tunneling and Transport Dynamics

Ultrafast studies have also made important contributions to tunneling and transport dynamics in semiconductors. Time-dependent imaging of the luminescing region of a sample excited by an ultrashort pulse provides information about spatial transport of carriers. Such a technique was used, for example, to demonstrate negative absolute mobility of electrons in p-modulation-doped quantum wells. The availability of near-field microscopes enhances the resolution of such techniques to measure lateral transport to the subwavelength case.

A different technique of 'optical markers' has been developed to investigate the dynamics of transport and tunneling in a direction perpendicular to the surface (vertical transport). This technique relies heavily on fabrication of semiconductor nanostructures with desired properties. The basic idea is to fabricate a sample in which different spatial regions have different spectral signatures. Thus if the carriers are created in one spatial region and are transported to another spatial region under the influence of an applied electric field, diffusion, or other processes, the transmission, reflection, and/or luminescence spectra of the sample change dynamically as the carriers are transported.

This technique has provided new insights into the physics of transport and tunneling. Investigation of perpendicular transport in graded-gap superlattices showed remarkable time-dependent changes in the spectra, and analysis of the results provided new insight into transport in a semiconductor of intermediate disorder. Dynamics of carrier capture in quantum wells from the barriers provided information not only about the fundamentals of capture dynamics, but also about how such dynamics affects the performance of semiconductor lasers with quantum well active regions.

The technique of optical markers has been applied successfully to investigate tunneling between the two quantum wells in an a-DQW (asymmetric double quantum well) structure in which two quantum wells of different widths (and hence different

energy levels) are separated by a barrier. The separation between the energy levels of the system can be varied by applying an electric field perpendicular to the wells. In the absence of resonance between any energy levels, the wavefunction for a given energy level is localized in one well or the other. In this nonresonant case, it is possible to generate carriers in a selected quantum well by proper optical excitation. Dynamics of transfer of carriers to the other quantum well by tunneling can then be determined by measuring dynamic changes in the spectra. Such measurements have provided much insight into the nature and rates of tunneling, and demonstrated phonon resonances. Resonant tunneling, i.e., tunneling when two electronic levels are in resonance and are split due to strong coupling, has also been investigated extensively for both electrons and holes. One interesting insight obtained from such studies is that tunneling and relaxation must be considered in a unified framework, and not as sequential events, to explain the observations.

Novel Coherent Phenomena

Ultrafast spectroscopy of semiconductors has provided valuable insights into many other areas. We conclude this article by discussing some examples of how such techniques have demonstrated novel physical phenomena.

The first example once again considers an a-DQW biased in such a way that the lowest electron energy level in the wide quantum well is brought into resonance with the first excited electron level in the narrow quantum well. The corresponding hole energy levels are not in resonance. By choosing the optical excitation photon energy appropriately, the hole level is excited only in the wide well and the resonant electronic levels are excited in a linear superposition state such that the electrons at $t=0$ also occupy only the wide quantum well. Since the two electron eigenstates have slightly different energies, their temporal evolutions are different with the result that the electron wavepacket oscillates back and forth between the two quantum wells in the absence of damping. The period of oscillation is determined by the splitting between the two resonant levels and can be controlled by an external electric field. Coherent oscillations of electronic wavepackets have indeed been experimentally observed by using the coherent nonlinear technique of four-wave mixing. Semiconductor quantum wells provide an excellent flexible system for such investigations.

Another example is the observation of Bloch oscillations in semiconductor superlattices. In 1928 Bloch demonstrated theoretically that an electron wavepacket composed of a superposition of states from a single energy band in a solid undergoes a periodic oscillation in energy and momentum space under certain conditions. An experimental demonstration of Bloch oscillations became possible only recently by applying ultrafast optical techniques to semiconductor superlattices. The large period of a superlattice (compared to the atomic period in a solid) makes it possible to satisfy the assumptions underlying Bloch's prediction. The experimental demonstration of Bloch oscillations was also performed using four-wave mixing techniques. This provides another prime example of the power of ultrafast optical studies.

Summary

Ultrafast spectroscopy of semiconductors and their nanostructures is an exciting field of research that has provided fundamental insights into important physical processes in semiconductors. This article has attempted to convey the breadth of this field and the diversity of physical processes addressed by this field. Many exciting developments in the field have been omitted out of necessity. The author hopes that this brief article inspires the readers to explore some of the Further Reading.

Further Reading

Allen, L., Eberly, J.H., 1975. Optical Resonance and Two-Level Atoms. New York: Wiley Interscience.
Chemla, D.S., 1999. In: Garmire, E., Kost, A.R. (Eds.), Nonlinear Optics in Semiconductors. New York: Academic Press, pp. 175–256 (Chapter 3).
Chemla, D.S., Shah, J., 2001. Many-body and correlation effects in semiconductors. Nature 411, 549–557.
Haug, H., 2001. Quantum kinetics for femtosecond spectroscopy in semiconductors. In: Willardson RK and Weber ER (eds) *Semiconductors and Semimetals*, vol. 67, pp. 205–229. London: Academic Press.
Haug, H., Jauho, A.P., 1996. Quantum Kinetics in Transport and Optics of Semiconductors. Berlin: Springer-Verlag.
Haug, H., Koch, S.W., 1993. Quantum Theory of the Optical and Electronic Properties of Semiconductors. Singapore: World Scientific.
Kash, J.A., Tsang, J.C., 1992. In: Shank, C.V., Zakharchenya, B.P. (Eds.), Spectroscopy of Nonequilibrium Electrons and Phonons. Amsterdam: Elsevier, pp. 113–168.
Mukamel, S., 1995. Principles of Nonlinear Optical Spectroscopy. New York: Oxford University Press.
Shah, J., 1992. Hot Carriers in Semiconductor Nanostructures: Physics and Applications. Boston: Academic Press.
Shah, J., 1999. Ultrafast Spectroscopy of Semiconductors and Semiconductor Nanostructures, 2nd edn. Heidelberg, Germany: Springer-Verlag.
Steel, D.G., Wang, H., Jiang, M., Ferrio, K.B., Cundiff, S.T., 1994. In: Phillips, R.T. (Ed.), Coherent Optical Interactions in Solids. New York: Plenum Press, pp. 157–180.
Wegener, M., Schaefer, W., 2002. Semiconductor Optics and Transport. Heidelberg, Germany: Springer-Verlag.
Zimmermann, R., 1988. Many-Particle Theory of Highly Excited Semiconductors. Leipzig, Germany: Teubner.

Band Structure and Optical Properties

W Zawadzki, Polish Academy of Sciences, Warsaw, Poland

Introduction

The purpose of this article is to describe the influence of the electronic energy band structure of semiconductors on their optical properties. By the band structure one understands the energy–wavevector dispersion relation $\varepsilon(\mathbf{k})$ for the electron energy bands and the relative positions of band maxima and minima in the Brillouin zone.

The main problem in calculating the band structure of a semiconductor is that of the periodic potential of the lattice in which the electrons move. It is, in fact, a self-consistent problem since the moving electrons partly screen the potential. Different approximations have been developed to deal with the question, and in all of them the symmetry of the lattice is of fundamental importance. Thus, in the so-called empty-lattice approximation one deals exclusively with the symmetry and periodicity of the lattice without specifying the potential. This can give qualitative ordering and symmetry of the bands but no quantitative results. In the opposite limit one uses the tight binding approximation, in which the bands are constructed from the atomic states of separate atoms. This method gives quite a good description of lower (valence) bands but poor results for higher (conduction) bands. A powerful way to describe real energy bands is obtained by various forms of pseudopotential calculation, where one approximates the potential near actual atoms by simple parametrized forms and then adjusts the parameters to fit experimental (mostly optical) data. The pseudopotential methods give a good overall description of various bands in the entire Brillouin zone, but they do not provide sufficiently precise results near the band extrema. A semi-empirical way to serve the latter purpose is called $\mathbf{k}\cdot\mathbf{p}$ theory which we describe below. Various methods of band structure calculation are reviewed in books on solid state physics as outlined in the suggestions for Further Reading at the end of this article.

In the majority of important semiconductor materials (Si, Ge, many III–V and II–VI compounds) the maxima of the valence bands are at the center of the Brillouin zone (Γ point). The minima of the conduction bands in Si and Ge are not at the Γ point. This means that the fundamental optical interband transitions in these materials are indirect (in the ε–$\mathbf{k}$ diagram), i.e., they require phonon assistance. On the other hand, the minima of conduction bands in important III–V and II–VI compounds are at the Γ point, so that the interband optical transitions are direct and they do not require phonon assistance. Both systems are utilized in opto-electronic devices, particularly for detectors. However, for emitters (light-emitting diodes and lasers) we are mostly concerned with the second case, on which we concentrate here.

Bloch States

The Bloch theorem states that if the potential $V(\mathbf{r})$ in which the electron moves is periodic with the periodicity of the lattice, then the solutions $\Psi(\mathbf{r})$ of the Schrödinger wave equation

$$\left[\frac{\mathbf{p}^2}{2m_0}+V(\mathbf{r})\right]\Psi(\mathbf{r})=\varepsilon\Psi(\mathbf{r}) \tag{1}$$

are of the form $\mathbf{\Psi}(\mathbf{r})=\exp(\mathrm{i}\mathbf{k}\cdot\mathbf{r})u_k(\mathbf{r})$, where $u_\mathrm{k}(\mathbf{r})$ is periodic with the periodicity of the direct lattice, and $\mathbf{k}$ is the wavevector ($\hbar\mathbf{k}$ is the pseudomomentum). The proof of this theorem can be found, for example, in undergraduate textbooks on solid state physics.

k·p Theory

The underlying idea of semi-empirical $\mathbf{k}\cdot\mathbf{p}$ theory is to describe the energy bands in the vicinity of a given point of the Brillouin zone (usually near a band extremum). Symmetry properties are used to minimize the number of unknown band parameters (effective mass, energy gap, etc.) which are then determined by experiment. The initial Schrödinger equation for an electron in the periodic potential $V(\mathbf{r})$ of the crystal lattice reads

$$[\mathbf{p}^2/2m_0+V(r)+H_\mathrm{so}]\Psi=\varepsilon\Psi \tag{2}$$

where m_0 is the free electron mass and H_so is the spin–orbit interaction. Since $H_\mathrm{so}(\mathbf{r})$ is also periodic with the lattice periodicity, the solutions of Eq. (2) are given by the Bloch states above.

We use $\mathbf{k}\cdot\mathbf{p}$ perturbation theory, as follows, to describe the energy bands near $\mathbf{k}=0$ (i.e., the so-called Γ point of the Brillouin zone). Since we are interested in small values of $\mathbf{k}$, we expand our unknown cell periodic functions, $u_{l\mathbf{k}}(\mathbf{r})$, in terms of the set of Γ-point functions, $u_{lo}(\mathbf{r})$, whose symmetry we know. One defines a representation $\Phi_{l\mathbf{k}}=\exp(\mathrm{i}\mathbf{k}\cdot\mathbf{r})u_{lo}(\boldsymbol{r})$, where $u_{lo}(\mathbf{r})$ is the periodic

k-independent function satisfying the equation

$$[\mathbf{p}^2/2m_0 + V + \mathrm{H_{so}}]u_{lo} = \varepsilon_{lo}u_{lo} \tag{3}$$

whose solution we know in terms of the band-edge energies, ε_{lo}. This representation is sometimes referred to as the Luttinger and Kohn (LK) representation. The u_{lo} functions are in general mixtures of spin up and spin down states because of the spin–orbit interaction. One looks for the solutions of **Eq. (2)** in the form

$$\Phi_{n\mathbf{k}} = \exp(\mathrm{i}\mathbf{k}\cdot\mathbf{r})\sum_l c_l^{(n)}(\mathbf{k})u_{lo}(\mathbf{r}) \tag{4}$$

where the sum is over all bands and $c_l^{(n)}(\mathbf{k})$ are the coefficients to be determined. Inserting the above form into **Eq. (2)**, performing the operation $\mathbf{p}^2$ (i.e., operating twice on the Bloch product function with the operator $\mathrm{p} = -\mathrm{i}\hbar\nabla$) and using the property **(3)** one eliminates the unknown potential $V(\mathbf{r})$. Multiplying the resulting equation from the left by $u^*_{l'o}$, integrating over the unit cell Ω and using the orthonormality of u_{lo}, one finally obtains

$$\sum_l \left[(\varepsilon_{lo} - \varepsilon')\delta_{l'l} + \frac{\hbar}{m_0}\mathbf{k}\cdot\mathbf{p}_{l'l}\right]c_l^{(n)}(\mathbf{k}) = 0 \tag{5}$$

for $l' = 1, 2, 3,\ldots$ Here $\delta_{l'l}$ is the Kronecker delta function, $\varepsilon' = \varepsilon - \hbar^2k^2/2m_0$, and $\mathbf{p}_{l'l}$ is the so-called matrix element of the momentum given by $\mathbf{p}_{l'l} = (1/\Omega)\int_\Omega u^*_{l'o}\mathbf{p}u_{lo}\,\mathrm{d}\mathbf{r}$. The index l' runs over all energy bands.

Eq. (5) formulates the famous $\mathbf{k}\cdot\mathbf{p}$ theory. The nondiagonal part $\mathbf{k}\cdot\mathbf{p}_{l'l}$ can be eliminated by the perturbation procedure, and the method is referred to as $\mathbf{k}\cdot\mathbf{p}$ perturbation theory. In the second order of perturbation the bands are separated and in each band the carriers are described by a dispersion relation $\varepsilon_l = \left(\hbar^2/2\right)\mathbf{k}\left(\widehat{1/m^*_{lo}}\right)\mathbf{k}$ where $\left(\widehat{1/m^*_{lo}}\right)$ is an inverse effective mass tensor at the band edge in question. The inverse mass is generally a 3×3 tensor, but for cubic materials in the center of the Brillouin zone it is a scalar, so that $\varepsilon_l(\mathrm{k}) = \hbar^2\mathrm{k}^2/2\mathrm{m}^*_{lo}$ (where $1/\mathrm{m}^*_{lo} \equiv 1/\mathrm{m}_0 + (2/\mathrm{m}_0^2)\sum_{l\neq l'}(p^2_{ll'})/[\varepsilon_{l0} - \varepsilon_{l'0}]$). We then say that the band is spherical and parabolic. The second order of perturbation is a good approximation if the energy ε (as counted from a nondegenerate band edge) is small compared to forbidden energy gaps. For the triply degenerate p-like valence band one has to do degenerate perturbation theory and treat the bands together as a 3×3 matrix equation.

In the same approximation the wavefunction for a given band is: $\Psi_{l\mathbf{k}}(\mathbf{r}) = \exp(\mathrm{i}\mathbf{k}\cdot\mathbf{r})\ u_{lo}\ (\mathbf{r})$. In the absence of spin–orbit coupling, H_{so}, the u_{lo} function would just have the symmetry of the parent band (i.e., labeled S for the s-like conduction band, and X, Y or Z for the triply degenerate p-like valence band – each with a single spin up or spin down component). In the presence of H_{so}, the triply degenerate valence band states become mixtures of X, Y and Z with mixed spin character and part of the degeneracy is raised. In atomic notation the presence of H_{so} has transformed the basis from the (l, s) to the (J, m_J) representation. This results in the well-known light and heavy hole bands (degenerate at $k=0$) and the spin–orbit split-off band.

Under the influence of a radiation field, of vector potential $\mathbf{A}'$ and frequency ω, the optical transition probability for an electron to be raised from state Ψ_i (initial) to Ψ_f (final) is proportional to the square of $\mathrm{M} = (e/mc)\ \langle\Psi_f|\mathbf{A}'\cdot\mathbf{p}|\Psi_i\rangle$ i.e., it is determined by the same momentum matrix element, p_{fi}, which governs the effective mass. For interband transitions one immediately gets the selection rule $\Delta k = 0$ (i.e., direct transitions). The polarization selection rules are determined by the $u_{lo}(r)$ components of the initial and final states, through the momentum matrix elements.

Narrow-Gap Semiconductors

Semiconductors such as InSb and $\mathrm{Hg_{1-x}Cd_xTe}$ (with $x<0.3$) have small energy gaps and are referred to as narrow gap semiconductors (NGSs). They are commonly used in opto-electronic applications which may be accessing energies in the conduction or valence band which are a significant fraction of the energy gap. Under these circumstances it is not valid to cut off the expansion to order k^2, and one has to deal with so-called nonparabolic bands. Thus, in NGSs (or indeed in any situation where the energy becomes an appreciable fraction of ε_{g}) the above procedure is not valid and an alternative approximation is in order. Following semidegenerate perturbation theory one includes exactly in **Eq. (5)** a finite number of bands (near each other in energy) and neglects distant bands. This is referred to as the Kane model, and the energy bands near $k=0$ are shown for InSb-type III–V compounds and for HgTe-type II–VI compounds in **Fig. 1**. In this case we include only the conduction and valence band in our set of states (**Eq. (5)**). Including spin and degeneracy of the Γ_8 symmetry the set **(5)** then has eight equations. They can be solved analytically and the resulting energies are given by

$$\varepsilon'(\varepsilon' + \varepsilon'_{\mathrm{g}})(\varepsilon' + \varepsilon'_{\mathrm{g}} + \Delta) - \left(\varepsilon' + \varepsilon'_{\mathrm{g}} + \frac{2\Delta}{3}\right)\kappa^2k^2 = 0 \tag{6}$$

where $\varepsilon' = \varepsilon - \hbar^2k^2/2m_0$. We see that for this restricted set of states the only unknown parameters are the energy gap, ε_g, the spin–orbit splitting energy, Δ, and $\kappa = -\mathrm{i}(\hbar/m_0)\langle S|p_z|Z\rangle$ (note that this is the only non-vanishing matrix element of the momentum). The fourth energy is $\varepsilon' = -\varepsilon_g$. All four energies are spin degenerate. In NGSs one may neglect the free electron term, i.e., set $\varepsilon' \approx \varepsilon$. For $\varepsilon \ll \varepsilon_{\mathrm{g}} + (2/3)\Delta$ the resulting quadratic equation for the conduction and the light-hole bands is

$$\varepsilon(1 + \varepsilon/\varepsilon_{\mathrm{g}}) = \hbar^2k^2/2m^*_0 \tag{7}$$

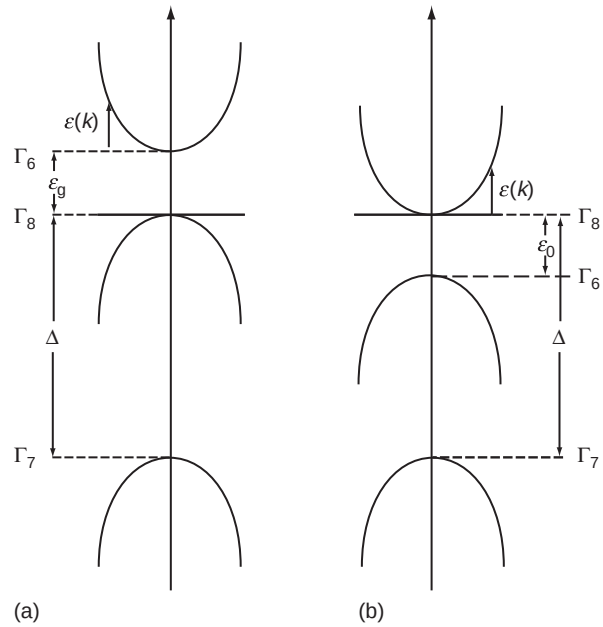

Fig. 1 Three-level model of band structure near the Γ point of the Brillouin zone: (a) InSb-type semiconductors; (b) HgTe-type semiconductors.

where $1/m_0^* = (4\kappa^2/3\hbar^2\varepsilon_g)(\varepsilon_g + 2\Delta/3)/(\varepsilon_g + \Delta)$ defines the effective mass at the band edge. The root for the heavy holes is $\varepsilon = -\varepsilon_g$ i.e., this band is not correctly described within the three-level model. The bands given by **Eq. (6)** are spherical but nonparabolic. for $\varepsilon \ll \varepsilon_g$ one recovers the standard dispersion, $\varepsilon = \hbar^2 k^2/2m_0^*$.

For HgTe-type materials, setting the zero of energy at the Γ_8 edge and replacing ε_g by $-\varepsilon_0$ (cf. **Fig. 1**), one obtains

$$\varepsilon'(\varepsilon' + \varepsilon_0)(\varepsilon' + \Delta) - \left(\varepsilon' + \frac{2\Delta}{3}\right)\kappa^2 k^2 = 0 \tag{8}$$

For $\varepsilon' \ll (2/3)\Delta$ the dispersion **(7)** is valid with $1/m_0^* = 4\kappa^2/3\hbar^2\varepsilon_0$. In this case the conduction and the heavy hole bands ($\varepsilon \equiv 0$) are degenerate at $k = 0$, i.e., the thermal gap is zero.

The important property of the above model is that it holds also in the limit of $\varepsilon_g \to 0$ (i.e., for mixed $Hg_{1-x}Cd_xTe$ crystals). The effective mass in **Eq. (7)** tends to zero, but the dispersion described by **Eq. (6)** is valid and it gives $\varepsilon = (2/3)\kappa k$, i.e., a linear band.

The electron and hole wavefunctions in NGSs resulting from semidegenerate perturbation theory are given by **Eq. (4)**, in which the sum runs over all the included states. Thus they involve truly **k**-dependent amplitudes of the Bloch states, cf. **Eq. (1)**. In addition, these functions are mixtures of spin-up and spin-down states. These features have important consequences for optical and dc transport phenomena.

If the conduction band minimum is not at the Γ point (Ge and Si), there are at least two different matrix elements of momentum and the resulting band is ellipsoidal.

Effective Mass

The nonparabolic dispersions $\varepsilon(k)$ in NGSs require generalizations of important relations. The momentum mass $\widehat{m}^*$ is introduced as a connection between the psuedomomentum and the velocity: $\hbar\mathbf{k} = \widehat{m}^*\mathbf{v}$. Since $v_i = \delta\varepsilon/\delta(\hbar k_i)$, one obtains for a spherical band

$$\frac{1}{m*} = \frac{1}{\hbar^2 k}\frac{\mathrm{d}\varepsilon}{\mathrm{d}k} \tag{9}$$

The above mass is a scalar, but it depends on the energy (or momentum). This mass is more useful than the one usually defined in textbooks, relating force to acceleration. The latter is not a scalar even for a spherical band. However, for the standard parabolic band, $\varepsilon = \hbar^2 k^2/2m_0^*$, both masses are the same and are equal to m_0^*. The mass **(9)** is related via velocity to the electric current, so that it enters naturally in transport phenomena (also free carrier optics). In particular, this mass enters the definition of mobility $\mu = q\tau/m^*$, where q is the charge and τ is the relaxation time. It enters the density of states (see below). Finally, it is the momentum mass **(9)** which is measured in cyclotron resonance.

For the dispersion **(7)** (the so-called two-level model), the mass **(9)** in the conduction band is

$$m^*(\varepsilon) = m_0^*\left(1 + \frac{2\varepsilon}{\varepsilon_g}\right) \tag{10}$$

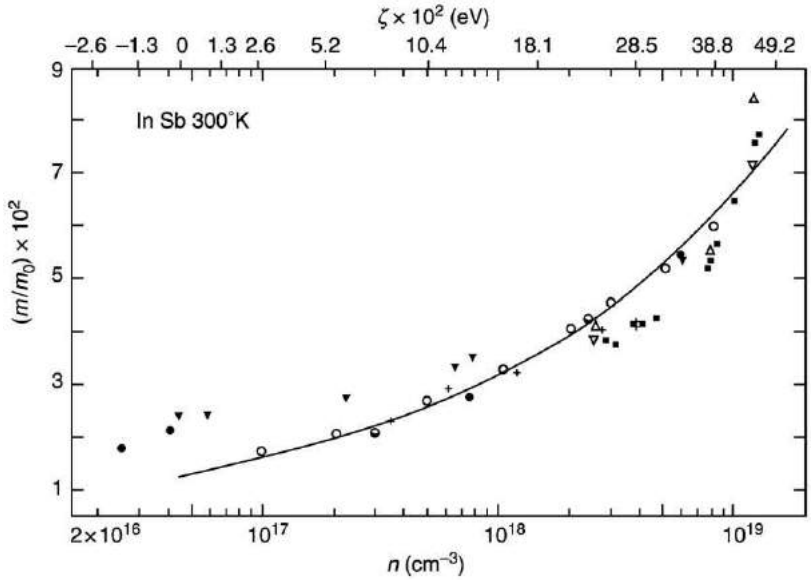

Fig. 2 Electron effective mass in InSb at room temperature versus free electron concentration. The solid line, calculated for the two-level model, represents the mass value at the Fermi energy, as indicated on the upper abscissa. The experimental results of various authors are also indicated. Reproduced from Zawadzki W (1974) Electron transport phenomena in Sm all-gap semiconductors. *Advances in Physics* 23: 435–512.

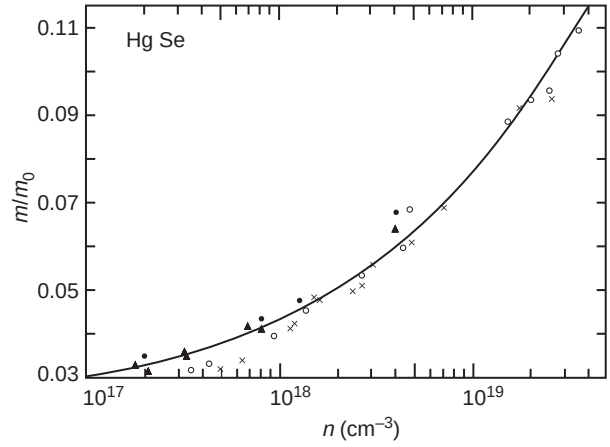

Fig. 3 Electron effective mass in zero-gap material HgSe versus free electron concentration. The solid line is calculated for the three-level model. Experimental results of various authors are also indicated. Reproduced from Zawadzki W (1974) Electron transport phenomena in Sm all-gap semiconductors. *Advances in Physics* 23: 435–512.

Again, for $\varepsilon/\varepsilon_g \ll 1$ the energy dependence of the mass is negligible, which represents the standard regime. **Figs. 2** and **3** illustrate energy dependences of the electron masses in InSb and HgSe, as measured by various optical and transport methods. The increase of the mass with energy is very strong in both cases.

Density of States

The density of states in energy space is calculated beginning with pseudomomentum space. For the spherical band one obtains

$$\rho(\varepsilon) = \frac{k^2}{\pi^2}\frac{\mathrm{d}k}{\mathrm{d}\varepsilon} = \frac{m^* k}{\pi^2 \hbar^2} \tag{11}$$

The momentum mass **(9)** enters naturally into this important quantity. For the parabolic case $m^* = \text{const.}$ and $k \sim \varepsilon^{1/2}$, so that the standard result is recovered. In the general case one should use **Eq. (10)** for $m^*(\varepsilon)$ and determine k from **Eqs. (6)** or **(7)**.

Electron–photon Interaction

The electron–photon interaction can be introduced into the theory on two levels. If one replaces in the initial Eq. (2) the momentum $\mathbf{p}$ by $\mathbf{p}+(e/c)\mathbf{A}'$, the interaction term is obtained in the scalar form $H_{\text{int}}=(e/m_0c)\,\mathbf{A}'\cdot\mathbf{p}$, where $\mathbf{A}'$ is the vector potential of radiation. The wavefunctions to be used for the calculation of matrix elements are of the form (4), i.e., they include the periodic LK amplitudes. In fact, the matrix elements of momentum for optical transitions are identical to those of the $\mathbf{k}\cdot\mathbf{p}$ theory, as pointed out earlier. Since (as noted above) the wavefunctions for all bands have the same form (4), differing only by the coefficients $c_l^{(n)}(\boldsymbol{k})$, the matrix elements for interband and intraband optical transitions have the same form.

The electron–photon interaction can also be introduced directly to the $\mathbf{k}\cdot\mathbf{p}$ theory of Eq. (5) replacing $\hbar\mathbf{k}$ by $\hbar\mathbf{k}+(e/c)\mathbf{A}'$. If the free electron term $\hbar^2\,k^2/2m_0$is neglected, the interaction matrix involving $\mathbf{A}'$ terms is a number matrix. In this procedure the wavefunctions for initial and final states are simply columns and rows of $c_l^{(n)}$ coefficients and LK amplitudes no longer come into play. Here the $\mathbf{p}_{l'l}$ elements of $\mathbf{k}\cdot\mathbf{p}$ theory determine directly the electron–photon interaction. Both procedures described above give the same results.

Quantum Wells

If charge carriers are placed in a quantum well described by a potential $U(z)$, the motion in the z-direction is quantized while the motion in the x–y plane remains free. One takes the same LK basis (cf. Eq. (3)), but the general form of solution is given by

$$\Psi=\sum_l f_l(\mathbf{r})u_{l0}(\mathbf{r}) \tag{12}$$

since, in the presence of an external potential, $\mathbf{k}$ is not a good quantum number and the envelope functions $f_l(\mathbf{r})$ are unknown. If $U(z)$ is a slowly varying function within the unit cell the potential appears only on the diagonal of the set (5) and $\hbar\mathbf{k}$ is replaced by $\mathbf{p}$. One obtains

$$\sum_l\left[\left(\varepsilon_{l0}+\frac{\mathbf{p}^2}{2m_0}+U-\varepsilon\right)\delta_{l'l}+\frac{\mathbf{p}_{l'l}\cdot\mathbf{p}}{m_0}\right]f_l(\mathrm{r})=0 \tag{13}$$

One can now eliminate the nondiagonal terms applying the standard Luttinger and Kohn scheme of second-order perturbation theory and arrive at the decoupled band equations with effective masses. This results in the standard quantization due to electric (and magnetic, if present) fields, and wavefunctions are of the form $\Psi_{lm}=f_m u_{l0}$ for the m-th sub-band of the l-th band.

For NGSs this procedure is a poor approximation and one should proceed as before including exactly a finite number of bands (cf. Eq. (12)). This results in a set of coupled differential equations for the envelope functions $f_l(\mathrm{r})$. In some important cases one can find the eigenenergies without finding the functions by using the semiclassical approximation (the so-called WKB method).

We shall omit the calculations with coupled differential equations by using a semiclassical procedure also in another sense. Namely, we shall generalize the nonparabolic dispersion relation (7) to include the presence of external fields. To this end we observe that, including the potential $U(\mathbf{r})$ on the diagonal of Eq. (13), one replaces $-\varepsilon$ by $-\varepsilon+U$. Further, if a magnetic field is introduced to the $\mathbf{k}\cdot\mathbf{p}$ theory one replaces $\hbar\mathbf{k}$ by $\mathbf{P}=\mathbf{p}+(e/c)\mathbf{A}$, where $\mathbf{A}$ is the vector potential of the magnetic field. Thus the semiclassical equation resulting from Eq. (7) is

$$\left[\frac{\mathbf{P}^2}{2m_0^*}-(\varepsilon-U)\left(1+\frac{\varepsilon-U}{\varepsilon_{\mathrm{g}}}\right)\right]\Psi=0 \tag{14}$$

It can be seen that for $\varepsilon-U\ll\varepsilon_{\mathrm{g}}$ one recovers the standard one-band approximation mentioned above. However, below we consider the general situation described by Eq. (14).

Let us first consider the case of no magnetic field, i.e., $\mathbf{P}=\mathbf{p}$. For $U(\mathbf{r})=U(z)$ one can separate the variables by looking for solutions in the form $\Psi=\exp(\mathrm{i}k_x x+\mathrm{i}k_y y)\Phi(z)$. One can now transform Eq. (14) into

$$\left[\frac{1}{2m_0^*}p_z^2-\frac{1}{\varepsilon_{\mathrm{g}}}(\varepsilon-\varepsilon_\perp-U)(\varepsilon_{\mathrm{g}}+\varepsilon+\varepsilon_\perp-U)\right]\Phi(z)=0 \tag{15}$$

where

$$\varepsilon_\perp=\frac{\varepsilon_{\mathrm{g}}}{2}+\left[\left(\frac{\varepsilon_{\mathrm{g}}}{2}\right)^2+\varepsilon_{\mathrm{g}}\frac{\hbar^2k_\perp^2}{2m_0^*}\right]^{1/2} \tag{16}$$

in which $k_\perp^2=k_x^2+k_y^2$. Eq. (15) is suitable for the WKB semiclassical quantization with which one determines the eigenenergies ε by one integration.

For a magnetic field applied along the z-direction (parallel electric and magnetic fields) the effects of both fields are separated and the form of Eq. (15) is valid, with $\hbar^2k_\perp^2/2m_0^*$ in Eq. (18) replaced by $\hbar\omega_{\mathrm{c}}(n+1/2)\pm(1/2)|g_0^*|\mu_{\mathrm{B}}B$. Here $\omega_{\mathrm{c}}=eB/m_0^*c$ is the cyclotron frequency, $n=0,\,1,\,2,\ldots$ is the Landau quantum number, g_0^* is the spin splitting factor, and μ_{B} is the Bohr magneton. This results in nonequal spacing of Landau levels for a fixed value of B and in nonlinear B dependence of the Landau energies as functions of B (similarly but not identically to the bulk case). These features are illustrated by the cyclotron resonance experiment and theory shown in Fig. 4.

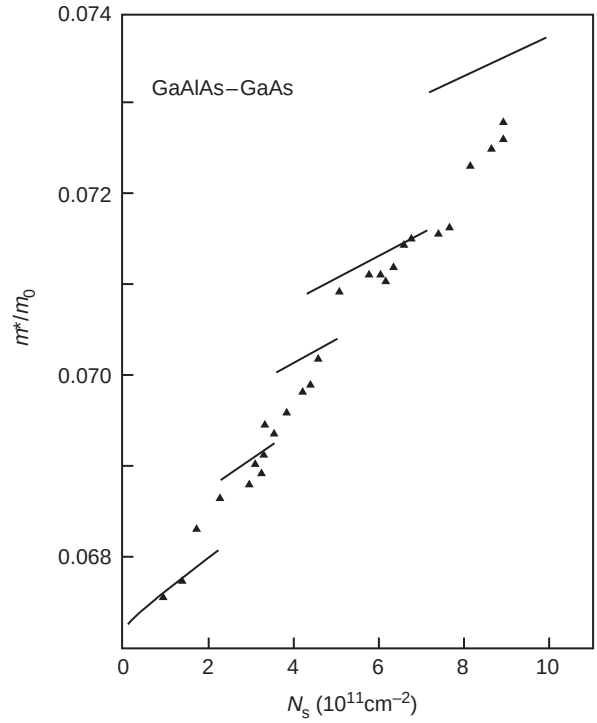

Fig. 4 Electron effective mass versus two-dimensional electron density in GaAs/GaAlAs heterostructures, as measured by far infrared cyclotron resonance. The solid lines are calculated using the effective two-level model for GaAs. Reproduced from Zawadzki W, *et al.* (1994) Cyclotron emission study of electron masses in GaAs-GaAlAs heterostructures. *Semiconductor Science and Technology* 9: 320–328.

Selection Rules for Intersub-band Optical Transitions

For the one-band effective mass approximation the electron–photon interaction is $H_{\text{int}} \sim \mathbf{A}' \cdot \mathbf{p}$ (cf. above), and the wavefunctions for the initial and final sub-bands are $\Psi_m = \exp(\mathrm{i}k_x x + \mathrm{i}k_y y)\Phi_m(z)$, in which the $\Phi_m(z)$ must be orthogonal to each other since they describe different energies. It is then clear that the matrix element $\langle fn|\mathbf{A}' \cdot \mathrm{p}|in\rangle$ is finite only if $\mathbf{A}'$ is polarized along the z-direction (also called the growth direction) since for $\mathbf{A}'$ parallel to x or y the integral over z is $\langle \Phi_{m'}|\Phi_m\rangle = 0$ giving a vanishing matrix element. Experimentally this is a serious problem for spectroscopic applications involving intraband (i.e., intersub-band) transitions in n-type semiconductors, since it is not possible to access these using normal incidence radiation.

This selection rule is somewhat relaxed for p-type semiconductors as a result of the mixed spin states of the complex valence band. In addition the narrow-gap band structure introduces interesting new possibilities into the selection rules. The wavefunctions have the form **(12)** and for the light polarizations A'_x or A'_y the momentum operators p_x or p_y act also on the periodic amplitudes $u_{l0}(\mathbf{r})$, which leads to a nonvanishing transition probability for intersub-band transitions. This has indeed been observed in narrow-gap materials.

Further Reading

Ashcroft, N.W., Mermin, N.D., 1976. Solid State Physics. Holt, Rinehart and Winston.
Kane, E.O., 1980. Band structure of narrow gap semiconductors. In: Zawadzki, W. (Ed.), Narrow Gap Semiconductors. Physics and Applications. Berlin: Springer, pp. 13–31.
Luttinger, J.M., Kohn, W., 1955. Motion of electrons and holes in perturbed periodic fields. Physical Review 97, 869.
Zawadzki, W., 1974. Electron transport phenomena in Sm all-gap semiconductors. Advances in Physics 23, 435.
Zawadzki, W., 1984. Electric and magnetic quantisation of two-dimensional systems – elementary theory. In: Bauer, G., *et al.* (Eds.), Two-Dimensional Systems, Heterostructures, and Superlattices. Berlin: Springer, p. 2.
Zawadzki, W., 1991. Intraband and interband magneto-optical transitions in semiconductors. In: Landwehr, G., Rashba, E.I. (Eds.), Landau Level Spectroscopy. Amsterdam: North-Holland.

Excitons

I Galbraith, Heriot-Watt University, Edinburgh, UK

In semiconductors electron–hole pairs can lower their energy by correlating their motion in an exciton state. Such states dominate the bandedge optical properties of direct-gap semiconductors. The resultant absorption peaks are affected by quantum confinement, electric, and magnetic fields and carrier scattering.

Introduction

In the simplest, single-particle picture, the ground state of a semiconductor consists of a full valence band and an empty conduction band. The lowest lying excitation above the ground state in such a scheme is produced by excitation of one electron from the valence band to the conduction band. To achieve this optically requires a photon energy greater than the bandgap, E_g. However in many semiconductor materials this simple picture is unable to explain the observed absorption spectrum in the neighborhood of the fundamental bandgap. The origin of the discrepancy lies in the neglect of the Coulomb interaction between the negatively charged excited electron and the hole which is left in the valence band.

The term exciton refers to a two-particle excitation, consisting of a bound electron and hole. Such excitations dominate the bandedge optical spectra of most direct gap semiconductors, especially at low temperatures. In particular there exists a series of hydrogen-like bound states lying below the bandedge. These states are bound by energies

$$E_b^{3D} = \frac{\mu e^4}{8\mathrm{h}^2 \varepsilon_r^2 \varepsilon_0^2} \frac{1}{n^2} \tag{1}$$

where $\mu = m_e m_h/(m_e + m_h)$, is the reduced mass, $m_{e(h)}$ the electron (hole) effective mass, e is the electronic charge, h is Planck's constant, ε_0 is the permittivity of free space, ε_r the relative permittivity of the material and $n = 1, 2, 3, \ldots\infty$ is the principal quantum number.

The binding energy of the lowest lying exciton state varies considerably from one semiconductor to another being, e.g., 0.5 meV in InSb and over 60 meV in Zn0. Similarly the spatial extent of the 1 s electron-hole relative wave function or effective Bohr radius is given by

$$a_0^{3D} = \frac{\hbar}{\sqrt{2\mu E_b^{3D}}} \tag{2}$$

and ranges from 750 μm in InSb to ≈2 nm in Zn0. Typically, in semiconductors the exciton spatial extent is much larger than the lattice constant of around 0.6 nm, and such excitons are classed Wannier excitons. At the opposite limit, which occurs in many molecular materials, the exciton is localized around a particular atomic site and is termed a Frenkel exciton. Wannier excitons are characterized by the hydrogen-like quantum numbers of their relative motion and an overall center of mass momentum which describes the wave-like motion of the bound entity through the crystal.

Another type of exciton is often referred to as a bound exciton. This consists of an electron–hole pair bound to a neutral impurity center. Such excitons are localized around the impurity and cannot move through the crystal in the same way as a free exciton.

Optical Properties

Excitons manifest themselves primarily as strong modifications to the bandedge optical properties of semiconductors. In particular the bound states give rise to discrete absorption lines at lower energy than the bandgap energy. Several peaks corresponding to the different hydrogen-like bound states (1s, 2s, etc.) can be seen provided the lines are sufficiently narrow. A theoretical calculation of this is shown in **Fig. 1**.

Associated with these absorption changes are concomitant changes in the refractive index. Contributions to the spectral width of the absorption peaks are characterized as either homogeneous or inhomogeneous. Homogeneous broadening refers to broadening associated with the finite lifetime of a particular state, e.g., due to phonon scattering. Inhomogeneous broadening is due to non-uniformity in the system, e.g., in a sample having a spatially non-uniform strain field.

There is also enhancement of the absorption at photon energies above the bandgap due to the residual influence of the Coulomb attraction on the electron-hole scattering states. Although not bound, electrons and holes in these states still have an enhanced probability of being found close together. This is called the Sommerfeld Enhancement Factor (see **Fig. 1**).

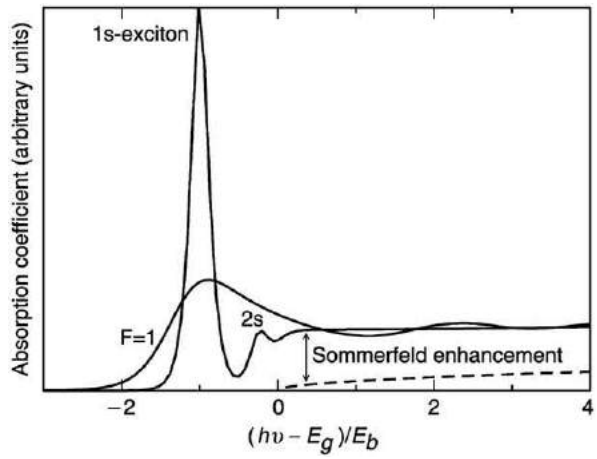

Fig. 1 Calculated bulk absorption spectra including (solid-line) and neglecting (dashed line) the Coulomb interaction. The Sommerfeld enhancement is seen as photon energies above the bandgap. Also shown is the spectrum with an applied electric field of one exciton Rydberg per Bohr radius.

It is important to note that although there may be an excitonic peak in the absorption spectrum this does not imply that excitons form a stable population in such a sample. Often, especially at room temperature, the lifetime of an optically generated exciton is determined by the scattering time with optical phonons which may be very short (<1 ps). In this case the excitons are quickly ionized and the quasi-thermal equilibrium which is formed is dominated by essentially plasma-like behavior. In photoluminescence experiments, where carriers are generated high in the band and the spectrum of the resulting emitted luminescence is measured, peaks associated with the free exciton are often observed and taken as a signature of the presence of excitons. This view has been challenged recently by theoretical calculations which show that a photoluminescence peak at the exciton energy can be also be produced from an uncorrelated plasma if proper account of the Coulomb interaction is taken. This issue remains controversial.

In bulk samples, experiments such as absorption and reflectivity are complicated by the existence of polariton effects. An exciton-polariton is a quantum mixture of the propagating photon inside the semiconductor sample and the material excitation. Where the photon dispersion and the exciton energy dispersion meet there is an anti-crossing and this is manifested in, for example, the appearance of two lines in the reflectivity spectrum.

It was hoped during the 1970s that laser action in wide bandgap material such as ZnSe would be possible based on excitonic lasing. In this process a quasi-equilibrium exciton population would form the injected excitation and stimulated recombination would occur with the associated emission of a scattering partner which would take up the necessary momentum to ensure overall momentum conservation. This leads to a light emission wavelength below the absorption edge which is advantageous for minimizing losses. In fact, this process has only been seen at low temperature for scattering with LO-phonons, electrons and other excitons. Each mechanism has its own characteristic lasing threshold and temperature dependence.

Influence of Quantum Confinement

The advent of Molecular Beam Epitaxy enabled the growth of semiconductor layers of thickness similar to the electronic de Broglie wavelength. By sandwiching a low bandgap semiconductor between two high bandgap layers, one can make quantum well structures in which the electronic motion is effectively confined to the plane of the layer. In such structures the excitonic properties are also radically altered by this confinement. An example of the absorption spectrum of a ZnCdSe quantum well embedded in ZnSe is shown in Fig. 2. The heavy and light hole excitons correspond to different spin quantum numbers for the electron and hole making up the exciton.

The exciton binding energy is enhanced by a factor of 4 compared to the bulk case and the spatial extent of the 1s wavefunction is also reduced by half:

$$E_b^{2D} = \frac{\mu e^4}{8\mathrm{h}^2\varepsilon_r^2\varepsilon_0^2}\frac{1}{(n-\frac{1}{2})^2} \quad n = 1, 2, 3\ldots \tag{3}$$

This enhancement can be understood since the particles are confined to lie in the plane and are hence closer together on average than in the three-dimensional case. In practice, a four-fold enhancement in the binding energy is never observed since the quantum wells are neither infinitely deep nor infinitely thin. Both these limitations lead to a reduction in the actual enhancement from 4 to about 2, depending on the structure. For example, the binding energy in a realistic quantum well corresponds to the bulk value of the well material for very wide wells and to the bulk value of the barrier material for very narrow wells. In between the

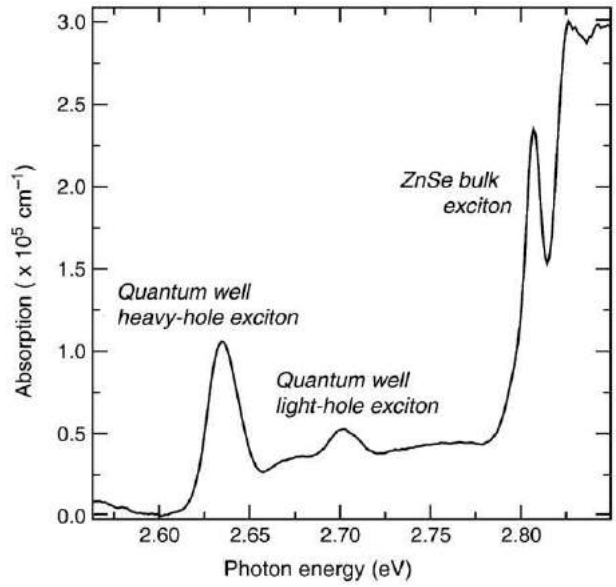

Fig. 2 Heavy and light hole exciton absorption spectrum for ZnCdSe quantum wells embedded between ZnSe barriers. Also shown is the bulk exciton of the ZnSe barriers.

value is enhanced by the quantum confinement. The maximum enhancement typically occurs when the well width is around half the bulk Bohr radius for the well material.

Accurate calculation of the exciton binding energy is complicated by the fact that the quantum confinement influences the band structure, making it strongly anisotropic and splitting heavy and light holes. A complete inclusion of these effects is at the limit of current computational power due to the need to calculate the electron-hole interaction matrix elements. Approximate schemes, including variational approaches, have proved successful in fitting a variety of samples, including GaAs and ZnSe based quantum wells. The limitation to such calculations is often the requirement for accurate values of the material parameters such as effective masses and energy band offsets.

The Sommerfeld enhancement of the absorption into the continuum states of quantum wells leads to a doubling of the absorption at the bandedge. This enhancement reduces only slowly as the photon energy is increased.

If multiple quantum wells with very thin barrier layers are used, the electronic states in the neighboring wells couple together to form a superlattice. The superlattice dispersion in the growth direction behaves like a heavy mass and the resulting exciton is three-dimensional but strongly anisotropic.

More extreme quantum confinement is possible making quantum wires, by etching or selective growth of quantum well samples. In these quasi-one dimensional structures, as in quantum wells, the exciton binding energy is again increased further. The Coulomb problem in one-dimension is pathological in that the ground-state binding energy diverges logarithmically, and higher states are pair-wise degenerate. However, accounting for the finite cross-section of realistic wires eliminates this problem and recovers a finite binding energy. One consequence of this is that there is no simple enhancement factor which corresponds to the four-fold enhancement in going from bulk to two-dimensional excitons. The density of states in one-dimensional systems diverges at the bandedge, implying the existence of an infinite absorption in a perfect quantum wire. However, the influence of excitonic correlations cancels this divergence, resulting in an enhanced, but not divergent, bandedge absorption feature.

The ultimate limit in quantum confinement can be achieved in quantum dots, where the carriers are confined in all three directions. In this case the term exciton should be used with some care as there are no bandedge states to compare with. The discrete energy levels dictated by the confinement are affected by the Coulomb interaction and the energy is renormalized – in some cases substantially. The absorption spectrum should consist of a series of sharp atomic-like transitions corresponding to s-s, p-p, d-d,… transitions and indeed luminescence spectra obtained from single quantum dots do display such fine structure. Even in the best controlled material systems, such as InAs quantum dots, the spectra are strongly inhomogeneously broadened due to variations in the dot size, shape, and alloy composition.

In large quantum wells, where the width is larger than the exciton Bohr radius the center-of-mass part of the wavefunction can be confined by the barriers. This leads to a quantized motion for the exciton as a whole, rather than for the individual electron and hole separately. One implication of this is that for such excitons the center-of-mass momentum is no longer a good quantum number.

The above effects are a controlled case of disorder induced effects in semiconductors. This is an area which is only now beginning to be understood. The presence of local potential fluctuations will influence the energy spectrum of excitons in that vicinity. Clearly both the depth and spatial extent of these fluctuations will be important – very short range (less than Bohr radius),

shallow potential fluctuations will not have any influence on the exciton. While deep potentials can lead to the center of mass confinement above.

Influence of Electric and Magnetic Fields

Just as an externally applied electric field affects the absorption edge via the Franz–Keldysh Effect, broadening the edge itself and inducing oscillations in the continuum part of the spectrum, so, in the case of an excitonic bandedge, there is substantial modification due to the presence of the field. At low fields, when the Coulomb interaction is strong compared to the potential drop across the exciton diameter, the exciton will be little affected by the applied field. At higher fields the exciton will become gradually more polarized by the field until the exciton is field ionized. This ionization process is manifested in the field broadening of the exciton line (see Fig. 1).

Of more technological importance is the influence of an electric field applied perpendicular to the plane of a quantum well. The field forces the electron and hole in opposite directions leading to a reduction in the exciton binding energy. This would lead to a blue shift in the exciton absorption peak were it not for an even larger reduction in the single particle electron and hole energy levels within the quantum wells. Overall a strong red-shift of the excitonic absorption peak is observed with increasing field. This is accompanied by a reduction in the oscillator strength due to the reduced electron-hole overlap. This is termed the Quantum Confined Stark Effect and is the basis for the operation of some semiconductor electro-optic modulators. An electric field applied in the plane of the quantum well gives rise to Franz–Keldysh oscillations analogous to the bulk case.

The influence of magnetic fields on excitons is more subtle and richer than the electric field. This is because the magnetic field seeks to induce a circular cyclotron orbit motion on the carriers, but this is influenced by their mutual Coulomb interaction. Two regimes exist where either the cyclotron energy:

$$\hbar\omega_c = \hbar(eB/\mu) \tag{4}$$

or the exciton binding energy dominates. In the former strong field case, one can treat the Coulomb interaction as a small perturbation on the electron and hole states in the presence of the magnetic field. In exciton states, which have a finite magnetic moment at zero field, we find a linear Zeeman shift in the magneto-exciton energy. In the other case we must use the magnetic field as the perturbation which produces a mixing of the zero-field exciton states. For example, for the ground state the magnetic field induces an admixture of the p-like excited state with the s-like ground state. The total angular momentum of the mixed state is proportional to B and as the energy of a magnetic dipole in a magnetic field is also proportional to B the shift of the magneto-exciton is proportional to B^2. This distinction between linear and quadratic shifts has been used to identify the nature of carrier populations (excitons or unbound free carriers) in a variety of samples. Clearly there exists an intermediate regime where both the magnetic and Coulomb energies are comparable and in this case the precise energy shifts need to be evaluated numerically.

In quantum well samples the orientation of the magnetic field with respect to the confinement direction is crucial in determining the physics. When the field is perpendicular to the confinement plane the cyclotron orbits exist as before and essentially the same phenomenon as in the bulk emerge. For other orientations, the behavior is much more complex and beyond the scope of this article.

Exciton Scattering

An electron and hole in an excitonic bound state execute a correlated motion which can be disturbed by scattering of either partner. This may lead to the ionization of the exciton and the destruction of the correlation or, alternatively, it may change the center of mass momentum of the exciton as a whole. Almost all of the important scattering mechanisms for excitons arise from the charged nature of the electron and hole. Via the Coulomb interaction, excitons can scatter with lattice phonons, other excitons, free carriers, and impurities. Each of these scattering processes has its own regime of dominance, dependent on, for example, temperature or material quality.

When the scattering rate with other particles is much larger than the recombination rate for electrons and holes, a quasi-equilibrium state is reached. The detailed nature of such an interacting electron/hole plasma remains a a long-standing open question. The reasons for this can be traced to interplay between the intrinsic Coulomb interactions within the plasma, which give rise to both bound and scattering states, and Pauli exclusion arising from the fermionic nature of the electrons and holes. This leads to a complex phase diagram which encompasses, e.g., electron-hole droplets, the ionized electron-hole plasma, biexciton phase, and the exciton gas which, some suggest, may undergo Bose–Einstein condensation. The parameters of this phase diagram are carrier density, temperature, and the semiconductor material parameters. Interest in this essentially fundamental question has remained high, stimulated by the stream of technological benefits that even partial answers have brought.

In this context it is worth mentioning that there have been two theoretical approaches to exciton physics which each have their advantages and problems. One approach is to treat excitons as bosonic quasi-particles and derive results from Hamiltonians formulated using bosonic operators. This approach has the appeal of simplicity and produces acceptable results in the low-density regime. It does, however, omit a key feature of the constituent particles making up an exciton namely the fermionic nature of

electrons and holes. A more rigorous approach based around electron and hole operators has been followed but this is considerably more complex and numerically involved.

Excitonic Molecules

Just as two hydrogen atoms can lower their total energy by forming a bound hydrogen molecule, so two excitons can make a bound complex which has lower energy than the two isolated excitons. Such complexes are called biexcitons and are mostly important in high quality material and at low temperatures. Their binding energy is less than that of the exciton by a factor of about 0.1–0.3, depending on the ratio of electron to hole effective masses.

Three particle complexes have also been observed consisting of two electrons bound to a hole. These are termed trions or charged excitons. Their binding energy lies intermediate between excitons and biexcitons.

See also: Foundations of Coherent Transients in Semiconductors

Further Reading

Bastard, G., 1998. Wave Mechanics Applied to Semiconductor Heterostructures. New York: Halstead John Wiley.
Haug, H., Koch, S.W., 1993. Quantum Theory of the Optical and Electronic Properties of Semiconductors. Singapore: World Scientific.
Klingshirn, C.F., 1997. Semiconductor Optics. Berlin: Springer.
Rashba, E.I., Sturge, M.D. (Eds.), 1982. Excitons. Amsterdam: North Holland.
Schmitt-Rink, S., Chemla, D.S., Miller, D.A.B., 1989. Linear and nonlinear optical properties of semiconductor quantum wells. Advances in Physics 38, 89–188.
Ueta, M., Kanzaki, H., Kobayashi, K., Toyozawa, Y., Hanamura, E., 1986. Excitonic Processes in Solids. Berlin: Springer.
Yu, P.Y., Cardona, M., 1999. Fundamentals of Semiconductors. Berlin: Springer.

Quantum Wells and GaAs-Based Structures

P Blood, Cardiff University, Cardiff, UK

Nomenclature

E energy
E_f Fermi energy
E_g bandgap
F envelope wavefunction
$f(E)$ Fermi–Dirac distribution function
g_{2D} two-dimensional density of states [energy]$^{-1}$ [area]$^{-1}$
$\hbar$ Planck's constant divided by 2π
I_{ph} optical field intensity
k wavevector
k_B Boltzmann's constant
m electron mass
n carrier density, expressed per unit area in quantum well structures
Q_c, Q_v heterostructure band offset ratios
r position vector
t time
T_N fraction of light transmitted at normal incidence through N quantum wells
$\langle \mathbf{x} \rangle$ unit vector along x-direction, similarly for y, z directions
α optical absorption coefficient in a bulk material, per unit length
γ fraction of light absorbed by a quantum well at normal incidence to the plane of the well
λ wavelength
Φ photon flux (photons per unit time crossing unit area)
Ψ electronic wavefunction
ν light frequency

Introduction

A quantum well is a potential minimum within a semiconductor structure which is sufficiently thin to localize charge carriers on a length-scale similar to their de Broglie wavelength which, for an electron in GaAs at room temperature, is about 30 nm. When electrons are localized in this way their electronic and optical properties are determined by quantum mechanical aspects of their behavior which are not apparent in larger-scale structures. The potential well is usually formed by a sandwich of a thin layer of narrow-gap semiconductor between two layers of a wider gap material as depicted in Fig. 1. Typical quantum wells have widths of about 5 nm and the key to their routine production has been the development of advanced epitaxial semiconductor crystal growth techniques such as metalorganic vapor phase epitaxy (MOVPE) and molecular beam epitaxy (MBE). The most significant

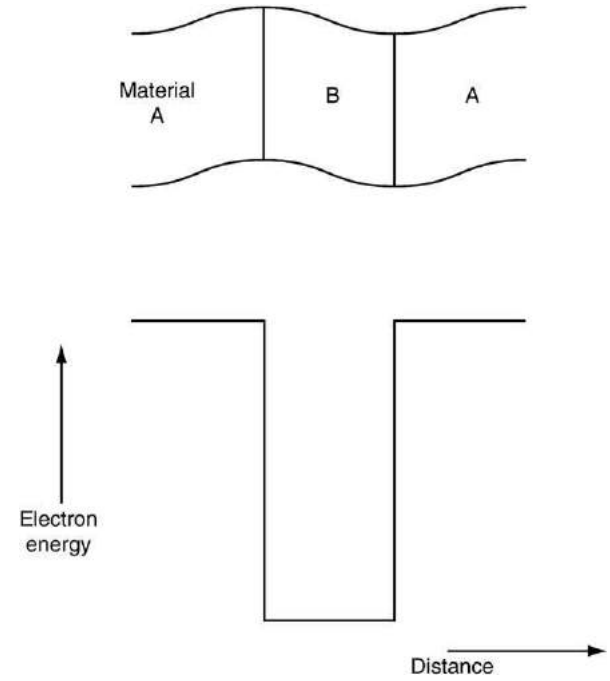

Fig. 1 Electron energy diagram illustrating the formation of a potential well by sandwiching a layer of narrow gap semiconductor material (B) between two wider gap layers (A).

 doi:10.1016/B978-0-12-809283-5.00619-4

application of these structures has been in opto-electronic devices, especially laser diodes, and in this article I describe how electrons behave in quantum wells with particular reference to optical properties.

The optical properties of semiconductors are usually considered in terms of transitions of electrons between quantum mechanical energy levels, as in the Bohr atom. An alternative description is to consider the spatial displacement of electrons in the oscillating electric field of the electromagnetic radiation. These two descriptions are linked through Schrödinger's equation. For each energy state there is an associated wavefunction which determines the spatial probability distribution of electrons in the state. Consequently, a change in quantum state implies a change in spatial distribution and energy of the electron system. Both approaches are used: here I adopt the viewpoint of transitions between energy states.

We begin with a quantum mechanical description of electrons in a 'bulk' material, as appears in many textbooks. Here electrons are constrained by a three-dimensional potential well of extent given by the sample dimensions, typically millimeters or centimeters and therefore large compared with the de Broglie wavelength. Then we examine what happens when the size of the sample is reduced in one direction: in the limiting case of a quantum well electrons are able to move in only two directions. Quantum confinement is also possible in two or three dimensions to produce a quantum wire or quantum dot, respectively, and it is easy to use the concepts developed in this article to determine the properties of such systems.

Electron States in Bulk Material

We first consider the behavior of an electron in one direction (the x-direction) within a potential well with large dimensions (side length L), then generalize this to three directions (see **Fig. 2**). The wavefunction ψ_n and energy E_n (measured with respect to the bottom of the well) are given by solutions of Schrödinger's equation, which within the well is:

$$\frac{\hbar^2}{2m}\frac{\mathrm{d}^2\psi_n}{\mathrm{d}x^2} = E_n \tag{1}$$

It is assumed that the potential is infinitely deep compared with the energy of the electrons. Within the sample the electrons are described by plane waves of the form $\psi_n = A_n \sin(k_n x)$, where k_n is a wavevector ($=2\pi/\lambda_n$, where λ_n is the wavelength), and substitution into **Eq. (1)** gives the corresponding energy eigenvalues as

$$E_n = \frac{\hbar^2 k_n^2}{2m} \tag{2}$$

where m is the mass of the electron. [Substituting $E_n = k_B T$ at room temperature (0.025 eV) and using an effective mass of $0.067m_0$ (for GaAs) in **Eq. (2)** gives $k_n = 2.1 \times 10^8\ \mathrm{m}^{-1}$ and $\lambda = 30$ nm.] Applying cyclic boundary conditions which permit traveling-wave solutions of **Eq. (1)** the wavelength must satisfy the condition $\lambda_n = nL$, i.e., $k_n = 2\pi n/L$, as illustrated in **Fig. 2**, where n is an integer. The dimensions of a typical sample are much greater than the 'size' of an electron so n can be a very large number. The energy eigenvalues are therefore given by

$$E_n = \frac{\hbar^2}{2m}\left(\frac{2\pi n}{L}\right)^2 \tag{3}$$

This treatment can be extended by solving **Eq. (1)** for motion in three orthogonal directions, x, y, and z. In this case electron motion is described by a wavevector $\mathbf{k}$ comprising components k_x, k_y, k_z, each of which satisfies the cyclic boundary condition in the respective direction. We take the sample to be of dimension L in each direction with no loss of generality. Thus (representing unit vectors by $\langle \mathbf{x} \rangle$, etc.)

$$\begin{aligned}\mathbf{k} &= k_x\langle \mathbf{x}\rangle + k_y\langle \mathbf{y}\rangle + k_z\langle \mathbf{z}\rangle \\ &= \frac{2\pi n_x}{L}\langle \mathbf{x}\rangle + \frac{2\pi n_y}{L}\langle \mathbf{y}\rangle + \frac{2\pi n_z}{L}\langle \mathbf{z}\rangle\end{aligned} \tag{4}$$

the energy is

$$E = \frac{\hbar^2}{2m}\{k_x^2 + k_y^2 + k_z^2\} \tag{5}$$

and the wavefunction takes the vector form

$$\psi_k(\mathbf{r}) = A_k \sin(\mathbf{k}\cdot\mathbf{r}) \tag{6}$$

$\psi_n^2(\mathbf{r})$ represents the probability of an electron being at the location $\mathbf{r}$. Each allowed electron state obtained by solution of Schrödinger's equation is specified by a unique combination of the numbers $(n_x n_y n_z)$, which take both positive and negative values corresponding to plane waves traveling in positive and negative directions. The Pauli exclusion principle states that only one electron of each spin can occupy a given quantum state so the amplitude of each wavefunction is normalized such that

$$\int \psi_k^2(\mathbf{r})\mathrm{d}\mathbf{r} = \int [A_k \sin(\mathbf{k}\cdot\mathbf{r})]^2 \mathrm{d}\mathbf{r} = 1 \tag{7}$$

where the integrals are evaluated over the volume of the sample. Since the amplitude of these solutions is constant throughout the sample electrons may be anywhere in the sample with equal probability. The motion of an *individual* electron of momentum $\hbar\mathbf{k}$ is represented by a wavepacket formed by combination of a number of wavefunctions having similar values of $\mathbf{k}$.

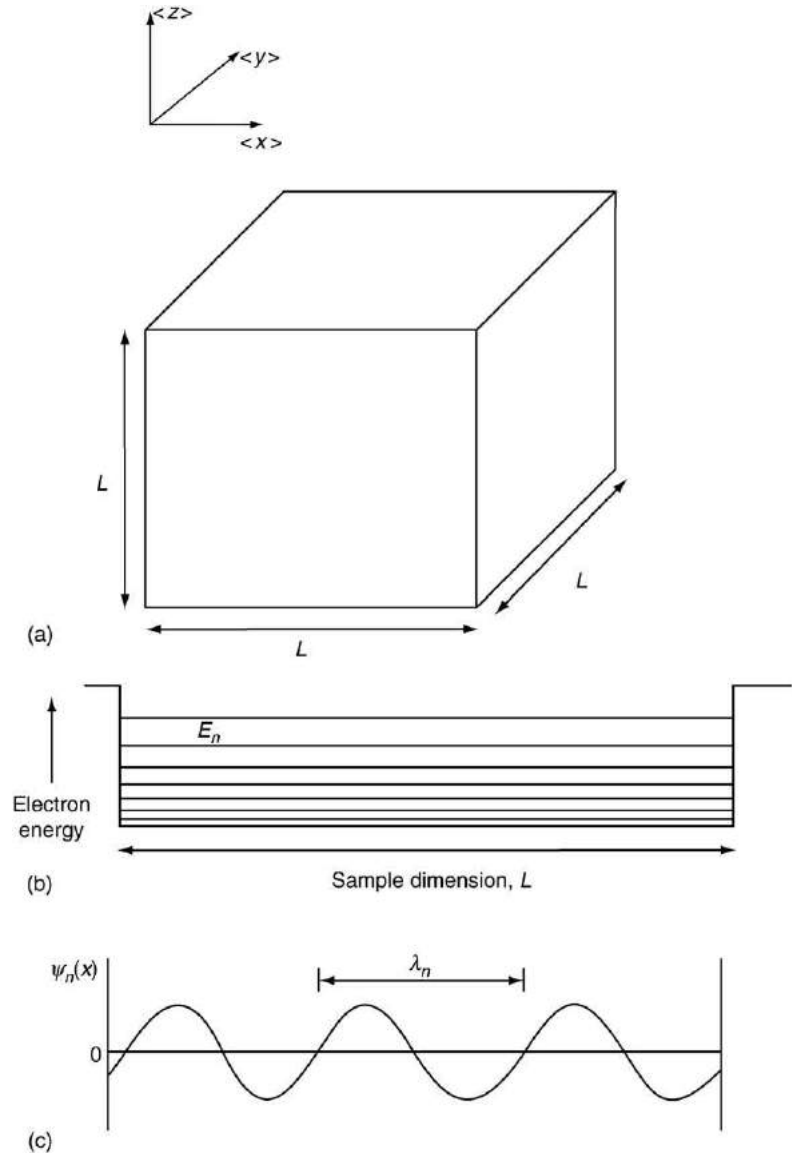

Fig. 2 The properties of electrons in a large cubic sample (a) can be obtained by solving Schrödinger's equation for the potential well formed by the sample in each direction (b). This potential well keeps the electrons within the material. The wavefunctions $\psi(x)$ shown in part (c) satisfy cyclic boundary conditions as described in the text, leading to a series of electron energy levels, E_n, given by **Eq. (3)**, shown in (b) of the figure.

Because the sample dimensions are large the energies of the states allowed by Schrödinger's equation take a series of very closely spaced values for increasing integer values of n as shown in **Fig. 2(b)**. Substituting a typical sample size of $L=1$ into **Eq. (3)**, the $n=1$ and 2 states are separated by only 7×10^{-11} eV in GaAs and these values are very small compared with thermal energies: at room temperature, $k_BT=0.025$ eV. The allowed values of $\mathbf{k}$ and E are discrete and there is a finite but very large number of allowed states in the sample. Since these are very closely spaced we can regard $\mathbf{k}$ and E as continuous variables and the allowed energy states form a continuum. We can calculate the number of allowed energy states dN in a small energy interval, dE, and hence determine the density of states in energy $g=(dN/dE)$ (the number per unit energy interval) at any value of energy.

Electron States in Quantum Wells

Now imagine a sample where the length of one side is reduced in the z-direction (L_z), as illustrated in **Fig. 3**. Gradually the allowed values of k_z become more widely spaced as a consequence of the boundary conditions (**Eq. (4)**). Eventually the width of the potential well becomes similar to the wavelength of the plane-wave state corresponding to the lowest energy level in the well. In these circumstances electron motion is not possible in the z-direction and the wavefunctions become standing waves with boundary conditions $n(\lambda_n/2)=L_z$ in an infinitely deep potential (where n_z takes only positive values) and the amplitude of the

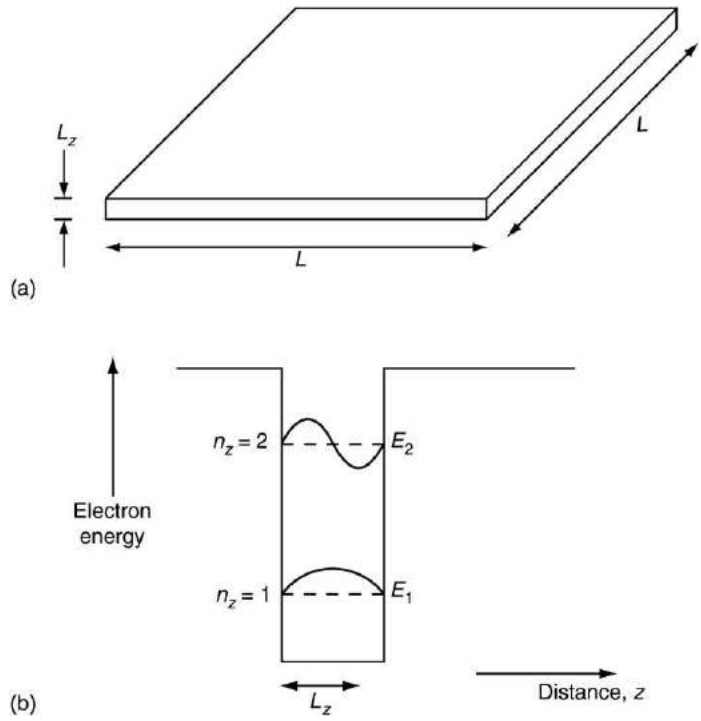

Fig. 3 Electron energy diagram (b) for a sample (a) in which one dimension, L_z, is very small. The electron motion in the z-direction is constrained and the associated energy levels become widely spaced, defined by the integers n_z=1,2,... in **Eq. (8)**. The wavefunctions in the z-direction, $\psi(z)$, are also illustrated for the first two electron levels.

wavefunction outside the well must be zero. The lowest energy state, given by **Eq. (5)** with $n_x=n_y=n_z=1$, no longer lies near the bottom of the potential well. Since the energies associated with motion in the x and y directions remain very small (of order 10^{-10} eV derived earlier) because L_x and L_y are large, the energy of the lowest state is effectively determined by the z-dimension because it is very small, so **Eq. (5)** gives

$$E_{n_z} = \frac{\hbar^2}{2m}\left(\frac{n_z\pi}{L_z}\right)^2 \tag{8}$$

for the lowest energy state in the well. For L_z=5 nm (and m=0.067) E_1 is 0.22 eV above the bottom of the well. **Eq. (8)** shows that this energy can be changed by choice of the well thickness. A simple interpretation of this behavior is as follows. When L_z becomes very small it is no longer possible for the z-component of the wavefunction of the lowest energy state to satisfy the boundary condition of zero amplitude at opposite sides of the sample. The condition can only be satisfied by wavefunctions which have a smaller wavelength, and consequently a higher energy.

Electrons can occupy states defined by all values of ($n_xn_yn_z$), thus for n_z=1 there is a continuum of allowed states at increasing energies corresponding to increasing values of (n_xn_y) and corresponding to motion in the (x, y) plane (**Fig. 4(a)**). Since L_x and L_y are both large these states are closely spaced and form a continuum. There is a similar continuum of energy states above the energy levels for n_z=2, 3, etc. Thus the energy level diagram is a series of sub-bands, each defined by n_z and with a continuum of states associated with motion in two dimensions in the (x,y) plane, as shown in **Fig. 4**. The number of continuum energy states in a given sub-band in a small energy interval gives the density of states for both spin directions

$$g_{2D}(E) = \frac{m}{\pi\hbar^2} \tag{9}$$

per unit energy interval per unit sample area.

The density of states is independent of n_z and therefore the same for all sub-bands, and is independent of energy. The density of states function is therefore a series of steps of height given by **Eq. (9)** at each sub-band energy as illustrated in **Fig. 4(b)**. The density of states per unit area within a given sub-band is independent of L_z as a consequence of dealing with a two-dimensional system.

In real samples the quantum well is not infinitely deep: typically the well depth is in the range 100–400 meV, consequently electrons are not totally confined to the well but are able to penetrate into the barrier material by quantum mechanical tunneling. Solution of Schrödinger's equation for a finite well (treated in many quantum mechanics textbooks) again yields a series of sub-bands, finite in number and with an energy spacing smaller than that given by **Eq. (8)**. In a typical GaAs well the lowest energy state may be about 100 meV from the bottom of the well. The density of states in each sub-band, which is due to the x,y) motion, remains unchanged. The wavefunction in the z-direction $F(z)$ comprises a sine wave-like standing wave within the well and an

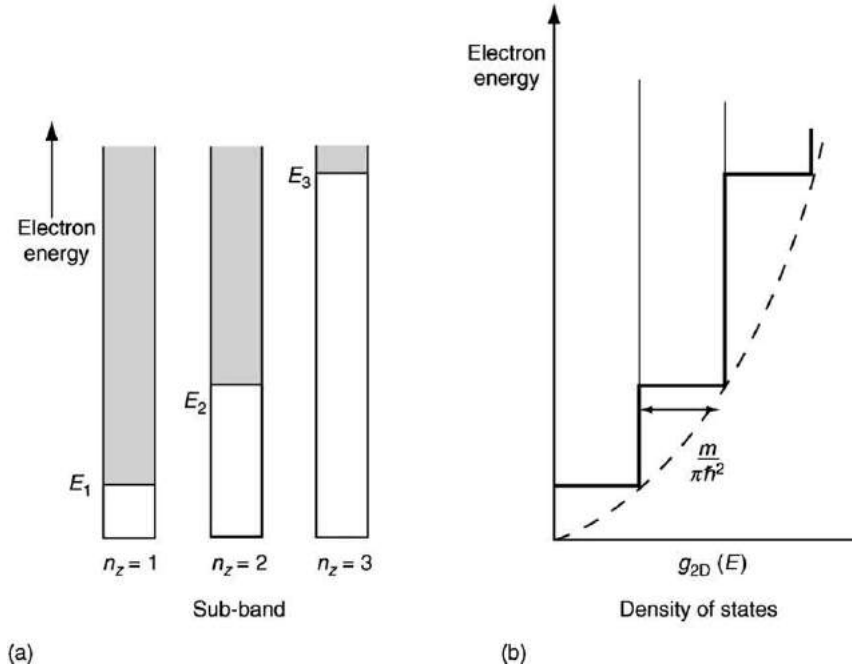

Fig. 4 (a) Electron energy diagram showing the energy states associated with unconstrained motion in the (x,y) plane for each sub-band formed by localizing the electrons in the z-direction, defined by $n_z = 1,2,3$, etc. When all these allowed states are summed at any energy and expressed as a number of states per unit energy interval we obtain the density of states function $g_{2D}(E)$ shown in (b). This has a series of steps corresponding to each sub-band edge.

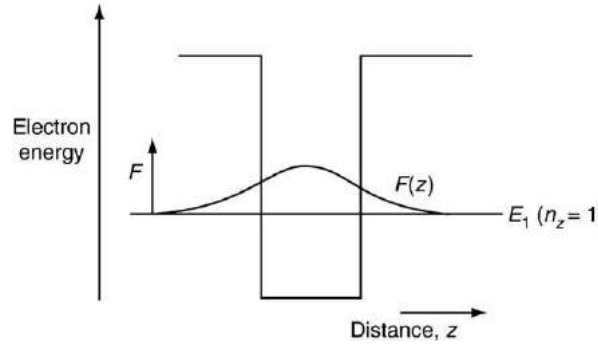

Fig. 5 Illustration of the envelope wavefunction $F(z)$ in the z-direction for an electron localized in the $n_z = 1$ sub-band of a well of finite depth. The electron is able to penetrate the barrier by tunneling and this is represented by the decaying part of the wavefunction outside the well.

exponentially decaying wave in the barrier representing tunneling (Fig. 5). The wavefunction is therefore made up of (x,y) plane-wave components and the localized z component and is of the form

$$\Psi_k(\mathbf{r}) = A_k \sin(\mathbf{k}_{xy} \cdot \mathbf{r}_{xy}) F(z) \tag{10}$$

where $\mathbf{k}_{xy}$ and $\mathbf{r}_{xy}$ are the wavevector and position vector in the (x,y) plane. This wavefunction is again normalized according to Eq. (7) such that each state is occupied by one electron of a given spin direction. We see from Fig. 5 that we cannot regard the electron as being localized in the well: the electron distribution in the z-direction is specified by the probability distribution $F^2(z)$.

Occupancy of States in Quantum Wells

The density of states function describes the energy distribution of allowed states in the quantum well. Whether or not particular states are occupied by electrons is determined by the electron concentration and the temperature through the thermodynamic properties of the system. In most cases electron–electron scattering is sufficiently rapid to bring the electron population into an internal equilibrium. Furthermore the interaction between the electrons and the crystal lattice is able to redistribute energy

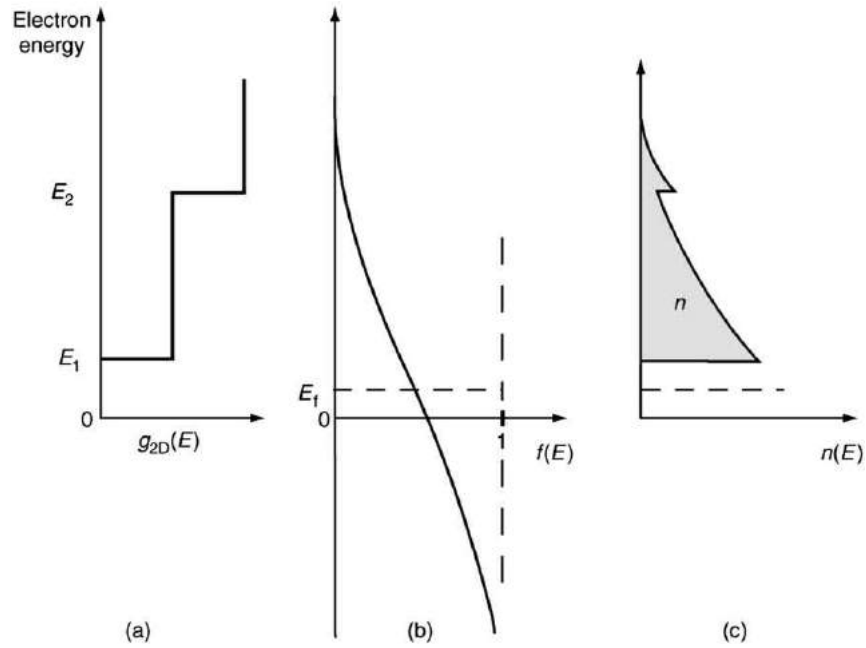

Fig. 6 (a) The density of states function for a quantum well, with the value of $g_{2D}(E)$ on the horizontal axis indicating how the density of states varies with increasing electron energy, (b) is an illustration of the probability that a state is occupied by an electron, showing how $f(E)$ varies with increasing electron energy, given by **Eq. (11)**. The actual density of electrons in a given energy interval at any energy $n(E)$ is given by the product of the density of available states and the probability that the state is occupied as shown in (c). The integral of $n(E)$ over energy gives the total density of electrons given for each sub-band by **Eq. (12)**.

between the two systems so that the lattice and the electron distribution have the same temperature. The electron distribution can therefore be described by Fermi–Dirac statistics for which the probability of a state at energy E being occupied with an electron is

$$f(E) = \frac{1}{1 + \exp\left\{\frac{E - E_f}{k_B T}\right\}} \tag{11}$$

where E_f is the chemical potential (which can be equated with the Fermi level for electrons), T is the lattice temperature and all energies are positive quantities measured with respect to the same arbitrary zero. The energy distribution of electrons in the well, $n(E)$, is then the product of the density of states function, $g_{2D}(E)$, and the occupation probability, $f(E)$, as depicted in **Fig. 6**. The total number of electrons, n, is the integral over energy. Combining **Eqs. (9)** and **(11)**, for a single sub-band gives:

$$n_a = \frac{m k_B T}{\pi \hbar^2} \ell n\left\{1 + \exp\left(-\frac{E_n - E_f}{k_B T}\right)\right\} \tag{12}$$

per unit area, single sub-band.

We cannot specify the carrier density in the well 'per unit volume' since the extent of the distribution in the z-direction is not defined because electrons may tunnel into the barrier material. The electron distribution is properly specified as a number per unit area having a probability distribution in the z-direction given by the function $F(z)$. Where more than one sub-band is populated the total electron density is obtained by adding the contributions from each sub-band each given by **Eq. (12)** with the same Fermi energy and the appropriate sub-band energy. We always use Fermi factors to specify the probability of occupation of a state by an *electron*. The probability that a state is empty is $(1 - f)$.

Formation of Quantum Wells

A quantum well is formed by sandwiching a narrow gap semiconductor layer (E_{g1}) between two layers of wider gap material (E_{g2}). To avoid the formation of defects and dislocations, which have a deleterious effect on many properties of the structure, the constituent materials must have the same lattice parameter. **Fig. 7(a)** is an energy band diagram for an n-type double heterostructure with layers of micron dimensions. At each interface there are discontinuities in the conduction and valence bands, ΔE_c and ΔE_v, respectively, which account for the difference in the band gaps of the two materials, ΔE_g. While the value of ΔE_g is known for any system, the manner in which this difference is apportioned between the two band edges is not known a priori and the values of ΔE_c and ΔE_v must be obtained from experiment. The discontinuities are often expressed as fractions Q_c, Q_v of the band gap difference:

$$\Delta E_c = Q_c \Delta E_g, \quad \Delta E_v = Q_v \Delta E_g \tag{13}$$

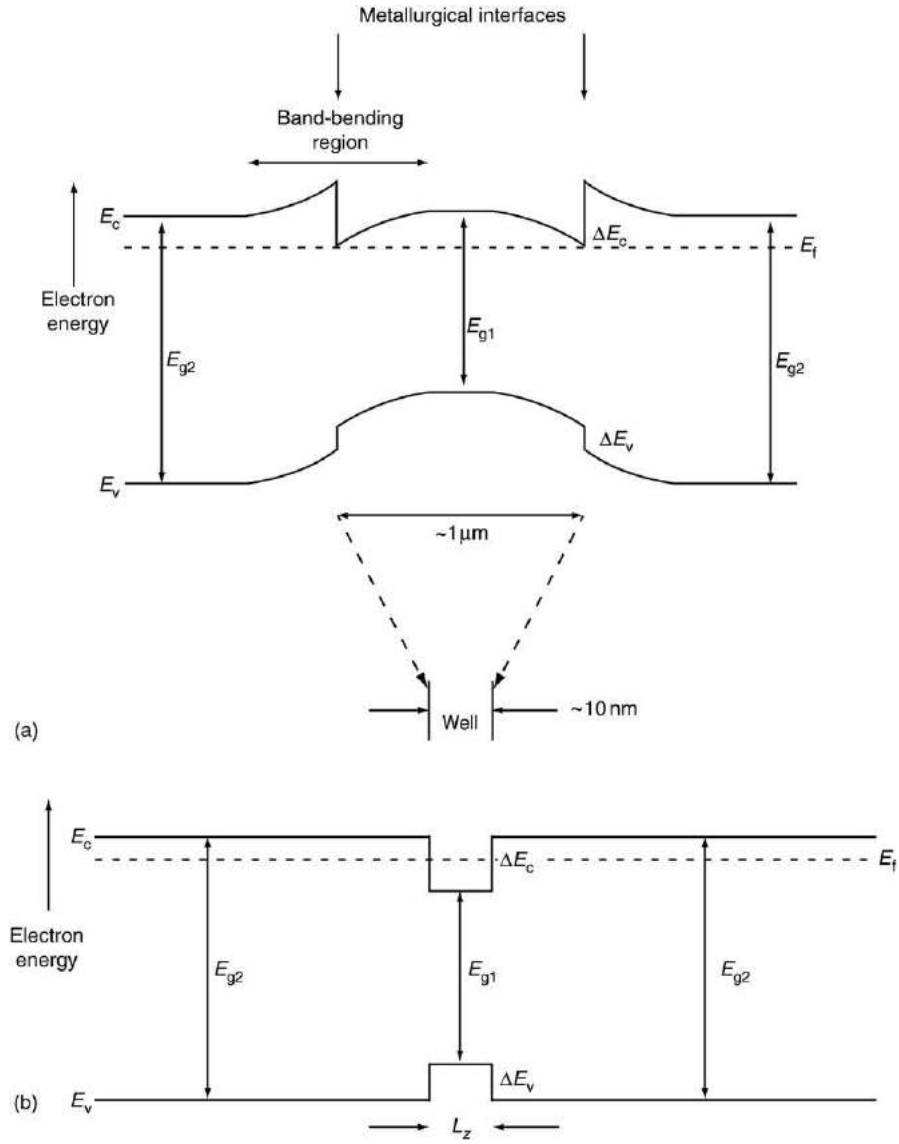

Fig. 7 Electron energy diagrams showing the formation of a quantum well as a very thin double heterostructure made up of materials of band gaps E_{g1} and E_{g2}. When the layers are thick the energy diagram is influenced by band bending at the interfaces as shown in (a). When the narrow gap layer is made thinner than the band-bending distances a rectangular well is formed in the conduction and valence bands with depths determined by the respective band offset as shown in (b). On the small distance scale of the well the bands are flat in each region of the structure.

In equilibrium the Fermi energy is constant across the diagram and there is a band-bending region on each side of the discontinuity such that the conduction band edges return to their equilibrium energy values at large distances from the interface. The extent of these regions is determined by the discontinuity and the doping level and is typically 200–1000 nm. At each single interface a triangular well is formed which localizes electrons and whose precise shape is determined by the doping density, the carrier distribution and discontinuity through simultaneous solution of Schrödinger's equation and Poisson's equation. Single heterointerfaces only localize the majority carrier. Quantum confinement of both types of carrier can be achieved by a very thin narrow gap semiconductor layer.

When the layer of narrow gap material is made thinner than the band-bending regions, the conduction band edge in this layer is not able to regain its bulk equilibrium position and the two triangular wells coalesce as illustrated in Fig. 7(b). When the well is sufficiently thin, say L_z less than 20 nm, the band-bending across this layer is negligibly small and the well becomes effectively rectangular. This well accumulates electrons from the surrounding barrier material and its potential relative to the Fermi level rises such that there is band bending in the barrier material on each side of the well inhibiting further accumulation. Since the band bending in the barrier occurs on a distance of hundreds of nanometers (being determined by the doping density) it is not apparent over the distance scale of the diagram of the quantum well, which is on the order of 10 nm. Consequently, as shown in Fig. 7(b), the well can be drawn as a rectangle to a good degree of approximation. This thin, narrow gap layer localizes both electrons and holes. The form of the well is determined by the respective band discontinuities, ΔE_c and ΔE_v, and the thickness of the layer.

Potential wells formed in this way are the basis for most quantum-confined device structures. Other forms are possible. In certain material combinations ΔE_c or ΔE_v may be negative, meaning that, at the interface, both discontinuities are in the same sense and only one carrier type is confined in the thin narrow gap material. These are known as 'type II' structures. If there is an electric field across the well due to doping or strain (piezoelectric) effects the well becomes triangular. Wells can also be engineered with steps in the potential profile and multiple well systems can be grown in which the states are quantum mechanically coupled. All these variants provide opportunities to engineer the properties of the structure.

One important development of the rectangular quantum well deserves mention. It is possible for the well material to have a different lattice parameter to that of the surrounding barrier material provided that the strain energy can be accommodated elastically within the structure. This means that the strain energy in the layer must be below the energy necessary for formation of dislocations and this translates into an upper limit on the layer thickness (the critical thickness) for a given mismatch. Incorporation of elastic strain is significant because relaxing the lattice match requirement widens the choice of well and barrier materials and because strain modifies the properties of the electronic states in the quantum well and these effects can be used to advantage in the design of devices.

Here we concentrate on the rectangular potential well. Such wells are widely used, particularly in opto-electronic devices, and this structure provides the basic concepts which lie at the heart of all quantum-confined systems. We consider the 'model' lattice-matched quantum well system: a GaAs well bounded by AlGaAs barriers.

GaAs/AlGaAs Quantum Well Structures

As Al is substituted for Ga in the $Al_xGa_{1-x}As$ alloy system the direct band gap increases from 1.424 eV in GaAs ($x=0$) to 3.018 eV in AlAs ($x=1$) (Fig. 8) while the lattice parameter remains unchanged. A quantum well is formed using a narrow GaAs layer sandwiched between barrier layers of $Al_xGa_{1-x}As$ having a composition chosen to give the desired well depth. The whole structure is grown on a GaAs substrate. The band discontinuity ratio between GaAs and $Al_xGa_{1-x}As$ is independent of alloy composition (ΔE_c and ΔE_v are constant fractions of ΔE_g as x is varied) with $Q_c=0.66$ and $Q_v=0.33$. For a typical barrier composition of $x=0.3$, $\Delta E_g=0.374$ eV and ΔE_c and ΔE_v are 0.247 eV and 0.127 eV, respectively.

Optical Properties of AlGaAs/GaAs Quantum Wells

General Principles

Fig. 9 illustrates a transition of an electron between an occupied state at E_1 in the $n_z=1$ sub-band in the conduction band and an empty state (i.e., a hole) at E_2 in the $n_z=1$ sub-band in the valence band, resulting in emission of a photon of energy $h\nu=E_1-E_2$ as required by energy conservation. This is the process of luminescence (or spontaneous emission), which requires external excitation such as illumination (photoluminescence) or biasing a p-n junction (electroluminescence, as in a light-emitting diode).

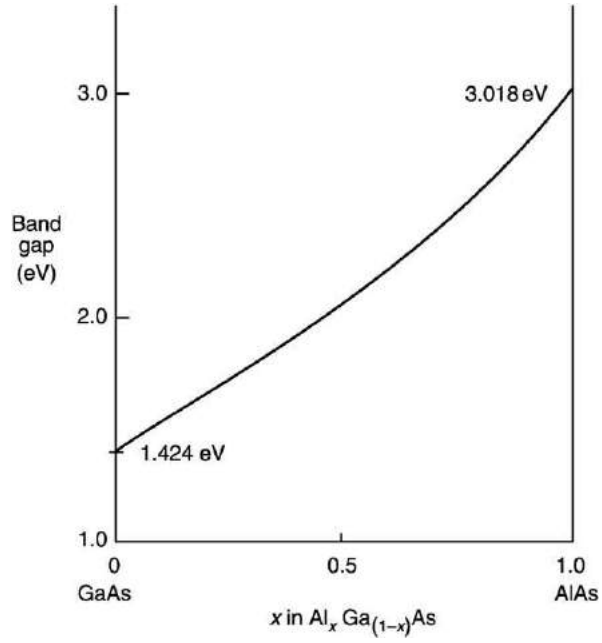

Fig. 8 Variation of the direct band gap of $Al_xGa_{(1-x)}As$ with Al content (x).

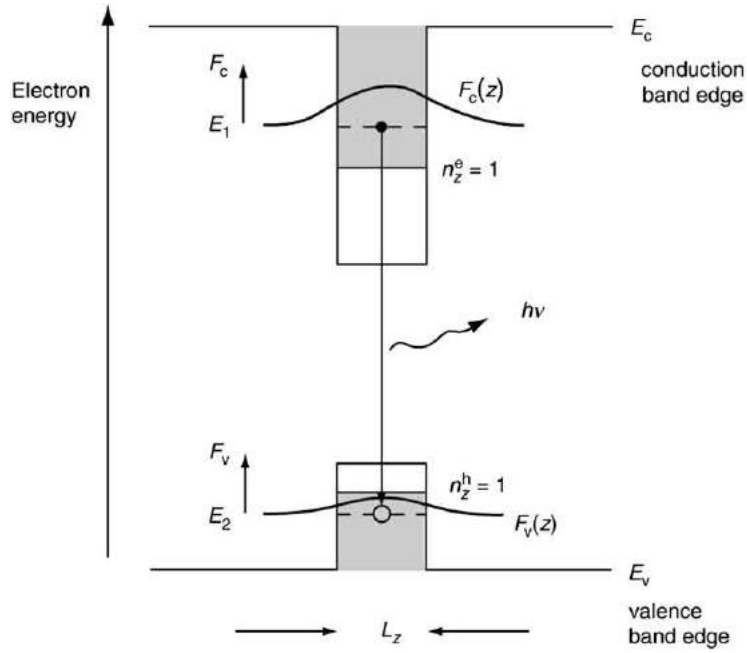

Fig. 9 Illustration of a downward optical transition of an electron from a state in the $n_z=1$ conduction sub-band to an empty state (hole) in the $n_z=1$ valence sub-band resulting in emission of a photon of energy $h\nu = E_1 - E_2$. The spatial probability distribution of the electron and hole is specified by their respective envelope functions $F_c(z)$ and $F_v(z)$ and the probability of the transition occurring is proportional to their overlap integral, **Eq. (14)**.

Alternatively, incident light may promote an electron from an occupied state in the valence band to an empty state in the conduction band with the consumption of the energy of a photon. This is the process of optical absorption.

At typical temperatures the electrons and holes are at energies near to their respective band edges so the photon energy $h\nu$ is close to the separation of the $n_z=1$ sub-bands. Since these energies are determined by the well width it is possible to change the photon energy for emission or absorption by changing the well width. As the well width is reduced the energy separation increases and the wavelength of light emitted by the structure is reduced. This ability to engineer the emission wavelength by simply adjusting the well width, without changing the chemical composition of the constituent materials, is one of the major attractions of quantum-confined structures. The shortest wavelength possible for a given material combination is determined by the band gap of the wide gap barrier material alongside the well.

The strength of the luminescence signal or the absorption is determined by the rates at which the electronic transitions occur. In general these rates are determined by the following intrinsic factors.

1. *The quantum mechanical probability of an electron making a transition between a full and an empty state*. This depends upon the 'matrix element' for the transition, which is a fundamental property of the well material and the overlap of the envelope functions. The electron and hole in the initial and final states must be in the same region of space and if the spatial distributions of electrons and holes, determined by F_c and F_v, do not overlap the probability for an electron to make a transition to the hole state is very low. The overlap is almost complete in a rectangular well (see **Fig. 9**), but is only partial in a triangular well. The transition probability is proportional to
$$I_{\text{overlap}} = \left\{ \int F_v(z) F_c(z) \mathrm{d}z \right\}^2 \tag{14}$$
2. *The probability of occupation of initial and final states by electrons, specified by the Fermi factors* (**Eq. (11)**). The initial state must be occupied and the Pauli exclusion principle prohibits an electron entering a final state which is already occupied. Thus for a downward transition resulting in emission of light the rate is proportional to $f_c(1-f_v)$.
3. *The density of initial and final states* (**Eq. (9)**). The more states there are within a given small energy interval, the greater the emission rate per unit energy.
4. *Photon distribution*. For processes such as optical absorption, which are 'induced' by the presence of a photon, the transition rate is also proportional to the photon density in the region of the well. Because the electrons are not wholly localized in the well, and because the spatial extent of the optical field is larger than that of the carriers, the coupling between the photon field (intensity $I_{\text{ph}}(z)$) and the carrier probability distributions is expressed as
$$I_{\text{coupling}} = \frac{\left\{ \int F_v(z) [I_{\text{ph}} z]^{1/2} F_c(z) \mathrm{d}z \right\}^2}{\int I(z) \mathrm{d}z} \tag{15}$$

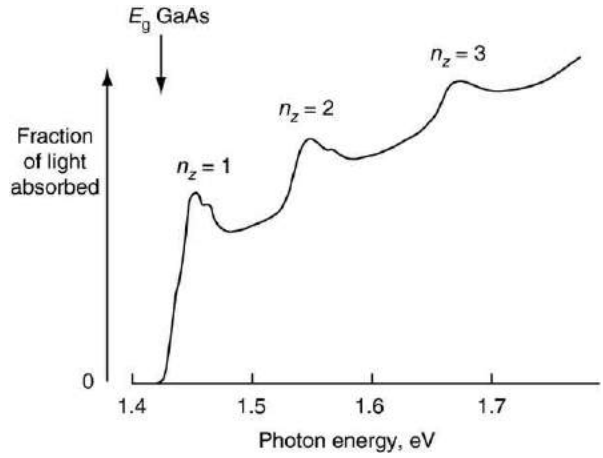

Fig. 10 Absorption spectrum of a multiple quantum well structure having 10 nm GaAs wells measured at room temperature. The absorption edges corresponding to transitions between n_z=1,2, and 3 pairs of sub-bands can be identified. The peaks are due to absorption by formation of excitons.

5. *Optical mode density*. The emission rate also depends upon the density of optical modes available to receive the recombination radiation. In most situations the emission occurs into the cavity of free space which is very large compared with the wavelength of the radiation so the modes are very closely spaced and the cavity can accept radiation at any transition energy (this is the well-known Raleigh–Jeans theory of cavity radiation). However, in some special situations the cavity has dimensions comparable with the wavelength of the radiation (microcavities) and then the modes are widely spaced. In such a microcavity, the optical cavity effectively controls the electronic transitions.

Optical Absorption

Fig. 10 is an optical absorption spectrum for a GaAs quantum well measured with light incident normal to the plane of the quantum wells. A series of steps can be seen which correspond to absorption at increasing photon energy between increasing orders of sub-bands, $E_{1e} \rightarrow E_{1h}$, $E_{2e} \rightarrow E_{2h}$, etc. Normally $n_z = 1$ to $n_z = 2$ transitions are forbidden in rectangular wells. At each step there is a pair of peaks (not always resolved) and above each step the absorption is roughly constant, following the step-like density of states function sketched in Fig. 4. The peaks are due to the formation of excitons which are weakly bound electron–hole pairs. In bulk materials excitons are only observed at low temperatures whereas in quantum wells the localization increases the binding energy (E_{bX}, typically 6–8 meV) such that they are observed at room temperature. The excitonic peak is at an energy E_{bX} below the sub-band separation. There are in fact two different kinds of electrons in the valence band with different masses so there are pairs of n_z=1,2,3,… sub-bands due to the different confinement energies for the different valence electrons. Spectra of this kind provided the first evidence for the distinctive sub-band structure of quantum well systems produced by carrier confinement. Absorption measurements provide data on the energy spacing of the sub-bands in the well, and it is possible to determine the well width and depths, from which the band offsets can also be determined.

The strength of the absorption is also of interest because it provides a measure of the optical matrix element if the Fermi factors f_c and f_v are known. If the incident optical beam is very weak such that the absorption process excites very few electrons then the system remains close to thermal equilibrium so the upper states are empty ($f_c = 0$) and the lower states are full ($f_v = 1$).

In *bulk material* the change in photon flux over a small distance Δx is proportional to the flux Φ and the distance traveled

$$\Delta\Phi = -\Phi\alpha\Delta x \tag{16}$$

where α is the absorption coefficient which has units L^{-1}. This results in an exponential decrease of photon flux with distance traveled through the material

$$\Phi(x) = \Phi_0 \exp(-\alpha x) \tag{17}$$

For a quantum well structure with light incident in the z-direction perpendicular to the plane of the quantum well, the fraction of light absorbed by a single well by transitions between a single sub-band pair (e.g., transitions between the $n_z = 1$ conduction and valence sub-bands) is independent of the thickness of the well because the density of states for a single sub-band is independent of well thickness. The strength of the absorption cannot be expressed by an absorption coefficient but by the fraction of light absorbed per well:

$$\Delta\Phi = -\gamma_w\Phi \tag{18}$$

This arises because the quantum well is effectively a sheet in the z-direction: changing its thickness does not change the density of states per unit area (Eq. (9)) and we cannot specify the electron concentration along the z-direction. This behavior also occurs because the well is much thinner than the wavelength of the light and it is not possible to specify the variation of photon flux on a

distance scale smaller than the wavelength. The barrier material surrounding the well has a wider band gap and does not absorb at the photon energies absorbed by the well. The well thickness does affect the sub-band spacing and at a fixed photon energy the absorption increases as the well width is decreased as light is absorbed by transitions between more sub-bands. Eq. (18) applies individually to each sub-band because the density of states is the same for each sub-band.

In a multiple well system, the fraction of light transmitted through N independent wells is

$$T_N = \{1 - \gamma_w\}^N \tag{19}$$

so the fraction of light absorbed by such a system is

$$\frac{\Delta\Phi}{\Phi} = -\left[1 - \{1 - \gamma_w\}^N\right] = -N\gamma_w \tag{20}$$

which is proportional to the number of wells when $\gamma_w \ll 1$.

The fraction of light absorbed by transitions between a single pair of sub-bands for a rectangular GaAs quantum well is predicted to be about 0.01, and a similar value has been obtained from absorption measurements. In a rectangular well the envelope function overlap is complete; however, in a triangular well of the same material the absorption is reduced by the smaller overlap of the envelope functions.

Photoluminescence and Spontaneous Emission

We turn now to discuss the optical emission from quantum wells as a result of external excitation. An external light source with photon energy above the effective band gap produces excess electron–hole pairs by excitation of electrons from the valence band. These rapidly lose their excess energy and thermalize to the sub-band edges to take up Fermi energy distributions. The electrons subsequently recombine to vacant valence band states by spontaneous emission. The low energy edge of this spectrum corresponds to the sub-band separation and the shape of the spectrum at higher energies corresponds to the thermal distribution of carriers in the bands. At very low excitation intensities where the excess electron population is small the luminescence may be due to recombination of electrons within an exciton and occur at an energy E_{bX} below the sub-band separation. Generally the sub-band separation is several k_BT so only the lowest sub-band is populated and photoluminescence is not seen from higher sub-bands. In this respect absorption measurements provide a more complete characterization of the structure.

While the energy position of features in the photoluminescence spectrum provides useful information, it is difficult to use the strength of the photoluminescence signal to provide quantitative information about the rate of optical transitions within the material because the excitation power and volume within the sample must be known and a known fraction of emitted light must be collected and measured with a calibrated detector. The relative intensity of photoluminescence under standard conditions does provide comparative information and the linewidth of the spectrum at low intensity provides a qualitative indication of 'quality' (e.g., well width variations in the structure). A quantitative determination of the internal recombination rate can be obtained by measuring the decay of the luminescence signal following short-pulse excitation. Fuller discussion can be found in books dealing with the characterization of semiconductor structures.

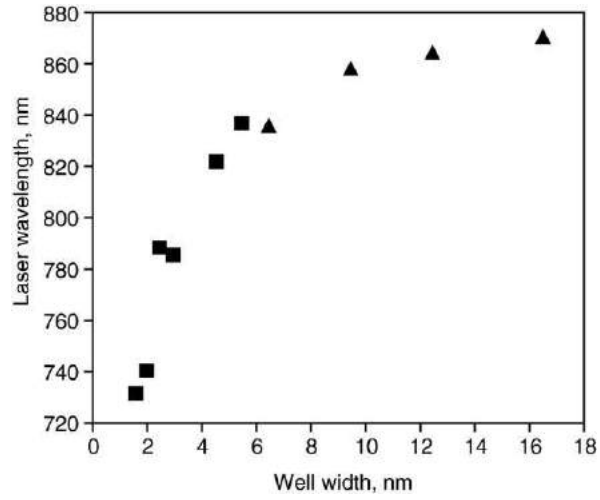

Fig. 11 Measured values of the emission wavelength of GaAs/AlGaAs quantum well lasers as a function of well width, showing how the wavelength can be engineered by choice of the well width. Data from: squares: Woodbridge K, Blood P, Fletcher ED and Hulyer PJ (1984) Short wavelength (visible) GaAs quantum well lasers grown by molecular beam epitaxy. *Applied Physics Letters* 45: 16–18; and triangles: Chen HZ, Ghaffari A, Morkoç H and Yariv A (1987) Effect of substrate tilting on molecular beam epitaxial grown AlGaAs/GaAs lasers having very low threshold current densities. *Applied Physics Letters* 51: 2094–2096.

Concluding Remarks

We have shown that an understanding of the electronic structure of quantum wells follows naturally from quantum mechanical treatments of electrons in bulk materials. When one of the dimensions of the structure is reduced to be similar to the de Broglie wavelength, quantization of the electron motion in this direction becomes apparent. The sub-bands which are formed have a profound effect on the electronic and optical properties when their spacing exceeds thermal energies such that only one band is significantly populated with carriers. The density of states function is a series of steps, each corresponding to a sub-band, and this produces characteristic step features in the optical absorption spectrum. The effective band gap of the structure is controlled by the width of the well. Quantum confinement also increases the binding energy of excitons compared with bulk materials and under low excitation conditions absorption and luminescence spectra are dominated by excitonic features even at room temperature.

One of the most successful application areas of quantum well structures has been in laser diodes. **Fig. 11** shows experimental data for laser emission wavelengths as a function of well width, showing the dramatic reduction which can be achieved simply by changing the width of the quantum well. Furthermore the step density of states function, characteristic of quantum wells, is one of the fundamental reasons for the reduction in threshold current of quantum well lasers relative to bulk GaAs devices.

We have considered GaAs because it is a model material system for production of quantum well structures. The concepts developed in this chapter are, however, quite general and can be applied to other material systems, other forms of quantum well and quantum wires and dots.

See also: Cavity QED in Semiconductors

Further Reading

Bastard, G., 1988. Wave Mechanics Applied to Semiconductor Heterostructures. New York: Les Editions de Physique, Halstead Press.
Bimberg, D., Grundmann, M., Ledentsov, N.N., 1999. Quantum Dot Heterostructures. Chichester, UK: Academic Press.
Blood, P., 1991. Heterostructures in lasers. In: Morgan, D.V., Williams, R.H. (Eds.), Physics and Technology of Heterojunction Devices. Stevenage, UK: Peter Peregrinus.
Blood, P., 2000. On the dimensionality of optical absorption, gain and recombination in quantum-confined structures. Journal of Quantum Electronics 36, 354–362.
Blood, P., Orton, J.W., 1992. The Electrical Characterisation of Semiconductors: Majority Carriers and Electron States. London: Academic Press.
Casey Jr, H.C., Panish, M.B., 1978. Heterostructure Lasers. San Diego, CA: Academic Press.
Dingle, R., Gossard, A.C., Weigmann, W., 1975. Direct observation of superlattice formation in semiconductor heterostructures. Physical Review Letters 34, 1327–1330.
Eisberg, R., Resnick, R., 1985. Quantum Physics of Atoms, Molecules, Solids, Nuclei and Particles, second ed. New York: John Wiley.
Hook, J.R., Hall, H.E., 1991. Solid State Physics. Chichester, UK: John Wiley.
Kelly, M.J., 1995. Low-Dimensional Semiconductors. Oxford, UK: Clarendon Press.
Orton, J.W., Blood, P., 1990. The Electrical Characterisation of Semiconductors: Measurement of Minority Carrier Properties. London: Academic Press.
Perkowitz, S., 1993. Optical Characterisation of Semiconductors. London: Academic Press.
Wolfe, C.M., Holonyak Jr, N., Stillman, G.E., 1989. Physical Properties of Semiconductors. Englewood Cliffs, NJ: Prentice Hall.

Recombination Processes

PT Landsberg, The University of Southampton, Southampton, UK

Introduction

In studies of recombination there occur considerable complications; many are associated with interactions between electrons, electrons and phonons, excitons, bi-excitons, impurity centers, etc. Most of these are not required in the present exposition. Modern work normally assumes that the basics of recombination physics are understood and the present exposition offers an appropriate outline.

Basic Assumptions

Electron–Electron Interactions for Electrons in Bands

The recombination problem in semiconductors is greatly complicated by the interaction of the electrons with each other. This allows one to speak only of the quantum states of the semiconductor crystal as a whole. However, as in metals, so also in semiconductors, a simplified picture is successful. In this, the electron interactions, and other interactions, are first neglected, but are later taken into account as a perturbation. Thus, the electrons are first treated as if they can move through each other; the fact that they collide and deflect each other by virtue of their Coulomb interaction is treated as a perturbation. The transitions are still described within the framework of the single-particle states of the unperturbed problem.

The Effect of Electron–Boson Interactions

Our first approximation is to neglect most (but not all) electron interactions. Later we take into account the two-electron transitions that arise. This two-particle recombination process is referred to as Auger recombination and its inverse as the generation of current carriers by impact ionization. In addition, electrons interact with the radiation and lattice fields and emit or absorb photons or phonons. These electron–boson interactions result in transitions of single electrons in an energy band scheme. We shall attach a superfix 'S' (for single-electron) to the recombination coefficients for such processes. These are then denoted by B^{S} or T^{S} depending whether only bands are involved or whether traps are also involved. They are illustrated in **Fig. 1**, where the solid horizontal lines represent the conduction and valence band edges and the dashed line refers to traps. Note that the two-electron transitions do not have the superfix 'S', while n and p refer to the electron and hole concentrations; N_0 and N_1 are the concentration of unoccupied and occupied trap states, respectively.

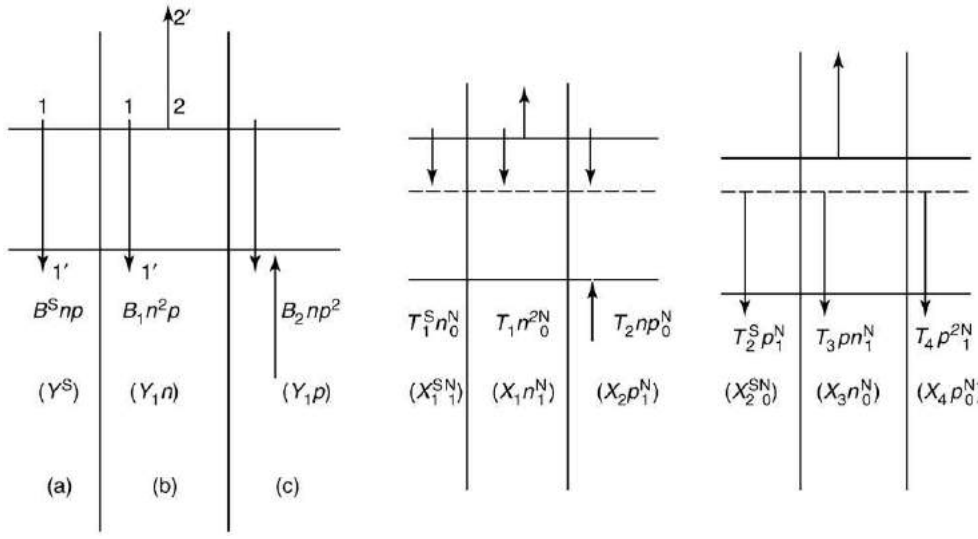

Fig. 1 Definition of recombination coefficients. Transition rates per unit volume are stated with each process, and, in brackets, for the reverse process. Thus B_1, B_2, T_1–T_4 refer to Auger processes; B^{S}, T_1^{S}, T_2^{S} refer to single-electron recombination; Y^{S}, X_1^{S}, X_2^{S} refer to carrier generation processes; Y_1, Y_2, X_1–X_4 refer to impact ionization processes. Arrows indicate transitions made by electrons.

 doi:10.1016/B978-0-12-809283-5.00618-2

Electron–Electron Interaction in Traps

Electron–electron interactions can at least formally be taken into account in connection with electrons trapped in a center: the spectrum is a function of the number, $r(=1, 2, 3,\ldots,M)$, of electrons captured. The 'irremovable' electrons can be included with the ion core. Given that the center captured r electrons (say), it can still be in a variety of quantum states, and they will be denoted by the symbol ℓ, yielding a set of quantum states (ℓ,r) for a center. The energy of such a state, divided by kT to make it dimensionless, is denoted by $\eta(\ell,r)$, yielding the canonical partition function

$$Z_r = \sum_{\ell} \exp[-\eta(\ell, r)] \tag{1}$$

The location of a center in space will here be considered to be of no significance.

Centers can capture several electrons. This brings in a need for the chemical potentials (or Fermi levels) for r-electron centers. They are here denoted by γ (when divided by kT) and the suffix 'eq' denotes an equilibrium value. Thus, the equilibrium probability of finding an r-electron center in state ℓ is given by:

$$P(\ell, r)_{\text{eq}} = \frac{\exp[r\gamma_{\text{eq}} - \eta(\ell, r)]}{\sum_{s=0}^{M} (\exp\ s\gamma_{\text{eq}}) Z_S} \tag{2}$$

One can identify the dominant energies from optical or electrical experiments.

To understand the properties of these centers, suppose we can arrange for the equilibrium Fermi level to rise from the valence band to the conduction band. At first practically no electrons are captured ($r=0$). As the Fermi level rises the states corresponding to $r=1$ begin to appear. The states ℓ of the center which correspond to the ground states in equilibrium are always more highly populated than the states ℓ corresponding to excited states, so that we can often confine attention to them. As the Fermi level rises, the ground states for $r=2$ become more important, and there may now be hardly any centers which have captured fewer electrons. If a larger number of electrons cannot be captured, by the time the Fermi level reaches the conduction band, then states of the center with $r>2$ can (normally) be neglected as unstable. That this is a satisfactory picture is our second approximation.

Assumptions for Nonequilibrium Statistics

The now much simplified recombination problem is still complex because of the many states available to electrons in bands and centers. A key simplification arises from the fact that it is often possible to talk about a small number of groups of quantum states: let $I(=1, 2,\ldots)$ label quantum states in a group i, let $J\ (=1, 2,\ldots)$ label quantum states in a group j, etc. Within each group it is supposed that the transitions are much more rapid than they are between groups. In that sense electrons in each group are in equilibrium among themselves. Their quasi-equilibrium is then characterized by what is called a quasi-Fermi level. The dimensionless version, obtained by dividing it by kT, is denoted by γ_i, γ_j, etc., for groups i, j, etc. That this is reasonable is our third assumption. Thus γ_c, γ_v are quasi-Fermi levels which refer to the quasi-equilibrium of the conduction and valence band, respectively, and γ_I to an I-electron center. Groups of states with different quasi-Fermi levels are not in equilibrium with each other.

Recombination problems can now be discussed by neglecting transitions within one (quasi-equilibrium) group, because they proceed at exactly the same rate as their reverse transitions. The number of transition types to be considered is thereby greatly reduced. Away from equilibrium we shall adopt **Eq. (3)** instead of **(2)** in order to allow for distinct quasi-Fermi levels:

$$P(\ell, r) = \frac{\exp[r\gamma_r - \eta(\ell, r)]}{\sum_{s=0}^{M} (\exp\ s\gamma_S) Z_S} \tag{3}$$

Eq. (3) is not always correct, but is an approximation arising from the quasi-Fermi level assumption. Among the improved theories are, for example, the 'cascade' theories.

The assumption of a quasi-Fermi level for each state of charge of a center implies that all the excited states of an r-electron center are populated according to **Eq. (2)**.

The Main Recombination Rates

General Results and The Two-Band Case

Recombination and its converse, generation, consist of a transition of an electron from one state to another. The observed rate is the net rate of recombination and is the algebraic sum of the recombination and generation processes. During these processes both total energy and total momentum must be conserved. This is achieved by creation or absorption of photons, or phonons, excitation of secondary electrons, etc.

The transition probability per unit time, S_{IJ}, for a single-electron transition from state I in a band to state J requires that state I should be occupied, with probability p_I say, and state J should be vacant with probability q_J say. The general expression for the average rate of the transition from I to J then takes the form $p_I S_{IJ} q_J$. For the reverse process electron state J has to be occupied and

state I vacant. Thus the rate of this reverse process is $p_J S_{JI} q_I$. The net recombination rate per unit volume of the process I to J can then be written as:

$$u_{IJ} = (p_I S_{IJ} q_J - p_J S_{JI} q_I) V^{-1} \tag{4}$$

By the principle of detailed balance, this expression vanishes at equilibrium. If one puts $X_{JI} \equiv S_{IJ} p_I q_J / S_{JI} p_J q_I$, one has $u_{IJ} = p_J S_{JI} q_I (X_{JI} - 1) V^{-1}$. In equilibrium $X_{JI} \to (X_{JI})_{\text{eq}} = 1$, and the recombination rate is zero.

One can say a little more if one assumes that states I and J are in conduction or valence bands, each with a quasi-Fermi level (divided by kT to make it dimensionless). If these are denoted by γ_{e} and γ_{h}, then $p_I = [1 + \exp(\eta_I - \gamma_{\text{e}})]^{-1}$ for a conduction band. For the total recombination rate per unit volume between the bands i and j one finds:

$$u_{ij} = \left[\sum_{I \in i} \sum_{J \in j} p_J S_{JI} q_I \right] \left[\exp\left(\frac{|e|\phi}{kT} \right) - 1 \right] V^{-1} \tag{5}$$

where ϕ is essentially the difference between the quasi-Fermi levels: $\gamma_{\text{e}} - \gamma_{\text{h}}$. The first factor depends on the bands involved and it has been assumed that the transition probability ratio S_{IJ}/S_{JI} is independent of the excitation.

A transition rate, when multiplied by the charge of current carriers, is a current, and when divided by the area of the surface involved, is a current density. If i and j denote the states of the conduction and valence bands of a semiconductor, excitation independence may often be assumed, and one then expects a current density proportional to $\exp(|e|\phi/kT) - 1$. This is characteristic of the current through pn junctions, metal semiconductor junctions, etc. In these configurations the Fermi level difference between the ends of the device determines the voltage across it. When radiation is involved, however, excitation dependence of some of the parameters introduced above (e.g. S_{JI}) tends to spoil this simple story.

The Case of Defects

So far we have considered only two bands. When a trap is involved matters are rather different. Because of the interactions among the electrons on a center it is not possible to talk of the same level being occupied or vacant. Consequently, identification of forward and reverse processes in terms of levels becomes impossible. Instead one deals with a center, say an r-electron center, as a whole; we then need the probability that a given center is an r-electron center. For example, the capture of an electron converts an $(r-1)$-electron center into an r-electron center. Thus the p's and q's must be replaced by more complicated expressions. It is convenient to denote u_{ij} by u_{cv} in the simple two-band case and its structure is given (with a sign change) by a recombination coefficient:

$$u_{\text{cv}} = B^{\text{S}} np [1 - \exp(\gamma_{\text{h}} - \gamma_{\text{e}})] \equiv B^{\text{S}} np - Y^{\text{S}} \tag{6}$$

using **Fig. 1(a)**. Here, n and p are the electron and hole concentrations. Auger effects in **Figs. 1(b)** and **1(c)** can also be included, at least formally. Then B^{S} has to be replaced in **Eq. (6)** by $B^{\text{S}} + B_1 n + B_2 p$. Analogous replacements are found for recombination processes involving defects.

In the simplest case of one type of localized defect one finds a steady-state recombination rate per unit volume of the form:

$$u = \frac{np - (np)_0}{\tau_{n0}(p + p_1) + \tau_{p0}(n + n_1)} \tag{7}$$

Here τ_{n0}, τ_{p0} are parameters with the dimension of time and p_1, n_1 are parameters with the dimension of concentration. The recombination increases with the defect concentration, which is actually in the denominators of τ_{n0} and τ_{p0}. This is an old and much used result associated with the names of W Shockley and W T Read.

Radiative Transitions

Spontaneous and Stimulated Emission

Quantum theory was initiated by Planck's law for black-body radiation at temperature T. This gives the spatial energy density as a function of frequency ν in this system as

$$f(\nu, T) = \frac{8\pi\nu^2}{c^3} \frac{h\nu}{\exp(h\nu/kT) - 1} \tag{8}$$

For low frequencies one finds the Rayleigh–Jeans law which makes **(8)** proportional to kT in agreement with the classical equipartition theorem. For high temperatures **(8)** diverges, as required.

The result **(8)** may also be obtained by writing

$$N = [A + Bf(\nu, T)] N_{\text{u}} \tag{9}$$

for the emission rate of photons by atoms, where N_{u} is the number of atoms in the upper of two states which are separated by the energy $h\nu$. The first term on the right-hand side is due to the normal decay of an excited state ('spontaneous emission'). The second term refers to additional emission ('stimulated emission') induced by the radiation of frequency $h\nu$ itself.

The first factor in (8) is due to the density of states and the second factor is due to the fact that the energy is considered. If one considers the number of photons of energy $h\nu$ at temperature T one comes up with

$$N_\nu = \left[\exp\frac{h\nu - \mu}{kT} - 1\right]^{-1} \quad \text{(Bose–Einstein distribution)} \tag{10}$$

Here μ is a possible chemical potential of the radiation which is non-zero only in nonequilibrium situations such as in a semiconductor laser.

In a pn junction the two bulk materials several diffusion lengths away from the junction are approximately in equilibrium even if a modest current is flowing. On one side one has then just one quasi-Fermi level, say μ_I, and on the other side one has just one quasi-Fermi level, say μ_J. Then the difference

$$\mu \equiv \mu_I - \mu_J = q\varphi \tag{11}$$

corresponds to the applied voltage φ. This is in agreement with Eq. (10), in that $\varphi = 0$ implies no current and hence thermal equilibrium is possible.

Statistical mechanics teaches one the rule that, in equilibrium, occupation probabilities of individual quantum states are always less for states of higher energies. This ensures that the total energy of a quantum system converges, the occupation probability p being a kind of convergence factor. However, if one is away from equilibrium, the above rule can be suspended and one can have population inversion.

For $h\nu = \mu$ there is trouble with (10) because the steady-state photon occupation diverges. This does not correspond to a 'death ray', but is the result of imperfect modeling; for example, the leakage of photons from the cavity may have been neglected. A formula of type (10) is also needed in connection with solar cells.

Donor–Acceptor Pair Recombination

A striking demonstration of radiative donor–acceptor transitions in GaP at low temperature (1.6 K) was revealed by sharp lines (Fig. 2) at photon energies $h\nu_i$ given by

$$h\nu_i = E_G - E_A - E_D + e^2/\varepsilon R_i \quad [-E] \tag{12}$$

where R_i is the distance between the ith-order nearest neighbors.

The discrete nature of the peaks is due to the fact that the impurities involved settle in general on lattice sites so that only definite separations R_i are possible. The energy gap is E_G; the value of $E_A + E_D$ can be inferred from the experimental lines by extrapolation to $R_i \to \infty$. The energy term E is sometimes neglected.

The ZnS phosphors were the first materials in which donor–acceptor radiation was hypothesized. However, it is hard to control its stoichiometry and its impurity content, and it has relatively large carrier mass and hence relatively large impurity activation energies. In such cases spectra such as those shown in Fig. 2 are hard or impossible to obtain. This applies in general to many II–VI compounds.

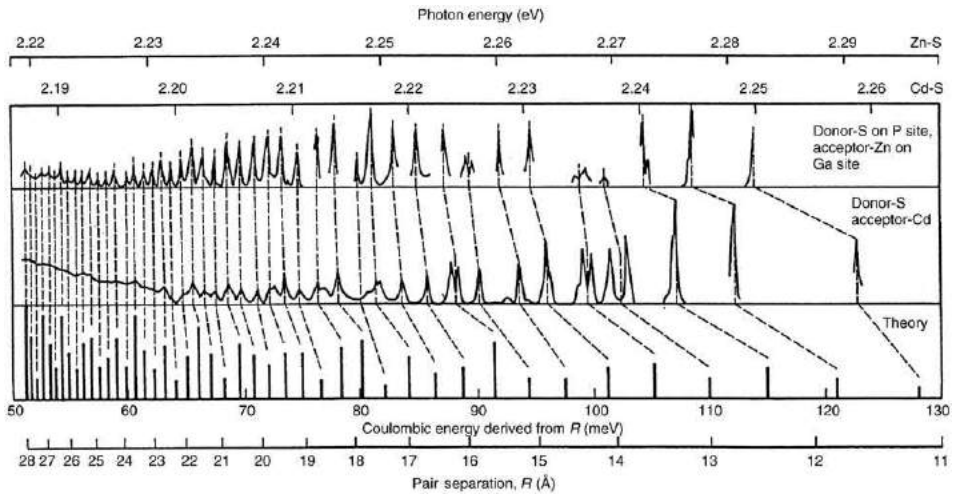

Fig. 2 Comparison of the positions and intensities of the sharp line spectra at 1.6 K corresponding to both ZnS and CdS acceptor–donor pairs with the predicted pair distribution. The lower scales show the pair separation (R) and the Coulombic energy derived from R. The emission energy scales for the two measured spectra are shown at the top. Reproduced with Gershenzon M, Logan RA, Nelson DF, *et al.* (1968) *Proceedings of the International Conference on Luminescence*. Budapest, Hungary: Akademia Kiada.

Quantum Efficiency

The quantum efficiency is the radiative recombination as a fraction of the total recombination. It can also be regarded as the average number of electrons produced per incident photon. Consider an intrinsic semiconductor, i.e., one in which the electron and hole concentrations are equal. Then with the notation of Fig. 1 the radiative band–band recombination rate is $B^s n^2$. The nonradiative rate is $(B_1 + B_2)n^3$. We shall add another nonradiative rate, An say, proportional to the injected carrier density $n \sim p$. The quantum efficiency is then

$$\eta = \frac{B^s n^2}{An + (B_1 + B_2)n^3 + B^s n^2} \tag{13}$$

Traps are here neglected, and one finds a maximum

$$\eta_{\max} = \left\{ 1 + \frac{2}{B^s} A^{1/2} (B_1 + B_2)^{1/2} \right\}^{-1} \tag{14}$$

For a high-quality epitaxial AlGaAs/GaAs double heterostructure of great purity we may take

$$\begin{aligned} &A \sim 0.5 \times 10^6 \ \text{s}^{-1}, B \sim 10^{-10} \ \text{cm}^3 \ \text{s}^{-1}, \\ &B_1 + B_2 \sim 10^{-29} \ \text{cm}^6 \ \text{s}^{-1} \end{aligned} \tag{15}$$

whence $n_1 \sim 2 \times 10^{17} \ \text{cm}^{-3}$, $\eta_{\max} \sim 0.96$.

Detailed Balance

It is clear that the radiative recombination rate from a material should be obtainable in terms of its optical 'constants', the absorption coefficient $\boldsymbol{\alpha}(\lambda)$ and the refractive index $\boldsymbol{\mu}(\lambda)$. Both are functions of the wavelength. This connection can be formalized by comparing the optical absorption process with the emission process and this can be done by using the principle of detailed balance. This equates the rate of disappearance by absorption of photons with the rate of production of photons by radiative recombination. The radiative recombination rate can thus be obtained as an integral over the optical functions. This relationship, pioneered by van Roosbroeck and Shockley in 1954, has been used a great deal because the optical quantities are often known with some accuracy.

This result corresponds to balancing the rate $B^s np$ and the rate Y^s in Fig. 1(a). Analogous detailed balance results are obtainable for the other pairs of processes in Fig. 1. Thus, the Auger recombination rate can be related to an integral over the impact ionization rate. However, this connection is less useful since impact ionization data are generally less well known than the optical absorption information.

Nonradiative Processes

Auger Effects

The Auger effect was discovered in 1925 in gases by Pierre Auger. An atom is ionized in an inner shell. An electron drops into the vacancy from a higher orbit and a second electron takes up the energy which is used to eject it from the atom. In solids the effect is roughly analogous. One of its characteristics is that it is not radiative. Since radiation is what is seen in many experiments, the Auger effect is suspected if and when there is less radiation than expected in the first place. It is rather harder to investigate than radiative transitions.

The effect has proved to be important as it limits the performance of semiconductor lasers, light-emitting diodes and solar cells, and it can be crucial in transistors and similar devices whose performance is governed by lifetimes. When heavy doping is required, as it is in the drive towards microminiaturization, its importance tends to increase, since the Auger recombination rate behaves roughly as $n^2 p$ or $p^2 n$ (see Fig. 1) compared with the radiative rate which behaves more like np. The inverse process is impact ionization, and is important in the photodiode, the impact avalanche transit time (IMPATT) diode, and hot electron devices.

Impact ionization can be regarded as an autocatalytic reaction of order one:

$$\text{e} \rightarrow 2\text{e} + \text{h} \quad \text{or} \quad \text{h} \rightarrow 2\text{h} + \text{e} \tag{16}$$

i.e., one extra particle is produced of the type present in the first place. This is a key feature for impact-induced nonequilibrium phase transitions in semiconductors.

In the theory one needs wavefunctions for four states of two electrons which are relevant after the many-electron problem has been reduced to a two-electron problem. In a matrix element calculation the four wavevectors imply a 12-fold integration in $\boldsymbol{k}$-space to cover all possible states of the two electrons. However, momentum and energy conservation usually reduce this to an eight-fold integration. Such a calculation is hard, for it requires (1) good wavefunctions and (2) accurate integrations.

Here we shall merely concentrate on the broad principles. In addition, the many-electron nature of the problem implies that another approximation is often inherent in the treatment apart from the use of perturbation theory. This is clear if one considers that the electron interactions are screened twice: once by an exponential screening factor and a second time by the dielectric constant. So there is some double counting, and the treatment of the effect as uncorrelated electronic transitions mediated by

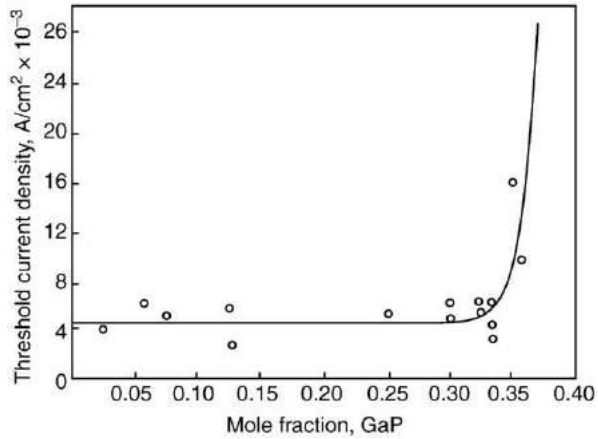

Fig. 3 Lowest values of laser threshold current density at 77 K as a function of mole fraction of GaP for $Ga(As_{1-x}P_x)$ junction lasers. Reproduced from Neill CJ, Stillman GE, Sirkis MD, *et al.* (1966) Gallium arsenide-phosphide: Crystal, diffusion and laser properties. *Solid-State Electronics* 9: 735–742, with permission from Elsevier.

screened Coulomb interactions is another approximation. The collective effects that enter require more sophisticated field theoretic methods which take care of electron–hole correlations, plasmon effects and the effect of free excitons, the so-called excitonic Auger effect. These calculations broadly confirm the results obtained by the simpler methods used for energy gaps large compared with the plasmon energies, provided the high-frequency dielectric constant is used and doping is not too heavy. Significant corrections are required for the narrow gap lead compounds, for example, and a considerable computational effort is required.

Let us now consider lifetimes for high carrier concentrations, as limited by band–band processes. They are best studied by looking at the emitted radiation. In this connection we recall the striking rise of the threshold current density J in a semiconductor laser as the material is changed from a direct material like GaAs to an indirect material, by mixing it with a compound such as GaP. **Fig. 3** shows this spectacular rise for $GaAs_{1-x}P_x$ at 77 K near $x = 0.38$. It is due to the drop in the radiative transition probabilities, as the substance becomes indirect, thus requiring a higher current density for threshold. So we should start with band–band processes in direct materials as most favorable for the experimentally accessible radiative transitions.

The essential point here is that the injected carrier density behaves as:

$$n_{\text{inj}} = \frac{J\tau}{qd} \sim \frac{(5 \times 10^3 \text{ A cm}^{-2})(10^{-9} \text{ s})}{(1.6 \times 10^{-19} \text{ C})(10^{-4} \text{cm})} \sim 10^{17} \text{ cm}^{-3} \tag{17}$$

where J/q is the particle current density at threshold, τ is the lifetime of the carriers, and d is the thickness of the active layer. As active layers are made thinner n_{inj} increases to 10^{19} cm^{-3} or so, far in excess of the defect concentration in the (undoped) material. This brings in band–band Auger effects. These have also been invoked to explain the undesirable increase in J with temperature. Thus the interest in direct band–band Auger effects in III–V compounds is fuelled by the need for better and smaller opto-electronic devices which work at long wavelengths (1.3–1.5 μm). It matches the interest in indirect band–band Auger and impurity Auger effects in silicon due to the importance of heavy doping in VLSI (very large scale integration) and transistor technology. **Fig. 4** shows a typical Auger process, called CHHS, as the conduction band (C) and the split-off band (S) are involved. Two relevant states are in the heavy hole band (H). State $2'$ is referred to as the Auger carrier, as it has more kinetic energy than the others.

Impact Ionization

The CHHS and CHCC processes and their inverses can be written as

$$e_C + 2h_H \leftrightarrow h_s \quad 2e_C + h_H \leftrightarrow e_C \tag{18}$$

where the suffix refers to the band. Viewed from left to right these are Auger processes. Viewed from right to left they are autocatalytic and impact ionizations.

Such processes are important for the impact-induced nonequilibrium phase transition in semiconductors. Also reaction rates with autocatalytic elements imply nonlinear equations in the concentrations and this gives rise to much interesting behavior as regards stability and bifurcation phenomena.

Momentum and energy conservation are very restrictive conditions on the four states involved in **(18)**, and it is easily seen that they cannot normally be satisfied if in a direct band semiconductor the recombining electron drops from the band minimum to the valence band maximum. This effect raises all the kinetic energies of possible processes. The Auger electron (on the right-hand sides of **(18)**) must also have a minimum energy in order that the impact-ionization process can proceed. This leads to an activation energy for the process. For **Fig. 4**, for example, and in the case of nondegeneracy, the energetic hole has a kinetic energy

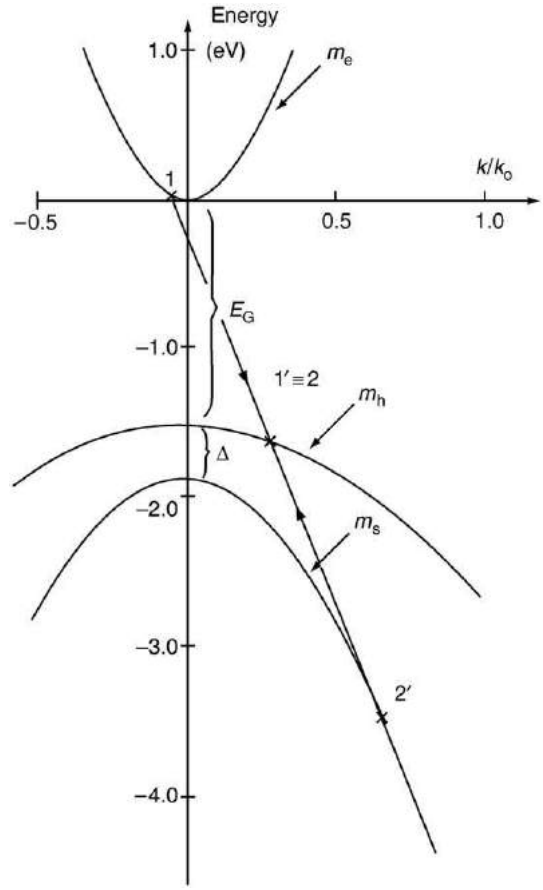

Fig. 4 Conduction band, heavy hole and split-off valence bands of GaAs, all treated as parabolic with $\hbar^2 k_0^2/2m_h \equiv E_G - \Delta$. *x* denotes a quadruplet of states for a most probable transition. The two states in the heavy hole band are not shown separately. The arrows indicate electron transitions. Reproduced from Neill CJ, Stillman GE, Sirkis MD, *et al.* (1966) Gallium arsenide-phosphide: Crystal, diffusion and laser properties. Solid-State Electronics 9: 735–742, with permission from Elsevier.

at threshold of

$$E_{th} = \frac{m_e + 2m_h}{m_e + 2m_h - m_s}(E_G - \Delta) \quad \text{(CHSS)} \tag{19}$$

As the band gap increases we see that E_{th} goes up and so the Auger rate for simple parabolic bands decreases. Values of E_{th} for other transitions can be obtained from **Eq. (19)**.

The total kinetic or threshold energy E_{th} can be converted to an activation energy by subtracting the basic energy which must under all conditions be part of the Auger particle energy. In the case of **Eq. (19)** one finds for $E_G > \Delta$:

$$\begin{aligned} E_a &= E_{th} - (E_G - \Delta) \\ &= \frac{m_s}{m_e + 2m_h - m_s}(E_G - \Delta) \end{aligned} \tag{20}$$

This can result in a strong temperature dependence in accordance with an Arrhenius factor $\exp(-E_{th}/kT)$. However, this is again lost, and the direct band–band Auger effect will certainly be important and only weakly temperature dependent, if $E_G \sim \Delta$. This can occur, for example, for InAs, GaSh and their solid solutions.

The connection between Auger processes and impact ionization has also been formalized. Of all Auger quadruplets of states in a set of nondegenerate bands, the most probable Auger transition involves the quadruplet, which yields the threshold for impact ionization. The three crosses in **Fig. 4** indicate such a quadruplet (the central cross represents two states).

Identification of Auger Effects

How does one know that an Auger effect has occurred? In pure but highly excited materials there is the original solid-state Auger effect which leads to a lifetime broadening of the electronic states at the band edge. For a semiconductor this effect causes fuzziness in the band edge which is a contributory factor to the overall bandgap shrinkage. The mechanism is illustrated in **Fig. 5**, which

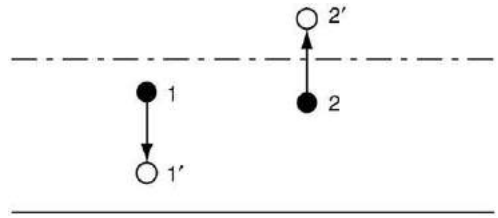

Fig. 5 The filling of a state k by the Auger effect. The arrows indicate electron transitions.

shows how the lifetime of a vacant electron state near the band edge is shortened by the Auger processes within this band. Under normal conditions this effect is small in a semiconductor. But under degenerate conditions the effect leads by lifetime broadening to a low-energy tail in emission.

The blanket term 'Auger effect' for all Coulombically excited two-particle transitions has been used here. In semiconductor device work the 'Auger effect' has become associated with transitions in which one electron bridges an energy gap. But there is no need to limit the concept in this way. Its competition with radiative effects means that it is often detrimental and a full discussion of means of suppressing it has recently been given by Pidgeon *et al*.

More spectacular and more convincing is the *detection* of the energetic Auger electron or hole. For this purpose one may look for the weak luminescence emitted when this carrier recombines radiatively. This has been done for the band–band process in Si and for the band–impurity process in GaAs. It leads to so-called $2E_G$-emission. Its rate-limiting step is the rate of the Auger process per unit volume, $B_1 n^2 p$, which populates the high-energy level. The radiative process then proceeds at a rate $B'_1 n^2 p^2$ per unit volume with $B'_1 \sim 10^{-54}$–10^{-51} cm^9 s^{-1}.

More usual is the identification of the band–band Auger recombination mechanism from the minority carrier lifetime τ, which behaves as

$$\frac{1}{\tau_p} = B_1 n^2 \quad \text{or} \quad \frac{1}{\tau_n} = B_2 P^2 \tag{21}$$

where B_1 and B_2 are Auger coefficients when the Auger particle is an electron or hole, respectively. A more complete expression allows also for trapping processes.

The departure from parabolic bands plays an important part for the energetic Auger particle (in state 2′). Theory shows that when this effect is taken into account, it tends to lower the band–band Auger coefficient. In fact in GaAs the CHCC process is ruled out altogether at 0 K for nonparabolic bands, suggesting that it should be an ideal material for radiative transitions. However, the possibility of phonon participation brings the effect back again, though it is still comparatively weak. The difficulty of conserving electron energy and momentum, which gives rise to the activation energies (**Eq. (20)**) is greatly alleviated if phonons can take up some of the momentum in a direct gap semiconductor.

We have seen that the activation energy barrier against the band–band Auger effect can be overcome by a suitable disposition ($E_G = \Delta$) of the three direct bands and by phonon participation. It can also be overcome if there is an indirect minimum near a Brillouin zone edge at about half the direct energy gap.

Acknowledgments

The author wishes to acknowledge support from NATO under their grant PST.CLG 975758.

Further Reading

Fraser, D.A., 1986. The Physics of Semiconductor Devices, fourth ed. Oxford, UK: Clarendon Press.
Hangleiter, A., 1985. Experimental proof of impurity Auger recombination in silicon. Physical Review Letter 55, 2976–2978.
Harrison, D., Abram, R.A., Brand, S., 1999. Characteristics of impact ionization rates in direct and indirect gap semiconductors. Journal of Applied Physics 85, 8186.
Landsberg, P.T., 1991. Recombination in Semiconductors. Cambridge: Cambridge University Press.
Nimtz, G., 1980. Recombination in narrow gap semiconductors. Physics Reports 63, 265.
Pidgeon, C.R., Ciesla, C.M., Murdin, B.N., 1997. Suppression of non-radiative processes in semiconductor mid-infrared emitters and detectors. Progress in Quantum Electronics 21, 361–419.

Coherent Terahertz Sources

L Wang, Chinese Academy of Sciences, Beijing, China

Introduction

Terahertz (THz) radiation is usually referred to as an electromagnetic wave with frequencies ranging from 1 to a few terahertz (1 THz = 10^{12} Hz). In the electromagnetic spectrum, terahertz radiation lies between the infrared and microwave, as shown in Fig. 1. The techniques for generation and detection of THz radiation bridge photonics and electronics in the sense of the concepts and the techniques discussed below.

Physical quantities corresponding to 1 THz are listed as follows;

- Frequency, 1 THz
- Wavelength, 300 μm
- Wavenumber, 33 cm^{-1}
- Energy, 4.1 meV
- Temperature, 48 K

In this spectral region, there exist many rich physical, chemical, and biological phenomena. For example, many phonon resonance and other elementary excitations in condensed matter fall in this region; many molecular vibrations and rotations occur at THz frequency; conformation-related collective vibrational modes in macromolecules, especially in bio-molecules such as proteins and DNA, also lie in THz frequency. However, the development of THz technique lags far behind that of photonics and electronics, due to lack of feasible, reliable, and economic coherent THz sources. The situation has been rapidly changing since the 1990s. Benefiting from advanced material science and the ultrafast laser techniques, various coherent THz sources are being developed and commercialization is soon to follow. We will introduce these coherent THz sources and their main features, and discuss their generation mechanisms. More details can be found in the listed books and articles in the Further Reading section at the end of this article.

Coherent THz Radiation

When a time varying polarization, $\mathbf{P}(\mathbf{r},t)$, is generated in a medium and oscillates at THz frequency, it will radiate a THz wave, $\mathbf{E}(\mathbf{r},t)$, according to Maxwell's equation:

$$\nabla \times \nabla \times \mathbf{E}(\mathbf{r},t) + \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\mathbf{E}(\mathbf{r},t) = -\mu_0 \frac{\partial^2}{\partial t^2}\mathbf{P}(\mathbf{r},t) \tag{1}$$

or

$$\nabla \times \nabla \times \mathbf{E}(\omega) - \frac{\omega^2}{c^2}\mathbf{E}(\omega) = \omega^2 \mu_0 \mathbf{P}(\omega) \tag{2}$$

in frequency domain, for a monochromatic wave or a Fourier component of a wave with a finite bandwidth. The properties of THz radiation are determined by the polarization source $\mathbf{P}(\mathbf{r},t)$. Incoherent THz sources exist in blackbody radiation, and are also detected in astronomy observations. If the phase evolution of the THz wave keeps in step, both spatially and temporally, it is called coherent radiation. A coherent THz source is more powerful and useful in spectroscopy, material characterization, imaging, and many other applications. It is clear that a coherent polarization source is first required for generation of a coherent THz wave. In practice, the working media for THz emitting can be free electrons, quasi-free electrons in solids, or bound electrons and other charge oscillations, such as plasmon oscillation, lattice vibration, etc. Classified by the excitation properties, the emitting process can be of resonance or nonresonance. Coherent polarization sources with different features can be created using various excitation techniques under different principles and device configurations, which lead to different coherent THz sources. Here we focus on coherent THz sources, primarily based on photonic techniques, and working in the spectral range from 1 THz to a few THz.

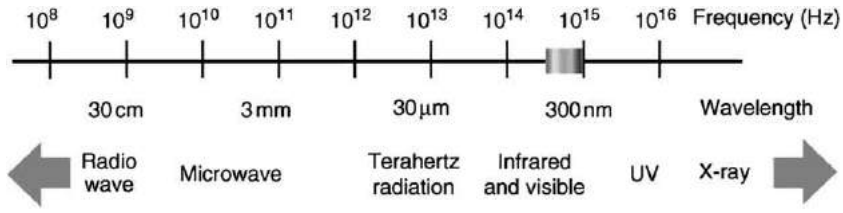

Fig. 1 The electromagnetic spectrum with THz radiation sitting between the microwave and the infrared.

 doi:10.1016/B978-0-12-809283-5.00857-0

Gas THz Laser

Molecules have many rotational and vibrational energy levels spaced at THz frequency. For polar molecules, the radiative transition between these levels can radiate THz wave. Following the standard laser technology, molecules pumped either optically or electronically can be used as laser gain media for coherent THz wave generation. For THz gas lasers, the very low THz photon energy poses a problem for lasing, because the coherent radiative transition can be disrupted by thermal depopulation and dephasing. An effective population inversion is difficult to set up and maintain. Therefore, it is important to carefully choose working medium and energy levels, as well as the way to excite the molecules. A number of media has been used for THz laser, such as methanol and HCN vapor, working in both continuous wave (CW) and pulsed modes. For example, methanol lasers operate under excitation of the rocking and asymmetric deformation modes of methanol excited by a CO_2 infrared laser. HCN laser is pumped electrically, which needs a large voltage and current to excite the molecules, a long cavity to obtain sufficient gain, and dedicated control of the temperature in the cavity for stable operation.

In general, THz gas lasers can be tuned discretely in thousands of lasing lines, ranging from a few tens to a few hundreds of micrometers, with power output from μW to mW. THz gas lasers have existed since the early 1970s, and are the only commercialized THz lasers to date. They are still subject to development, for less bulky volume, enhanced tunability and efficiency, and reduced costs.

Free Electron THz Laser

A free electron laser (FEL) uses free electrons as working medium, rather than bound atomic or molecular states in a conventional laser. A typical FEL consists of three parts: (i) an electron source that generates an electron beam with high current; (ii) an accelerator to raise the electron energy; and (iii) a lasing cavity. After the electron beam is generated, electrons are first accelerated to a relativistic velocity, typically with energy of hundreds of MeV, and then enter the cavity consisting of end mirrors and a spatially periodic magnets array called a wiggler (Fig. 2). Due to the Lorentzean force imposed by the periodic magnetic field, the high-energy electrons move in the cavity along a sinusoidal path, and emit coherent radiation with a wavelength determined by the spacing between magnets and the electron velocity, as well as the magnetic induction. With the feedback provided by the cavity mirrors, electrons are accelerated or decelerated continuously by the optical field, and are bunched via the resonant interaction. The collective motion of the electron bunches radiates powerful coherent synchrotron radiation. In a standard configuration, the wavelength of the radiation can be expressed as:

$$\lambda = \frac{\Lambda}{2\gamma^2}\left[1 + (k\Lambda B_0)^2/2\right] \tag{3}$$

where Λ is the spatial period of the wiggler magnets, γ the Lorentz factor of the electron beam, B_0 the peak magnetic induction, and k a constant. With a suitable arrangement of the magnet array, as well as other related factors, the facility can work in THz spectral region.

An FEL generates tunable, coherent, high-power THz radiation. It is the most powerful coherent source of THz radiation available. Up to 50 W of average THz power has been generated in the Jefferson Lab recently. The free electron THz laser is becoming an important tools for studying THz radiation and its interactions with condensed matter and biological materials. On the other hand, the FEL is a complicated and expensive facility, and needs dedicated maintenance.

Semiconductor THz Laser

In the twentieth century, semiconductor devices have achieved tremendous success, both in electronic and photonic regime, such as transistors and diode lasers. The semiconductor industry has resulted in revolutionary changes to our world and everyday life. To fill the THz gap between electronics and photonics, great efforts have been made to meet the demand in compact, economic, tunable, and highly efficient THz sources. Encouraging advances have been made since we stepped into the new millennium.

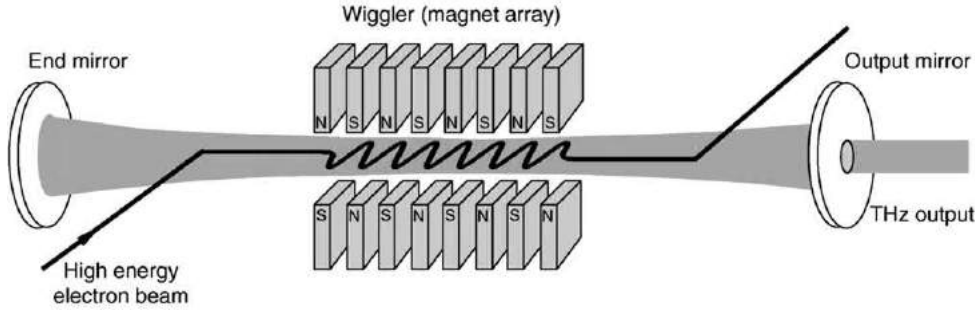

Fig. 2 A schematic of THz generation in a free electron laser.

In a semiconductor quantum well, the energy difference between a pair of sublevels can be artificially designed at THz frequency. As early as 1971, the concept of far infrared lasers based on intersub-band transition in semiconductor superlattices was proposed. Although the sublevels can be populated by electrical or optical pumping, the low photon energy at THz frequency causes serious difficulties for lasing, because the two levels sit so close to each other that many thermal processes could destroy the population inversion. Various designs and configurations have been proposed to tackle the problem. An attractive idea is so-called quantum cascade heterostructure. In this kind of device, the electrons are injected into a series of coupled quantum wells electrically. When the electrons are driven by the biased voltage, they can make radiative transition between a pair of sub-bands in a well. Subsequently, electrons enter the next well through resonant tunneling, and so on. However, this kind of device needs the materials of very high quality and suffers severely from fast depopulation of the excited states, even at very low temperatures.

As the advance in material science and unremitting efforts in pursuing the dedicated design of the device structures continued, the real breakthrough came in 2001, when Köhler *et al.* at Pisa demonstrated THz lasing in a quantum cascade heterostructure. Many improvements have been made since then; for example, the idea of using phonon resonance to selectively deplete the population of the lower sub-band has been successful, so that much stabler and robuster population inversion can be set up. The operation temperature has been raised to 136 K, well above the liquid nitrogen temperature, which opens the way for semiconductor THz lasers stepping into many more daily applications. These semiconductor devices are pumped with low voltages and currents at mA levels, and generate narrow bandwidth radiation with mW output at a frequency of a few THz. It is also probable that a four-level lasing system could even working at room temperature.

Another promising device is the p-germanium (Ge) laser developed recently. The Ge laser operates through the electrical excitation of hot holes in p-doped Ge. The laser cavity is formed by polishing the surfaces of the Ge crystal. These lasers could run at 5% duty cycles, with several mW output power. The output wavelength can be continuously tuned from 1 to 4 THz. At the moment, the devices have to operate at 20–30 K and consume about 10–20 W for refrigeration.

It is safe to predict that the feasibility, low cost, and compact semiconductor THz laser will emerge in the near future.

Beside the THz lasers, coherent THz radiation can also be generated by coherently exciting suitable media in other ways. Since the mid-1990s, the development of ultrafast laser techniques has led to two important pulsed coherent THz sources, i.e., photoconductive antenna and optical rectification THz sources. A unique feature of the pulsed coherent sources is broad bandwidth, which is useful in spectroscopic applications. **Fig. 3** is a typical setup for pulsed THz generation and spectroscopic study, using photoconductive antenna or optical rectification as THz sources.

THz Photoconductive Antenna (PCA)

Different from a conventional antenna working in radio and microwave frequency that are driven directly by electrical current, a photoconductive antenna for THz wave generation works with ultrafast pulsed lasers. When the photoconductive gap is irradiated by an ultrafast laser pulse, photocarriers are generated in the semiconductor. These photocarriers are driven by a DC voltage bias and form a current transient. The antenna couples the electromagnetic field associated with the time varying current into free space, and produces a THz pulse. The THz pulse usually has a duration of a few picoseconds, with a broad frequency bandwidth centered at THz frequency. The output THz power scales with both the optical pulse power and the DC bias field:

$$E_{\mathrm{THz}} \propto \frac{\partial J}{\partial t} \propto \frac{\partial^2 n_e}{\partial t^2} E_{\mathrm{bias}} \tag{4}$$

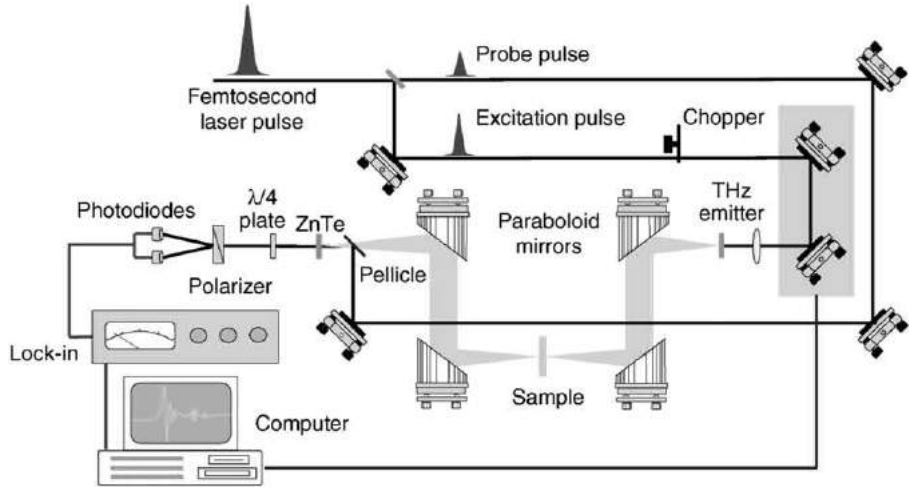

Fig. 3 A typical setup for broadband pulsed THz generation and spectroscopic study with a femtosecond laser as excitation source.

The output power, frequency bandwidth, and polarization of the THz radiation depend on the photoconductive features of the semiconductor layer and the geometric structure of the antenna. For an optimal performance, the substrate should have a suitable bandgap for optical interband excitation, a high electron mobility and short carrier lifetime to support an ultrafast current transient, and high dielectric breakdown threshold and to sustain high bias voltage. High-power PCAs have been demonstrated with semi-insulating GaAs, ion-implanted GaAs, and In-GaAs. In the aspects of antenna structures, coplanar strip lines and large aperture emitters are the most effective. Average output power from a good system can reach 30–40 μW.

A broad bandwidth is a special characteristic of a THz source, which is particularly desirable in the spectroscopic study as a coherent probe source. From the property of Fourier transform, we know that the shorter the THz pulse duration, the broader the spectral band. PCAs made of GaAs typically have a useful bandwidth extending from $\sim$100 GHz to 3 THz, which can be extended to 4 THz by tuning the pumping laser wavelength close to the bandedge. Up to 6 THz bandwidths have been reported for PCAs. The pulse duration and, therefore the achievable bandwidth, are ultimately limited by carrier mobility and TO phonon resonant absorption which is around 8.3 THz in GaAs.

In many applications, THz radiation is coupled to free space from the antenna using a closely attached silicon hemispherical lens. This practice increases the system's output and also provides control of the radiation pattern. The radiation pattern for the common dipole antenna is essentially dipolar, with a weak quadrupole component perpendicular to the bias field. In the far-field, the THz beam has a Gaussian cross-section with high-frequency components concentrated in the center.

As the first practical pulsed coherent THz source, PCA has been widely used for many applications in scientific research and material characterization. Combined with commercial ultrafast optical fiber lasers, full functional and self-contained THz spectroscopy systems with PCAs as emitters, have been demonstrated and are under commercialization.

Optical Rectification

Optical rectification is a second-order nonlinear optical effect originating from susceptibility $\chi^{(2)}$, and was first observed in 1962 using two intense monochromatic nanosecond laser pulses. In this process, the interaction between laser and nonlinear optical crystal produces a frequency difference polarization:

$$P(0) = \chi^{(2)} E(\omega_1) E(-\omega_1) \tag{5}$$

When a nonlinear optical crystal is irradiated by an ultrafast laser pulse, the different frequency components in the laser pulse will be coupled with each other by $\chi^{(2)}$, to induce frequency difference polarizations as:

$$P(\Omega) = \chi^{(2)}(\omega_1 - \omega_2 = \Omega) E(\omega_1) E(-\omega_2) \tag{6}$$

A femtosecond laser pulse usually has a spectral linewidth of at least a few nm, the induced polarization $P(\Omega)$ has a finite frequency distribution centered at THz frequency in the far field, and results in THz radiation $E_{\mathrm{THz}} \propto \partial^2 P(t)/\partial t^2$.

THz radiation in free space by optical rectification was obtained in the early 1990s. One of the differences from a PCA device is that the THz output by optical rectification scales with both the optical pulse power and the DC bias field, while the THz radiation generated by optical rectification solely results from the incident laser. The optimal performance depends on several factors. The generation efficiency first depends on the magnitude of the $\chi^{(2)}$ and phase-matching condition between the optical and THz pulses. The effective magnitude of the $\chi^{(2)}$ coefficient varies with the crystal cut and orientation. Up to now, the most popular material is ZnTe, for its physical durability and excellent phase matching, when a Ti:sapphire femtosecond laser is used as the excitation source. With a moderately focused optical pump beam and a ZnTe crystal, nW T-ray average power can be generated by optical rectification. In general, the generation efficiency may be increased by increasing the laser power, but there are two factors limiting the effect. First, the incident laser power cannot exceed the damage threshold. Second, other second-order nonlinear processes may compete with the optical rectification, such as second-harmonic generation of the pump beam at high optical flux.

For THz radiation generated by optical rectification, the ultimate bandwidth is primarily determined only by the linewidth of the pump laser pulse. Because it is a nonresonance process in most cases, optical rectification can reach broader bandwidths than PCA. Using ultrashort optical pulses and thin nonlinear crystal, THz pulses have been generated with spectra extending to the mid-infrared, well beyond the phonon band of the materials. Like the photoconductive antenna, optical rectification has become a popular technique to generate coherent broadband THz radiation in various applications. Beside the conventional bulk inorganic crystals, many different media have been used for THz generation, such as organic crystal DAST, biased quantum wells, periodically poled LiNbO3 (PPLN), poled polymers, super-conducting thin films, etc.

Other Coherent THz Sources

When two light sources with slightly different wavelengths interact together with a nonlinear optical crystal, a beating polarization will be generated and will radiate a THz wave. It is called difference frequency generation, or three-wave mixing. Optical parametric processing is another way to generate a THz wave. In this process, a near-infrared pump beam generates a second NIR idler beam in a nonlinear crystal, and THz radiation can be generated from the beating of the pump and the idler. The merit of this technique is the continuous frequency tunability of the THz output, which is valuable in spectroscopic applications, and relatively cheaper

nanosecond lasers can be used. If the nonlinear crystal is further placed in a cavity and the idler beam is amplified by feedback from end mirrors, a THz optical parametric oscillator is formed. A number of improvements have been made since the mid-1990s with this kind of device. More recently, THz parametric generation with an injection seeded idler beam has shown a reduced linewidth ($\Delta\nu/\nu \approx 10^{-4}$) and a peak THz power of over 100 mW for 3.4 ns pulses. A narrow linewidth combined with reasonable tunability make this kind of THz source attractive in spectroscopic studies.

Besides using photonic techniques, coherent THz radiation can also be generated by electronic techniques through increasing the output frequency of the microwave devices. For example, an electrically driven microwave generator, backward wave oscillator, can generate CW THz radiation up to 2 THz. A backward wave oscillator, running under optimal conditions, can provide up to 300 mW of polarized, narrowband radiation with a limited tunability about 30% of its central frequency.

See also: Strong-Field Terahertz Excitations in Semiconductors. Terahertz Physics of Semiconductor Heterostructures

Further Reading

Grüner, G., 1998. Vol. 74 of Topics in Applied Physics. Berlin: Springer-Verlag.
Kaiser, W., 1988. Vol. 60 of Topics in Applied Physics. Berlin: Springer-Verlag.
Khurgin, J.B., 1994. Optical rectification and terahertz emission in semiconductors excited above the bandgap. Journal of the Optical Society of America B 11, 2492–2501.
Rullière, C., 1988. Femtosecond Laser Pulses. Berlin: Springer-Verlag.
Saldin, E.L., Schneidmiller, E.A., Yurkov, M.V., 2000. Advanced Texts in Physics. Berlin: Springer-Verlag.
Siegel, P.H., 2002. Terahertz technology. IEEE Transactions on Microwave Theory and Techniques 50, 910–928.
Tredicucci, A., Gmachl, C., Capasso, F., *et al.*, 2001. Novel quantum cascade devices for long wavelength IR emission. Optics Materials 17, 211–217.
Tsen, K.T., 2001. Vol. 67 of Semiconductors and Semimetals. New York: Academic Press.

Using Ultrafast Optical Spectroscopy to Unravel the Properties of Correlated Electron Materials

Rohit P Prasankumar, Dmitry A Yarotski, and Antoinette J Taylor, Center for Integrated Nanotechnologies, Los Alamos, NM, United States

Published by Elsevier Ltd.

Introduction

Conventional metals, semiconductors, and insulators form the backbone of modern technology. This is possible because the properties of these materials are well understood and can largely be described in a single electron picture, where interactions between electrons do not play a significant role. For metals in particular, this is quite surprising, since one would think that their typical carrier densities of $\sim 10^{23}$ cm^{-3} (~ 1 electron/unit cell) would make it difficult to ignore the Coulomb interactions between electrons. Nevertheless, band theory can accurately predict the properties of these conventional materials in a quasi-free-electron picture, facilitating their use in a wide range of applications.

Despite the many technological successes achieved with these materials, researchers are continually searching for new materials with novel functionalities. Correlated electron materials (CEM) represent one of the most promising opportunities, since the electron–electron interactions governing their properties give rise to a host of unique phenomena, including magnetism, ferroelectricity, metal–insulator transitions (MIT), heavy fermion behavior, and superconductivity (SC). In addition, relatively small external perturbations (e.g., temperature, magnetic field, doping) can tune CEM between competing phases (e.g., SC and antiferromagnetic (AFM) order), further increasing their appeal. Intense interest in the fundamental properties of these materials, along with their great potential for applications, has thus driven decades of research into their properties.

The driving force for the unique properties of most CEMs is the competition between electron itinerancy and localization. For example, in conventional metals (with valence electrons in s or p orbitals), the electron hopping amplitude, t, is large, resulting in itinerant electrons with large kinetic energy that can travel freely throughout the material. However, when the potential due to Coulomb interactions between electrons (U) becomes comparable to or larger than t, electrons can be localized, even when band theory would predict them to be itinerant. This strongly influences the properties of CEM, which tend to have valence electrons in narrow bandwidth, low kinetic energy (<1 eV) d or f orbitals that are more susceptible to localization. The ratio between U and t is thus one of the first factors to consider when determining whether a material is strongly correlated.

The relatively low energies of electrons in CEM also make interactions with other degrees of freedom (DOF) (e.g., spin, charge, orbital and lattice) more significant, since they have comparable energy scales. This makes CEM extremely sensitive to small changes in external parameters, such as temperature, pressure, and magnetic field, due to these multiple competing energy scales; it also makes them remarkably flexible, since it is relatively easy to tune a given material between different ordered states.

A typical phase diagram for a representative CEM is shown in Fig. 1. This shows the different phases exhibited as a function of doping (x) and temperature (T) (although similar diagrams would appear when varying other parameters, such as magnetic field

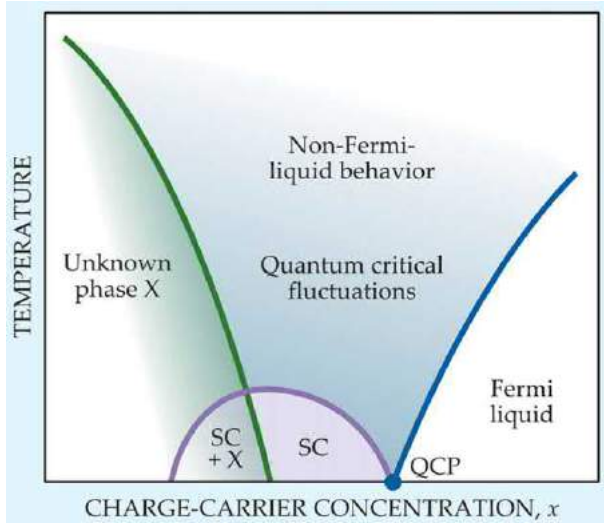

Fig. 1 Typical phase diagram for a hypothetical CEM. Reprinted with permission from Orenstein, J., 2012. Ultrafast spectroscopy of quantum materials. Physics Today 65, 44.

or pressure). It is clear from this schematic that small changes in T or x can result in transitions between different phases. In fact, the interplay between various DOF often leads to nanoscale phase separation, where different phases can simultaneously exist at a given temperature, although that is beyond the scope of this article.

Much insight into CEM can be obtained from optical spectroscopy (both time-integrated and time-resolved), which has several advantages over other experimental techniques for probing their properties. First, the energy scales relevant for CEM, ranging from meV (low energy excitations such as magnons and phonons) to eV (chemical bonds) are readily accessible using optics. In addition, optical techniques are non-contact and are therefore unlikely to introduce spurious effects in a given measurement. Furthermore, signatures of different ordered phases are present in the optical conductivity spectrum $\sigma(\omega)$; for example, the imaginary part ($\sigma_{\text{imag}}(\omega)$) in a superconductor has a $1/\omega$ dependence on the frequency (ω), while low energy excitations show up as peaks in the real part, $\sigma_{real}(\omega)$.

The addition of femtosecond (fs) temporal resolution to optical measurements offers additional benefits when probing CEM, particularly in unraveling the coexistence and interactions between different DOF. In many materials, various DOF relax on different timescales; for example, electrons typically lose their energy to phonons within $\sim$1 picosecond (ps) and the phonons subsequently interact with spins (in a magnetic material) on a timescale of tens of ps. Time-resolved measurements of these relaxation processes (typically using pump–probe techniques) can therefore be used to extract fundamental material parameters, such as the electron–phonon and spin–lattice coupling constants. More recently, ultrafast optical techniques including time-resolved second harmonic generation (SHG), Faraday/Kerr rotation, and X-ray diffraction have been used to *selectively* probe ferroelectric, magnetic, and structural dynamics, respectively. This is an important advance, because one can selectively photoexcite one DOF with a femtosecond optical pulse and probe the resulting transient response of another DOF to unravel the interactions between them, in a manner that simply cannot be done using conventional, static electronic, magnetic and structural characterization techniques.

In this article, our goal is to describe the ability of ultrafast optical spectroscopic techniques to provide deep insight into correlated electron materials that often cannot be attained using any other technique. We will begin with an overview of the essential properties and phenomena exhibited by various classes of CEM, underlining the many similarities in the phenomena exhibited by different materials. A brief description of the use of linear optical spectroscopy to study CEM will then be given, focusing on the Drude and Lorentz models as well as the various low energy excitations present in optical conductivity spectra.

This will set the stage for the main focus of the article, ultrafast optical spectroscopy (UOS) of CEM. We will begin that section by briefly describing the results from all-optical pump–probe spectroscopy, which was the first ultrafast optical technique used to study CEM in the early 1990s and is often still the first UOS technique used to study a new class of materials. We will then move on to ultrafast spectroscopy at mid-to-far-infrared (IR) frequencies, which has been particularly fruitful due to the rich spectrum of low energy excitations in this frequency range, and has recently garnered particular attention with the advent of intense terahertz (THz) and mid-IR pulses that can be used to drive these low energy excitations. UOS experiments at high frequencies (ultraviolet (UV) to X-rays) are somewhat less common, but have still provided much insight into magnetic and structural dynamics, and will thus be described in the following subsection.

The final section will give an overview of other ultrafast optical techniques that have been used to study CEM, going beyond conventional pump–probe techniques to directly examine different DOF. These include time-resolved SHG, scanning near-field optical microscopy, angle-resolved photoemission spectroscopy, and X-ray/electron diffraction. We will conclude with a future outlook, describing directions that we feel will be particularly promising as new UOS techniques and correlated materials are discovered.

We note that it is impossible for us to cover all of the exciting work done in this important field, and we apologize in advance to any of our colleagues whose work we have overlooked. We have focused here on a few specific classes of CEM with which we have direct experience; however, there are other interesting and important classes of CEM (e.g., charge/spin density wave materials), as well as other materials showing hallmarks of electronic correlations (e.g., carbon-based materials), where UOS techniques have made important contributions. Some of these areas are covered in the reading list at the end, which both overlap with this article and cover areas that we have neglected. Our hope here is that the reader will gain an appreciation of the many advantages of ultrafast optical techniques for studying CEM, and the deep insight they can provide into the unique properties and phenomena exhibited by these fascinating materials.

Classes of Correlated Electron Materials

Electron–electron correlations are important in many different classes of materials, ranging from simple metals like Fe and Pb (exhibiting ferromagnetic (FM) order and superconductivity, respectively) to exotic *f*-electron systems such as URu_2Si_2, which exhibits a variety of known and hidden phases. Here, we will focus on CEM whose properties are strongly influenced by multiple interacting degrees of freedom (unlike, e.g., simple ferromagnets) and thus generally display a rich and complex phase diagram similar to that in **Fig. 1**. We will begin this section by describing the properties of transition metal oxides (TMOs), arguably the most extensively studied class of CEM, and one that displays nearly all of the phenomena manifested in these materials (e.g., Mott physics, high-T_c superconductivity, metal–insulator phase transitions, etc.). This will be followed by a description of heavy fermion materials. In this section, our primary goal is to provide a basic understanding of the electronic, structural, and magnetic properties of CEMs, as this is critical for interpreting the results of ultrafast optical experiments. However, an in depth discussion of CEM properties is beyond the scope of this article, and we encourage the interested reader to peruse the reading list.

Transition Metal Oxides

Oxides are ubiquitous in our daily lives, as this class of compounds includes rust, sand, glass, precious gemstones such as ruby and sapphire, and the zinc oxide found in sunscreen. When transition metals are added to these materials, however, a multitude of novel and extremely interesting phenomena can appear, as mentioned above. Research on these materials has been ongoing since at least 1937, when Peierls and Mott pointed out that crystals with partially filled *d* orbitals (such as NiO) should be metallic based on band theory, but instead are insulators. We know now (in no small part due to the efforts of Peierls and Mott themselves) that electron–electron correlations are the origin of their insulating behavior. However, for a long time TMOs were neglected, partially due to the difficulty in fabricating them as compared to metals and semiconductors. This changed in 1986 upon the discovery of high-T_c superconductivity in layered copper oxides (where T_c is the critical temperature below which superconductivity exists), stimulating an enormous amount of effort in this area and leading to the discovery of additional effects, including colossal magnetoresistance (CMR), magnetoelectric coupling, and interfacial two-dimensional electron gases (2DEGs) in other TMOs.

The parent compounds of many TMOs (including the most well known examples, high-T_c superconductors (HTSC) and CMR manganites) are antiferromagnetic Mott insulators, with the chemical formulas ABO_3 or A_2BO_4 (where A is a rare earth ion and B is a transition metal (TM) ion). These materials possess an odd number of electrons/site and thus are expected to be metals from band theory, but Coulomb repulsion impedes electron transport, making them insulators (i.e., $U > t$ in the terminology used above).

This competition between electron itinerancy and localization in CEMs occurs at relatively small energy scales, as described in the previous section, amplifying the influence of spin, orbital, charge, and lattice degrees of freedom on their physics. In TMOs, this can readily be seen from the simple example of a transition metal ion placed in the cubic ABO_3 perovskite crystal structure (probably the most common TMO structure) (Fig. 2(a)). In this structure, the TM (e.g., Mn or Fe) occupies the corners of the unit cell. The cubic crystal field splits the 5-fold degenerate 3*d* energy levels of the TM ion, causing two e_g levels to go up in energy and three t_{2g} levels to go down in energy. This occurs because the TM ions are surrounded by oxygen octahedra; the e_g orbitals are

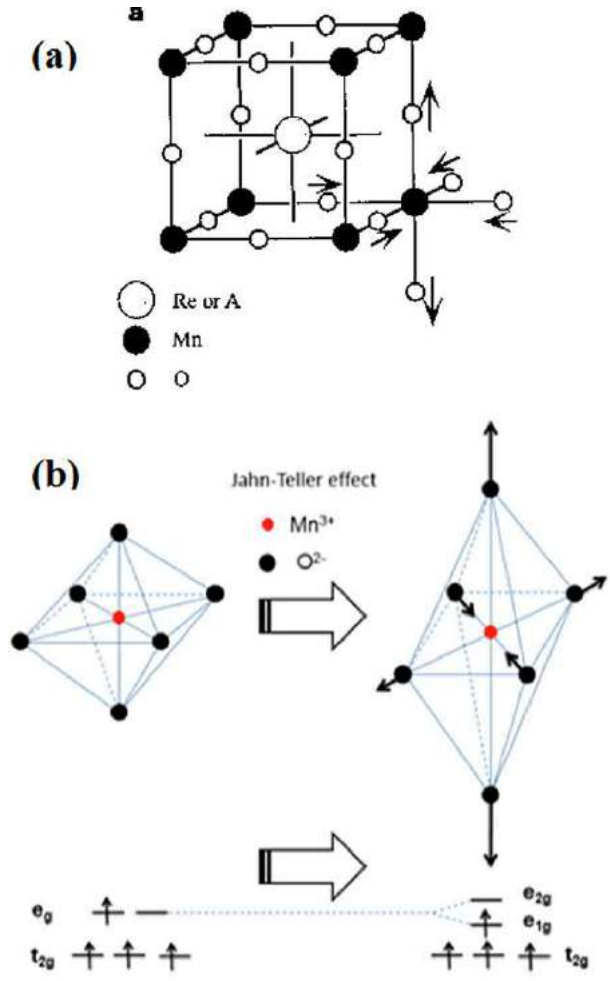

Fig. 2 Schematic of (a) the perovskite crystal structure (reprinted with permission from Millis, A.J., 1998. Nature 392, 147) and (b) the $Mn^{3+}O_6$ octahedron distortion and associated energy level splitting due to the JT effect.

directed towards the negatively charged oxygen ions and thus increase in energy (due to Coulomb repulsion), while the t_{2g} orbitals have lobes pointing in between the oxygen ions and decrease in energy. This illustrates the importance of the orbital DOF.

The lattice DOF also plays an important role, since the oxygen octahedra can elongate or compress along different axes to stabilize a given occupied orbital. If these distortions are strong enough, they can form a potential minimum that traps carriers, known as a lattice polaron. The most common distortion found in TMOs is due to the Jahn–Teller (JT) effect, where four oxygen ions move in one direction and two move in the other, creating a JT polaron that further splits the e_g energy levels (Fig. 2(b)). This has a significant influence on electron transport, as will be shown in the following sections. Finally, the spins of the localized electrons interact with one another through virtual hopping to their nearest neighbors (known as superexchange), leading to AFM spin order that typically alternates from site to site (or from plane to plane along a given crystal axis).

The most interesting physics occurs when TMO parent compounds are doped, typically by substituting a fraction of the rare earth A atoms with another alkaline ion that adds holes to the system. This tunes the carrier concentration (x) through the phase diagram of Fig. 1, initially suppressing AFM order and leading to SC or FM phases for $\sim 0.1 < x < \sim 0.4$ in the cuprates and manganites, respectively. A variety of other ordered phases can also emerge, often coexisting and competing with one another, as described in more detail in the following sections for specific materials. However, it is clear from this brief overview of TMO physics that the interplay between the charge, spin, lattice, and orbital DOF drives the distinctive phenomena observed in these materials.

High-T_c superconductors

The discovery of high-T_c superconducting cuprates revolutionized condensed matter physics and materials science, stimulating a worldwide effort to understand and optimize their properties that continues to this day. HTSC cuprates ($A_{2-x}B_xCuO_4$) have a layered tetragonal structure, with CuO_2 planes separated by (A,B)O planes. The CuO_2 planes are responsible for the superconducting behavior; the interleaved (A,B)O planes essentially serve to add electrons or more often, holes to the CuO_2 planes, which suppresses the AFM order of the parent A_2CuO_4 compound. The first HTSC cuprate to be discovered was $La_{2-x}Ba_xCuO_4$, with $T_c = 29$ K, followed within a year by $YBa_2Cu_3O_7$ (YBCO) ($T_c \sim 93$ K). The latter discovery was particularly significant since T_c was above the temperature of liquid nitrogen (77 K), making practical applications more feasible.

Two of the most hotly investigated topics after the discovery of HTSC cuprates included the symmetry of the pairing wave function and the mechanism underlying pairing in HTSC. In conventional superconductors, such as Pb, the paired electrons ("Cooper pairs") have an isotropic s-wave spin singlet wave function, resulting in a superconducting energy gap (Δ) proportional to T_c with equal magnitude throughout momentum (k) space. Pairing in these materials is mediated by phonons. In contrast, the Cooper pair wave function in HTSC cuprates was found to have d-wave symmetry, leading to an energy gap along certain directions in k space and gapless "nodes" along other directions. This, in combination with the complex phase diagram and proximity to AFM order, implied that phonons may not be responsible for electron pairing in the cuprates, and indeed, calculations show that the maximum T_c for phonon-mediated pairing is well below the experimentally observed T_c. Other possibilities for the pairing "glue" include AFM spin fluctuations or electron–electron correlations themselves. However, the nature of the pairing mechanism in HTSC cuprates remains one of the biggest unsolved problems in condensed matter physics to this day.

It is worth mentioning that in the last decade, other unconventional HTSC have also been discovered that are not TMOs. In 2008, iron-based superconductors (known as "pnictides"), with T_c up to 56 K, were first reported. This was quite surprising, since superconductivity was not believed to be possible in materials with strong magnetic ions like iron. These materials have attracted much attention because of their similarities and differences with the cuprates; i.e., they exhibit similar phase diagrams in which SC order emerges when doping an AFM parent compound. However, the parent compounds are metallic (in contrast with cuprates) and the pairing wave function has a specific type of s-wave symmetry. Despite the relatively recent discovery of these materials (along with the related iron chalcogenides, which have a T_c possibly as high as 109 K), the extensive experimental and theoretical expertise that has been developed from studying the cuprates for the past three decades has provided much insight into their physics. In fact, more is arguably known about their underlying physics than that of the cuprates; for example, the pairing mechanism, although still in question, is believed to be either due to spin/orbital fluctuations or electron correlations.

Finally, conventional (phonon-mediated) superconductivity was unexpectedly discovered in sulfur hydride (H_2S) at high pressures >160 GPa in 2015. Although the implications of this discovery are still being explored, it raises our hopes for finding a room-temperature superconductor at ambient pressure in the near future.

Vanadium dioxide and related materials

Vanadium dioxide (VO_2) is a canonical metal–insulator transition (MIT) material that has been studied using nearly every ultrafast optical technique, due to the longstanding controversy about the underlying origin of the MIT. At high temperatures, VO_2 is a metal with a rutile structure; upon cooling below $T_c = 340$ K, it becomes an insulator, with a concomitant change to a monoclinic structure that has a unit cell twice the size of that above T_c. The high temperature of the MIT has made VO_2 very attractive for potential applications. Importantly, the structural and metal–insulator transitions, though occurring at the same temperature, are not necessarily linked, and the question of whether the structural transition drives the electronic MIT or they occur independently has been the subject of much contention for decades.

Vanadium atoms in the high temperature rutile phase are arranged in a tetragonal lattice and surrounded by oxygen octahedra. Upon cooling below T_c, the vanadium atoms form dimers that tilt away from the c-axis, effectively doubling the unit cell; this structural change is known as the Peierls distortion, and has been suggested to directly lead to the insulating bandgap. However,

Mott believed that electron–electron correlations were essential to the MIT; as the system is cooled below T_c, the electron hopping parameter t decreases below the Coulomb interaction potential U, causing the system to become a Mott insulator.

Many researchers have used UOS techniques to address this problem, under the premise that the structural and electronic transitions could be disentangled in the time domain after ultrafast photoexcitation. However, to date the results have been mixed, as will be shown below, and the origin of the MIT is thus still an open question.

Magnetically ordered oxides

The discovery of colossal magnetoresistance in Mn-based oxides was perhaps the most significant offshoot of the intense interest devoted to HTSC cuprates. These materials, known as manganites (with the chemical formula $A_{1-x}B_xMnO_3$, where A and B are rare and alkaline earth ions, respectively), exhibit not only CMR, but also a variety of AFM, FM, orbital and charge-ordered phases depending on temperature and doping, with a phase diagram similar to that in **Fig. 1** and crystal structures similar to **Fig. 2**. The CMR effect occurs when a magnetic field is applied near the Curie temperature T_c, which separates the high temperature paramagnetic (PM) phase from the low temperature FM phase. With no applied field, the PM to FM transition is typically accompanied by an insulator-to-metal transition, although this depends on the specific material. However, if a magnetic field of a few Tesla (T) is applied near T_c, changes of $>1000\%$ in resistance can be observed (**Fig. 3**). This stimulated a substantial amount of effort aimed at both understanding the origin of this effect, as well as harnessing it for applications.

In the simplest picture, the CMR effect occurs when the magnetic field aligns the Mn spins between different sites. This influences the conduction of electrons in e_g levels, since their site-to-site-hopping depends on the alignment of their spins with that of the localized t_{2g} electrons; this is known as the double exchange mechanism. However, it was shown that double exchange cannot fully account for the dramatic decrease in resistivity observed in these materials; the strong electron–phonon coupling also plays a significant role by localizing carriers into polaronic states (particularly due to the JT effect described earlier). The interplay between these effects is believed to be the primary origin of CMR in the manganites, although nanoscale phase inhomogeneities are also thought to play an important role, perhaps even dominating the physics in certain parameter regimes.

The full spectrum of phases exhibited by the manganites, as with TMOs in general, depends on the ratio of U and t. As with many other TMOs, their parent compounds are AFM insulators, with the formula $AMnO_3$; doping at the A site gives the chemical formula $A_{1-x}B_xMnO_3$, which introduces holes into the system that can conduct (leading to CMR for $x \sim 0.15 - 0.4$). More specifically, manganites are often classified based on their electronic bandwidth, W (directly proportional to t), which depends on the Mn–O–Mn bond angle θ. This angle is determined by the ionic size of the A and B ions; the larger those ions are, the closer θ is to 180° (and the higher the bandwidth). Low bandwidth manganites, such as $Pr_{1-x}Ca_xMnO_3$ (PCMO), do not exhibit a metallic phase at any temperature without an applied magnetic field. Furthermore, their low bandwidth makes the interactions with other DOF more significant, leading to complex charge and orbital ordered (CO/OO) or magnetically ordered states for specific temperature and x ranges. In contrast, large bandwidth manganites, such as $La_{1-x}Sr_xMnO_3$ (LSMO), exhibit a FM metallic phase above room temperature and are much less sensitive to interactions with other DOF. Intermediate bandwidth manganites like $La_{1-x}Ca_xMnO_3$ (LCMO) exhibit the largest variety of phases and the greatest sensitivity to external parameters.

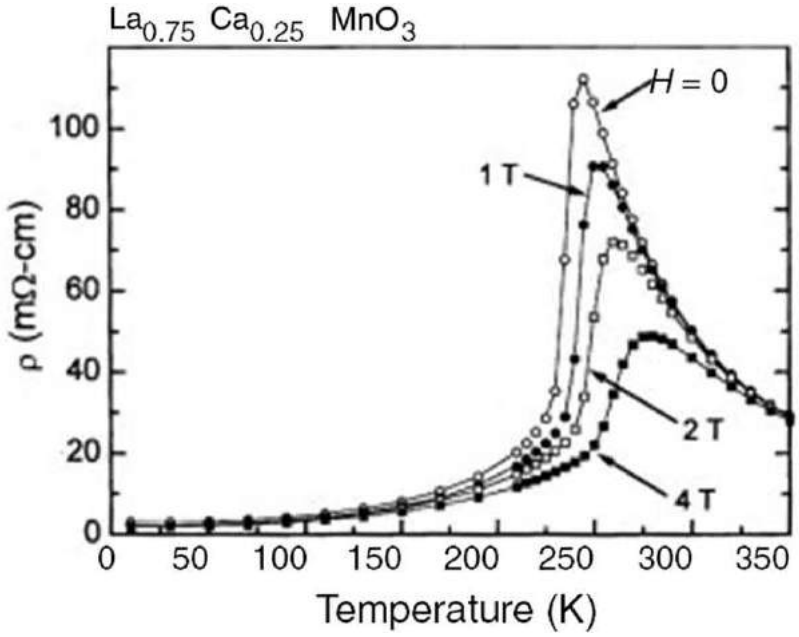

Fig. 3 Colossal magnetoresistance in LCMO. Reprinted with permission from Millis, A.J., 1998. Nature 392, 147.

Although the majority of manganites have a cubic perovskite structure, they can also adopt a layered structure (much like that of the cuprates), in which one or more conducting Mn oxide layers are separated by insulating layers. Layered manganites show many of the same complex magnetic and CO/OO phases as the cubic manganites, with the important distinction that the electronic and magnetic order is quasi-two-dimensional. This leads to a variety of interesting phases in which the system may exhibit magnetic and/or charge/orbital order in 2D and/or 3D.

Finally, although we have focused on the manganites here, it is important to note that other magnetically ordered oxides, such as the iron-based ferrites and the nickel-based nickelates, have similar structures (including layered phases) and display a similarly wide range of complex ordered phases originating from the interplay between charge, spin, lattice, and orbital DOF. These materials, particularly the ferrites, have been extensively studied in their own right using both conventional and ultrafast optical techniques; some examples will be given in the following sections.

Ferroelectrics

In analogy to magnetically ordered oxides, ferroelectric (FE) materials have a spontaneous electric polarization (P) below a critical temperature (T_c) that can be reversed by applying an external electric field (E). This property makes them useful for a wide variety of applications, including electronic memory, field-effect transistors, and capacitors, with a particular drive towards non-lead-based ferroelectrics in recent years. All ferroelectrics are non-centrosymmetric, piezoelectric, and pyroelectric in the FE phase. As with magnetic materials, FE materials exhibit domains (in which the orientation of P changes from domain to domain) and hysteresis (where P does not necessarily return to the same value as E is swept from positive to negative values and back). This enables much of the same theoretical machinery to be used to understand and model these materials.

Oxides with the perovskite structure (e.g., $BaTiO_3$ (BTO)) are perhaps the most common FE materials and will be our focus here. These oxides are a broad class of materials and thus exhibit different mechanisms for generating FE order. The most common mechanism is ionic displacement, where positive metallic ions (at the B site in the perovskite structure of **Fig. 2**) are displaced relative to the surrounding negatively charged oxygen octahedra (**Fig. 4(a)**); this occurs in the most well-known ferroelectrics, BTO and $Pb(ZrTi)O_3$ (PZT), resulting in an infrared-active soft mode phonon that can be observed in the optical response. Other mechanisms include charge ordering and octahedral tilting, although this is more rare. Finally, as will be discussed in the next section, magnetic order can also generate FE order in certain materials.

Magnetoelectric multiferroics

Magnetoelectric (ME) multiferroics, materials that simultaneously possess both magnetic and electric order, have garnered substantial interest since they can exhibit ME coupling between these two order parameters; notably, this could enable control of the FE polarization with a magnetic field or magnetism with an electric field. This would facilitate the exploration of novel physical phenomena in these materials as well as the creation of new multifunctional devices (e.g., multiferroic elements in which information is stored in either the magnetization or the polarization and could be written and read by both electric and magnetic fields). However, despite substantial effort in this area over the past decade, strong ME coupling in single-phase materials has only been realized at low temperatures, primarily due to a lack of knowledge regarding the underlying mechanisms. Obtaining a deep understanding of ME coupling in different multiferroics has thus been a focus of research in this burgeoning field.

Multiferroics are generally divided into two groups, depending on the nature of the coupling between magnetic and electric order. In type-I multiferroics, magnetic and electric orders originate from different mechanisms, and therefore they tend to be weakly coupled. However, both ferroelectricity and magnetism often appear well above room temperature, and the FE polarization can reach fairly high values ($\sim$10–100 C/cm^2). $BiFeO_3$ (BFO) is one of the most well known examples of this class of multiferroics, as FE order, appearing below $T_c = 1100$ K, originates from the ionic displacement mechanism discussed above (**Fig. 4(a)**) (and is further stabilized by the "lone pair" valence electrons of the Bi ion), while AFM order with a spiral spin structure arises below the Néel temperature of $T_N = 640$ K. Growth of BFO in thin film form causes a significant increase in the FE polarization, likely due to strain induced by the lattice mismatch between the film and substrate; this large polarization and room temperature coexistence of FE and AFM order has made BFO arguably the most heavily studied multiferroic material.

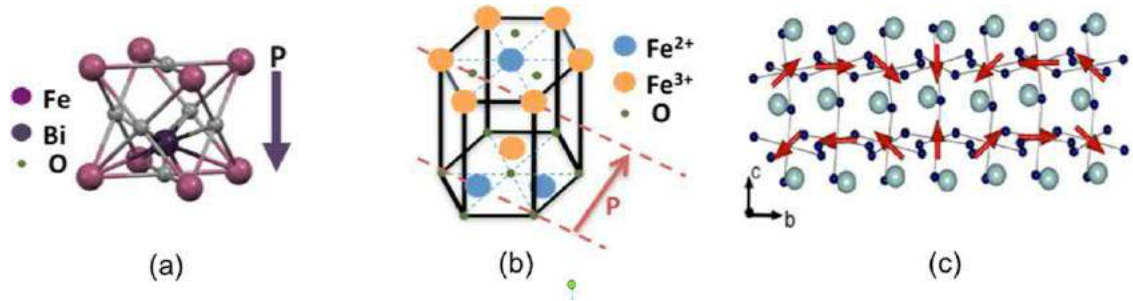

Fig. 4 (a) Structure of BFO, showing how the displacement of the Bi ion contributes to the FE polarization. (b) Charge ordering and FE polarization in $LuFe_2O_4$. (c) Spin spiral order in $TbMnO_3$. Reprinted from Kimura, T., 2007. Annual Review of Materials Research 37, 387.

Type-I multiferroicity can also occur through charge ordering; if a material contains transition metal ions with different valences, then FE order can emerge if the ions are ordered such that they can sustain a macroscopic polarization below the charge ordering temperature T_{CO}. $LuFe_2O_4$ (LFO) is perhaps the best known example of this class of multiferroics, with charge ordering below $T_{CO}=320$ K and AFM order below $T_N=240$ K that originates from its bilayered structure, in which charge ordering causes the FE polarization to appear in the direction between the two layers (Fig. 4(b)). However, it is worth noting that recent studies have indicated that $LuFe_2O_4$ may have nonpolar charge order, with no macroscopic FE polarization, calling the multiferroic nature of this material into question.

In contrast, in type-II multiferroics, ferroelectricity originates from magnetic order (through the Dzyaloshinskii–Moriya interaction), and therefore the two parameters are strongly coupled. However, the FE polarization is typically smaller than in type-I multiferroics, and strong ME coupling only occurs at low temperatures. The canonical example of a type-II multiferroic is $TbMnO_3$, which is AFM ordered below $T_N=41$ K, with ferroelectricity emerging below $T_c=28$ K. In this material, Mn moments orient along the crystallographic b axis in a sinusoidal pattern at T_N and in a cycloidal spiral pattern at T_c (where the spins rotate around the a axis) (Fig. 4(c)). Furthermore, application of a strong magnetic field along the b axis can cause the FE polarization to "flop" from the c axis to the a axis, demonstrating the strong link between the FE polarization and spiral magnetic order in $TbMnO_3$.

Finally, multiferroics display a number of low energy excitations that appear in the optical response, including IR-active phonons and magnons. Perhaps the most interesting excitations are electromagnons, magnetic excitations driven by the electric field of light that typically appear at low temperatures when magnetism and ferroelectricity are strongly coupled. These low energy ($\sim$1–10 meV) excitations were initially observed in $TbMnO_3$ and thus were thought to be intimately related to the mechanism underlying ME coupling. Electromagnons have since been observed in other materials, such as BFO and the non-multiferroic material $Ba_2Mg_2Fe_{12}O_{22}$. This last demonstration significantly expands their prospective impact, as one could potentially use light to control magnetic and/or FE order in any material supporting EMs (not only multiferroics), making them the subject of much interest in recent years.

Oxide heterostructures

Despite the many novel phenomena displayed by CEM in general and TMOs in particular, their discovery has often relied on serendipity, and the inability to fabricate crystalline TMO films with quality comparable to that of semiconductors has limited their applicability. However, this is rapidly changing, as TMO heterostructures can now be fabricated with a level of control and purity that is approaching that of semiconductor heterostructures. In these hybrid structures, two or more materials with a particular set of competing orders are interfaced in 1D, 2D and more recently, 3D patterns in order to tease apart the multiple DOF involved in the collective behavior that produces material functionality. Moreover, the interfaces often exhibit emergent properties that cannot be obtained in the individual constituents. For example, FM states in superlattices composed of AFM insulators have been observed, as well as extremely high charge mobility and superconductivity at the interface between insulating compounds. Interfaces also enable or enhance coupling between different order parameters that could not be achieved in single-phase materials, for example, strain-mediated ME coupling in multiferroic nanocomposites of magnetostrictive and electrostrictive materials. This has made the development and characterization of TMO heterostructures an extremely active field in current materials science and condensed matter physics.

One of the first discoveries in this field was that of a high mobility, superconducting 2DEG at the interface between insulating $LaAlO_3$ and $SrTiO_3$. This stimulated a substantial amount of effort aimed at understanding the origin of this phenomenon and developing heterostructures composed of materials with different functionalities. In this vein, an important focus has been the development of multiferroic heterostructures, where ME coupling can be engineered by proper choice of the interface geometry and constituent materials. For example, a magnetostrictive/electrostrictive interface enables tuning the magnetization of FM $CoFe_2O_4$ through the strain generated by an electric field in FE BTO. Alternatively, several groups have grown heterostructures consisting of multiferroic BFO and a FM material such as LSMO or CoFe (Fig. 5(a)), although the microscopic mechanism governing the coupling across the BFO/FM interface was not unambiguously identified. Finally, heterostructures composed of superconducting (e.g., YBCO) and FM or AFM (e.g., various manganites) layers have displayed novel interfacial phenomena, including orbital reconstruction and suppressed superconductivity.

More recently, researchers have gone beyond TMO bilayers to develop novel superlattice and vertically aligned 3D heterostructures that offer additional control over functionality (Fig. 5(b)). As will be seen in Section "Time-Resolved Optical Spectroscopy of Correlated Electron Materials," ultrafast optical spectroscopy is an essential tool for disentangling the interactions between the different DOF in these heterostructures through selectively photoexciting and probing their response in the time domain.

Heavy Fermion Materials

Heavy fermion (HF) materials display many of the phases encountered in other correlated electron materials, including SC, FM, and AFM order, and are thus an excellent testing ground for exploring these phenomena. In HF systems, a coherent heavy electron state forms below a characteristic temperature T^*, due to hybridization of the conduction electrons with localized f-electrons. This gives rise to a coherent "Kondo liquid" state in which the electron mass is enhanced and the resistivity drastically reduced. A wide spectrum of states, such as magnetism and superconductivity, can then emerge at low temperatures, depending on the interactions between the coherent heavy electrons and spins related to the localized f-electrons. These materials thus possess a complex phase diagram (Fig. 6), with phases that are both common to other CEM and unique to HF materials.

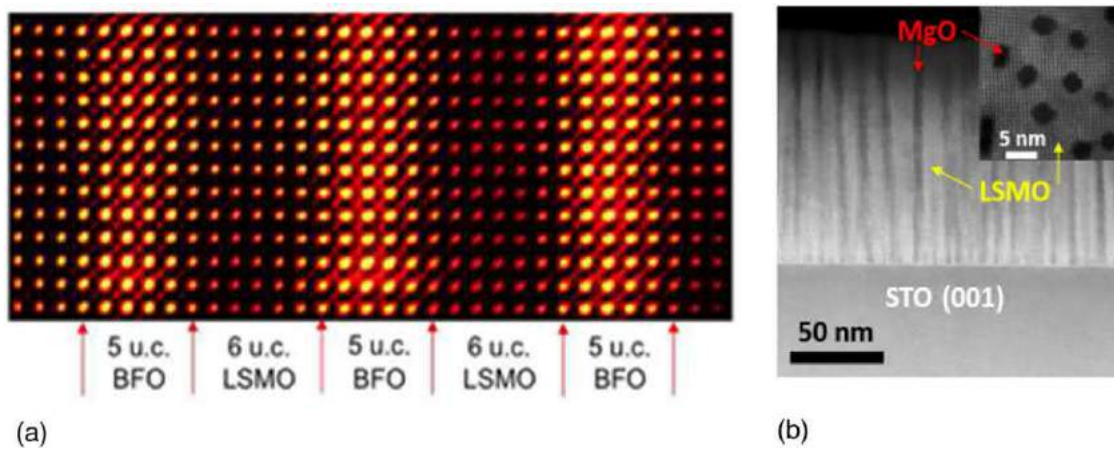

Fig. 5 Atomically sharp interfaces in (a) 1D superlattices (reprinted with permission from Singh, S., Haraldsen, J.T., Xiong J., e*t al.*, 2014. Physical Review Letters 113, 047204 (supplementary info)) and (b) 2D TMO heterostructures used to control meso-to-macroscale functionality (reprinted with permission from Chen, A., Hu, J.M., Lu, P., *et al.*, 2016. Science Advances 2, e1600245).

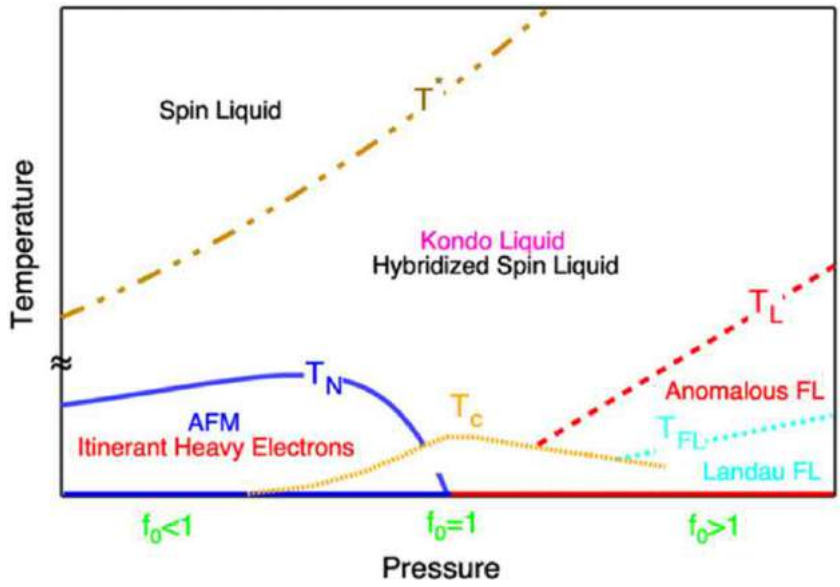

Fig. 6 Generic phase diagram for heavy fermion materials. Reprinted with permission from Yang, Y.-F., Pines, D., 2012. Proceedings of the National Academy of Sciences 109, E3060.

For most purposes, these materials may be treated as being primarily governed by the interactions between localized and itinerant electrons, making them a clean system for exploring strong correlations. However, the microscopic factors that govern the hybridization strength, V, and how it determines the resulting ground state are not well known. Optical measurements can shed light on this, since hybridization causes a gap (Δ) to appear in the electronic structure. This is manifested in the optical conductivity as a Drude peak centered at zero frequency, due to the reduction of scattering in the coherent state, as well as an IR peak due to transitions across the gap. Both time-integrated and time-resolved optical spectroscopy have thus been important in providing new insight into heavy fermion materials, as will be shown below, although the application of advanced UOS techniques (beyond all-optical pump–probe spectroscopy) to these systems is still relatively unexplored.

Time-Integrated Optical Spectroscopy of Correlated Electron Materials

Optical spectroscopy has been an invaluable tool for understanding the properties of correlated electron materials. Many of the extraordinary phenomena displayed by these materials exhibit unique signatures in the optical response, particularly at lower infrared photon energies ($\hbar\omega < \sim 1$ eV, where ω is the frequency). For example, optically active phonons and magnons have resonant peaks at mid-to-far-IR frequencies that can be used to probe their intrinsic properties. Similarly, metals exhibit a

characteristic increase in the conductivity at low frequencies (described by the Drude model), and semiconductors and insulators have a gap in their absorption at photon energies ranging from $\sim$0.3 to 10 eV that leads to interband transitions in the optical response. Superconductors and heavy fermion materials also exhibit gaps at much smaller ($<$100 meV) energies. Importantly, most optical techniques do not require contacts to the sample, making them relatively immune to the experimental artifacts that can affect data obtained by other characterization techniques.

In this section, we will give an overview of the optical response of CEMs. We will begin by describing the information contained within a generic optical conductivity spectrum for a CEM. We will then briefly describe the Drude-Lorentz model, arguably the most common starting point when fitting these spectra. It is worth noting that we will focus on information obtained from Fourier transform infrared (FTIR) and ultraviolet-visible (UV–visible) spectroscopy, and will not discuss other time-integrated optical techniques, such as Raman, photoluminescence (PL), and photoconductivity spectroscopy, although those techniques have often provided valuable insight into CEM as well. The discussion in this section will set the stage for understanding the results from UOS experiments discussed in the remainder of this article; as before, this is a general overview, and more detail is given in the reading list.

General Features of the Optical Response in Correlated Electron Materials

Fig. 7 displays a generic optical conductivity spectrum for correlated electron materials. Most CEM will exhibit some, if not all, of the features in this spectrum, making it worth a general description. At low photon energies (far-IR), materials that are metallic or semi-metallic display a Drude response, i.e., an increasing real part of the optical conductivity with decreasing frequency; this will be quantitatively described in the next section. Increasing the photon energy reveals low energy optically active magnon, phonon, and electromagnon excitations, typically modeled with a Lorentzian response (also described in the next section). These quasiparticle excitations are often obscured by the Drude response in metals, but can usually be seen in semiconductors and insulators. At still higher photon energies, interband transitions between occupied and unoccupied bands appear.

The optical conductivity can be measured as a function of various parameters, including temperature, pressure, and magnetic field. This enables researchers to track the evolution of CEM properties as a function of these parameters, giving insight into their properties. For example, tracking the Drude response as a function of temperature can indicate the appearance of a metal-insulator transition. Similarly, magnon excitations can appear when a system goes from paramagnetic to FM or AFM order, and soft mode phonons can emerge upon the onset of FE order. These measurements are also sensitive to the onset of polarons in, for example, CMR manganites, manifested through suppression of the Drude response at low frequencies and enhancement of the mid-IR resonant polaron absorption (i.e., spectral weight transfer). This further emphasizes that optical techniques are an important, non-contact probe of CEMs that can often give quantitative insight into their properties.

The Drude–Lorentz Model

The complex optical conductivity is given by $\sigma(\omega)=\sigma_1(\omega)+i\sigma_2(\omega)$. Experimentally, this is measured using FTIR spectroscopy (typically from low to intermediate frequencies ($\sim$0.01–5 eV)) and UV–visible spectroscopy at higher frequencies ($\sim$0.1–10 eV). For metallic systems (or any system with non-interacting carriers in a partially filled parabolic band), the Drude model is typically

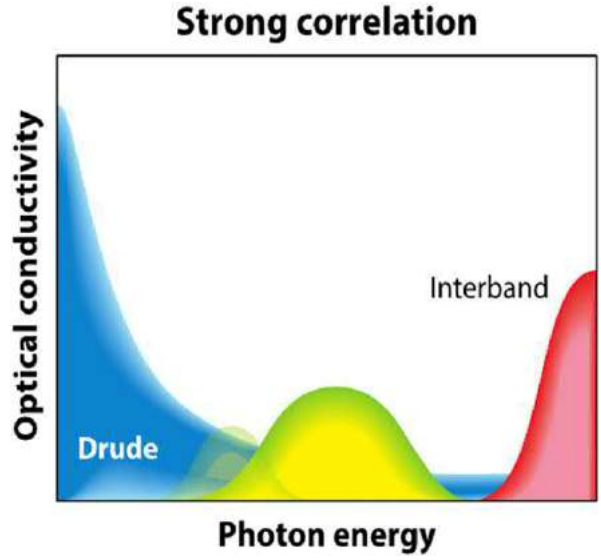

Fig. 7 Generic optical conductivity spectrum for CEM. Reprinted with permission from Zhang, J., Averitt, R.D., 2014. Annual Review of Materials Research 44, 19.

the first model used to fit the low frequency optical conductivity. This is obtained by classically considering a free charge driven by an electric field in a system with a scattering time, τ, between carrier–carrier collisions. At zero frequency, the DC conductivity is given by $\sigma_0 = \frac{Ne^2\tau}{m}$, where N is the carrier density and m is their mass. This can be generalized to finite frequencies:

$$\sigma(\omega) = \frac{\sigma_0}{1 - i\omega\tau}$$

This model yields a dependence of the optical conductivity on frequency that resembles the low frequency portion of Fig. 7. Although this is a classical model, it often provides a good approximation to the conductivity for a wide range of metallic systems. Furthermore, the Drude conductivity is often expressed in terms of the plasma frequency, $\omega_p = \sqrt{\frac{4\pi Ne^2}{m}}$:

$$\sigma(\omega) = \frac{\omega_p^2}{4\pi}\frac{\tau}{1 - i\omega\tau}$$

The Drude response is thus completely determined by the two frequencies ω_p and $1/\tau$, and by considering the ratio between these quantities one can often quickly get insight into material properties.

At higher frequencies, quasiparticle excitations (such as phonons and magnons) and interband transitions can be classically fit using a Lorentzian oscillator with a frequency ω_0. Combining this with the Drude model gives the Drude–Lorentz conductivity,

$$\sigma(\omega) = \frac{Ne^2}{m}\frac{\omega}{i(\omega_0^2 - \omega^2) + \omega/\tau}$$

Again, although this is a classical model, it is surprisingly useful in modeling the response of a broad range of materials, particularly when electron–electron interactions are not significant (and even in some cases where they are, as will be shown below). Finally, it is worth noting that multiple oscillators can be used in situations when there are several quasiparticle excitations contained within the optical conductivity spectrum, making this approach quite general for fitting optical conductivity spectra at low to intermediate photon energies.

Time-Resolved Optical Spectroscopy of Correlated Electron Materials

Ultrafast optical spectroscopy provides the ability to temporally resolve phenomena at the fundamental timescales of atomic and electronic motion. From a materials science perspective, 10–100 fs temporal resolution combined with spectral selectivity enables detailed studies of electronic, spin, and lattice dynamics, and crucially, the coupling between these degrees of freedom, an especially important aspect for correlated electron materials. In this sense, UOS complements both "static" measurement techniques, such as time-integrated optical spectroscopy, and dynamic techniques, such as inelastic neutron scattering. Furthermore, UOS offers exciting possibilities to investigate non-equilibrium phenomena, such as selective photo-doping, to create a metastable state that would not, for a given material, be thermally accessible. It is now possible to routinely generate and detect 10–100 fs pulses in the visible, mid-, and far-IR portions of the electromagnetic spectrum, enabling novel time-resolved spectroscopic investigations. Further, the development of time-gated detection techniques (especially at THz frequencies) has enabled direct measurement of the electric field, which permits the determination of the complex conductivity $\sigma(\omega)$ over the spectral bandwidth of the probe pulse. In the context of CEM, spectral selectivity from approximately 0.001 to 4.0 eV is especially important since many relevant quasiparticle excitations lie in this range, including electron–phonon coupling, charge ordering, hybridization phenomena, phonon and polaron dynamics, and condensate relaxation and recovery.

Fig. 8(a) depicts, in general terms, how this is accomplished. An incident pump pulse induces a change in the optical properties of a sample on a 10–100-fs time scale. Subsequently, this perturbation is probed with a second ultrashort pulse, which can, depending on the wavelength and experimental setup, measure the pump-induced change in the reflectivity, transmission, conductivity, or polarization. By changing the relative delay of the pump and probe pulses it is possible to temporally measure the subsequent relaxation dynamics. The ensuing photoinduced changes can be described in terms of the complex dielectric function ($\varepsilon(\omega) = \varepsilon_1 + \varepsilon_2$) of the sample. For example, the pump induced change in reflectivity (R), typically denoted as $\Delta R/R$, can be related to ε as follows:

$$\frac{\Delta R}{R} = \frac{\partial(\ln R)}{\partial\varepsilon_1}\frac{\partial\varepsilon_1}{\partial\kappa}\Delta\kappa + \frac{\partial(\ln R)}{\partial\varepsilon_2}\frac{\partial\varepsilon_2}{\partial\kappa}\Delta\kappa$$

where κ is a parameter that is modulated by the pump pulse and could be, for example, the temperature (T) or carrier density (N). Thus, the dynamics can be related to the complex permittivity (or equivalently the complex conductivity), which in turn is related to specific microscopic models. There are, of course, variations and more advanced approaches, but the ultimate goal is to relate dynamic changes in the dielectric function to specific microscopic phenomena. Approaches for modeling these dynamics range from simple rate equations, to density matrices, to Boltzmann transport equations.

Most often, the pump pulse creates an initially non-thermal electron distribution (Fig. 8(b) $1 \rightarrow 2$) fast enough that (to first order) there is no coupling to other degrees of freedom. During the first $\sim$100 fs, the non-thermal distribution relaxes primarily by electron-electron scattering (Fig. 8(b), $2 \rightarrow 3$). Coherence effects due to the impulsive nature of the photoexcitation can also lead to a phase-coherent collective response, resulting in coherent magnons or phonons. Subsequently, the hot Fermi-Dirac distribution thermalizes through coupling to the other DOF ($3 \rightarrow 1$) through, for example, electron-phonon coupling, phonon or magnon

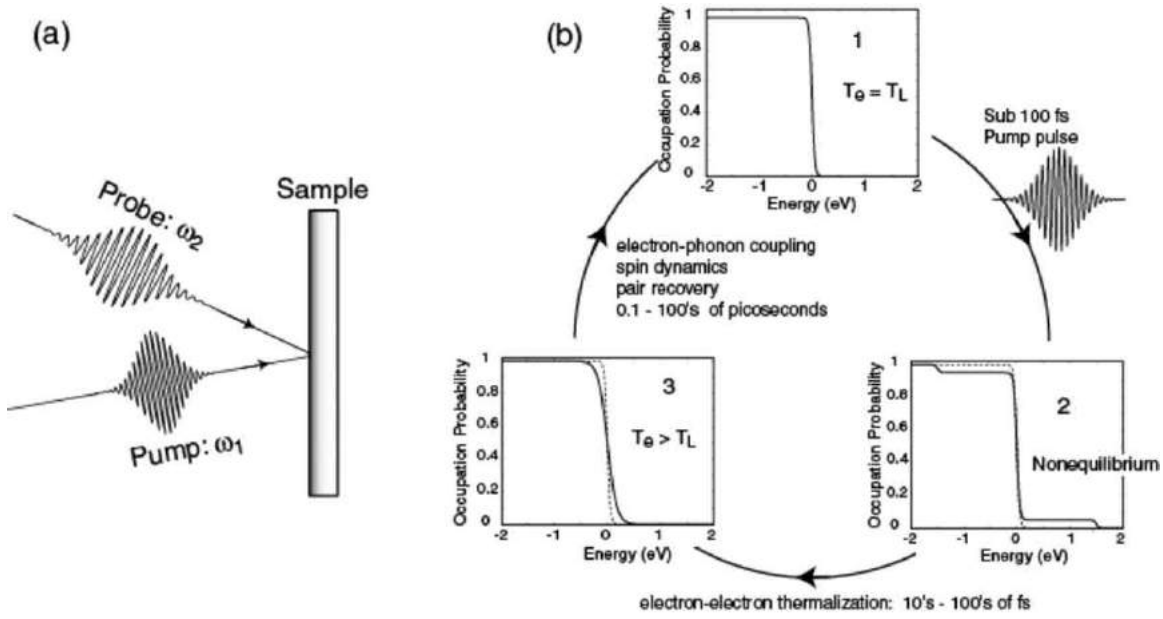

Fig. 8 (a) General pump–probe experimental setup. (b) Simple description of a pump–probe experiment in terms of changes in the occupation probability. 1. All DOF are initially in equilibrium. 2. Following ultrashort pulsed excitation, a non-equilibrium electron distribution is created which, primarily through electron-electron-scattering, relaxes to a Fermi–Dirac distribution, albeit at an elevated temperature in comparison to other DOF. 3. The quasiparticles subsequently relax through various (potentially competing) pathways in CEM, including electron–phonon relaxation, recombination/pair-recovery, or spin wave emission. Reprinted with permission from Averitt, R.D., Taylor, A.J., 2002. Journal of Physics: Condensed Matter 14, R1357.

emission, or recombination/pair-recovery. In some variations of UOS, a probe pulse is not required, since the pump pulse creates a nonlinear polarization in the material, leading to emission of radiation at THz frequencies that can be subsequently detected.

All-Optical Pump–Probe Spectroscopy of Correlated Electron Materials

Illustrating the dynamics in Fig. 8(b), electron phonon (e-ph) relaxation in conventional metals can be revealed through single color optical-pump, optical-probe (OPOP) experiments. Early experiments in noble metals, such as gold or silver, showed that relaxation occurs on a timescale of ~ 1 ps and is relatively temperature independent, as described by the two-temperature model. To explain these experiments, P.B. Allen proposed that electron energy relaxation occurs through their coupling to the lattice bath, connecting the measured relaxation time to the dimensionless e-ph coupling parameter λ, an important parameter in the BCS theory of superconductivity, with $\lambda \sim 1$ indicative of a potential superconductor. Early OPOP experiments at room temperature thus enabled the determination of λ, revealing $\lambda = 0.08$, 0.13 and 0.13 for the non-superconducting metals Cu, Au and Cr, respectively, while for the superconducting metals Nb, Pb, and NbN $\lambda = 1.16$, 1.45, and 0.95, respectively.

Following these measurements of e-ph relaxation and coupling in normal metals, OPOP experiments were used to characterize quasiparticle recombination in cuprate superconducting single crystals, with a view towards understanding pairing mechanisms in these unconventional superconductors. A systematic study performed on $Y_{1-x}Ca_xBa_2Cu_3O_{7-\delta}$ as a function of temperature and doping using OPOP spectroscopy at 1.5 eV revealed i) a temperature-dependent slow (2–6 ps) relaxation process, corresponding to superconducting pair recovery across the superconducting gap, with pair breaking by phonons increasing the recovery time and ii) a temperature-independent fast (0.5 ps) component associated with the pseudogap. Two-color pump–probe measurements with a mid-IR probe (1.5 eV pump, 0.06–0.2 eV probe) on $YBa_2Cu_3O_{7-\delta}$ (YBCO) thin films, in addition to confirming this behavior, observed that the amplitude of the signal followed the temperature dependence of the antiferromagnetic 41 meV peak observed in neutron scattering, supporting coupling between charge and spin excitations in these materials.

Detailed OPOP measurements (as well as optical-pump, THz-probe measurements, to be discussed in the next section) were also performed as a function of doping on films and single crystals of $Y_{1-x}Ba_2Cu_3O_{7-x}$ and $Bi_2Sr_2Ca_{1-y}Dy_yCu_2O_{8+\delta}$, revealing the importance of bimolecular recombination kinetics, consistent with the interaction of opposite spin quasiparticles as they recombine into Cooper pairs. A sharp transition in the quasiparticle dynamics was observed at optimal doping in $Bi_2Sr_2Ca_{1-y}Dy_yCu_2O_{8+\delta}$, with underdoped samples revealing bimolecular kinetics (recombination rate proportional to quasiparticle density) and with overdoped samples exhibiting a faster decay that is independent of excitation density.

More generally, the original motivation behind studying the recovery dynamics in superconductors on an ultrafast timescale was to understand the fundamental process of quasiparticle recombination (i.e., how free quasiparticles form into Cooper pairs). However, this process is complicated by the excitation of high energy ($>2\Delta$) phonons during the photoexcitation process, since

these phonons can also break Cooper pairs. The Rothwarf-Taylor (RT) rate equations comprise a simple model describing many of the features of quasiparticle recombination in superconductors:

$$dn/dt = \beta N - Rn^2 + I_0$$

$$dN/dt = \frac{1}{2}\left[Rn^2 - \beta N\right] - [N - N_T]/\tau_\gamma$$

where n is the quasiparticle density, N is the excess density of high energy ($>2\Delta$) phonons, τ_γ is the phonon relaxation time, R is the bare quasiparticle recombination rate, N_T is the equilibrium number of phonons at a temperature T, and β is the pair breaking coefficient for $>2\Delta$ phonons. When N_T is small, the n^2 term dominates the recovery, but in many experiments the phonon decay significantly affects the recovery dynamics, a phenomenon described as the "phonon bottleneck." The RT model can account for the intensity and temperature dependence of both the photoinduced quasiparticle density and the pair-breaking/superconducting state recovery dynamics in conventional as well as cuprate superconductors.

For example, an OPOP study of the photoexcited quasiparticle dynamics in the HTSC $Tl_2Ba_2Ca_2Cu_3O_y$ revealed that, deep into the superconducting state (below 40 K), a dramatic change was observed in the temporal dynamics, associated with photoexcited quasiparticles rejoining the condensate. An analysis based on the RT equations suggested that these dynamics resulted from the entry into a coexistence phase which opens a gap in the density of states (in addition to the superconducting gap) that competes with superconductivity, resulting in a depression of the superconducting gap. Similarly, competing orders and phase separation were revealed through a RT analysis of the quasiparticle relaxation dynamics in the underdoped HTSC $(Ba,K)Fe_2As_2$, obtained through all-optical ultrafast pump–probe experiments. Specifically, it was observed that spin density wave (SDW) order forms at 85 K, followed by superconductivity at 28 K. Additionally, a normal-state order emerges, suppressing the SDW around $\sim$60 K, which was hypothesized to constitute a precursor to superconductivity.

Adaptations of the Rothwarf-Taylor equations have been used to interpret ultrafast pump–probe experiments in other gapped CEMs. For example, the ultrafast carrier dynamics in heavy fermion systems exhibits dramatic nonlinear behavior as a function of temperature and fluence. **Fig. 9** depicts the relaxation time τ obtained using OPOP spectroscopy at 1.5 eV in the prototypical HF compound $YbAgCu_4$, which has a Kondo temperature $T_K \sim 100$ K and an electronic specific heat coefficient $\gamma = 210$ mJ/molK^2. Surprisingly, below T_K, τ increases by more than two orders of magnitude (to $\sim$100 ps), in dramatic contrast to the nonmagnetic analog $LuAgCu_4$, where $\tau < 1$ ps at all temperatures. This behavior of the relaxation dynamics is consistently observed in various HF metals and Kondo insulators. The temperature and fluence dependence of the observed dynamics in all of these systems can be modeled using a RT framework, where the dynamics are governed by the presence of the hybridization gap (see Section "Heavy Fermion Materials"). In addition, OPOP spectroscopy, combined with a RT analysis, was utilized to investigate carrier dynamics in the "hidden order" HF compound URu_2Si_2 (discussed in more detail below in Section "Time and Angle-Resolved Photoemission Spectroscopy"). The amplitude and decay time of the photoinduced reflectivity were found to increase in the vicinity of the coherence temperature $T^* \sim 57$ K, consistent with the presence of a 10 meV hybridization gap. However, at 25 K, a crossover regime is manifested as a new feature in the carrier dynamics that saturates below the hidden-order transition temperature of 17.5 K, indicating the formation of a 5 meV pseudogap that separates the normal Kondo lattice state from a hidden order phase.

Finally, in the itinerant HF antiferromagnet $UNiGa_5$, OPOP measurements revealed a sharp increase in τ at the Neel temperature, $T_N = 85$ K, and exhibited a temperature dependence of τ below T_N that is consistent with the opening of a SDW gap

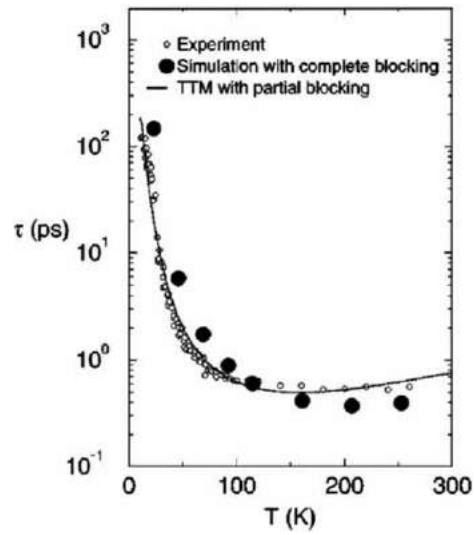

Fig. 9 e-ph relaxation time in $YbAgCu_4$ versus temperature. The solid line is a fit using the phenomenological two-temperature model, while the filled black circles are calculations using a semiclassical Boltzmann approach. Reprinted with permission from Demsar, J., Averitt, R.D., Ahn, K.H., *et al.*, 2003. Physical Review Letters 91, 027401.

leading to a quasiparticle recombination bottleneck, as revealed by the RT model. Following these results, ultrafast dynamics in the SDW phase of the similar itinerant HF antiferromagnet $UPtGa_5$ were also studied, revealing two relaxation components: (a) a slow component whose amplitude appears below T_N and relaxation time τ_{slow} exhibits an upturn near T_N, and (b) a fast component (τ_{fast}) that persists at all temperatures, also exhibiting an upturn near T_N. The differences with the dynamics observed in $UNiGa_5$ were explained using a RT framework and were primarily due to $UPtGa_5$ having A-type (rather than G-type) AFM order, resulting in partial Fermi surface nesting, partial gapping, and consequently a finite density of states at the Fermi surface.

Ultrafast Infrared Spectroscopy: THz to Mid-IR

Dynamic investigations of correlated electron materials in the mid-to-far-infrared spectral range are of particular interest, because the multitude of low energy excitations in this frequency range (such as charge, spin or superconducting gaps, polaronic excitations, and magnon or phonon modes) can be directly probed with ultrafast pump–probe techniques. Low energy photons also enable dynamical measurements of the optical conductivity, relevant, for example, to understanding the competing degrees of freedom involved in metal-insulator phase transitions. Recent advances in femtosecond pulse generation and amplification have enabled the efficient generation of intense ultrashort pulses from ultraviolet to mid-IR (50–500 meV) wavelengths using nonlinear optical parametric amplification and difference frequency generation processes. These pulses can be used as both pump and probe in ultrafast optical experiments, and their phase and amplitude can be manipulated to either coherently drive low-energy modes or monitor the evolution of relevant degrees of freedom following photoexcitation.

Ultrafast mid-IR spectroscopy has been applied to a broad range of CEMs, including high-temperature superconductors and colossal magnetoresistive oxides. As mentioned earlier (see Section "Transition Metal Oxides"), nanoscale inhomogeneities and phase coexistence play an important role in determining the electronic properties and macroscopic response to external electric or magnetic fields in CMR manganites. Ultrafast optical spectroscopy is sensitive to these inhomogeneities; for example, optical excitation can undress JT polarons by transferring electrons between Mn^{3+} and Mn^{4+} ions (**Fig. 2(b)**), which subsequently reform on a sub-picosecond timescale. Therefore, mid-IR spectroscopy, which can probe the polaron absorption peak observed in the optical conductivity, provides a powerful approach to study nanoscale phase separation in manganites by determining the evolution of the polaronic response with temperature. In particular, the persistence of specific polaronic signatures as the temperature is reduced below T_c should strongly indicate multiple phase coexistence.

Ultrafast mid-IR measurements of polaron dynamics were performed on the manganite $Nd_{0.5}Sr_{0.5}MnO_3$ (NSMO), revealing a mixed phase response. NSMO undergoes multiple phase transitions, from PM to FM ($T_c \sim 250$ K) to charge ordered (CO) states ($T_{CO} \sim 160$ K), accompanied by AFM ordering below T_{CO} (**Fig. 10**). Upon cooling towards T_c, correlations develop between polarons, resulting in the formation of nanoscale charge-orbital ordered clusters. The amplitude of the fast response, due to polarons, decreased as the temperature was lowered below T_c, but remained relatively large all the way down to T_{CO} (**Fig. 10**). This indicated coexistence of the semiconducting PM phase with the FM metallic one, in agreement with both X-ray and neutron scattering measurements. Furthermore, at lower temperatures the amplitude of the fast polaronic response scaled with the volume fraction of the CO phase, further demonstrating that UOS can be a sensitive probe of nanoscale inhomogeneities in CEMs.

A contrasting example is provided by magnetoresistive pyrochlores, such as $Tl_2Mn_2O_7$, which lack JT-active Mn^{3+} sites (unlike CMR manganites), so that charge-lattice interactions (either polaronic or thermally induced) do not play a major role in the

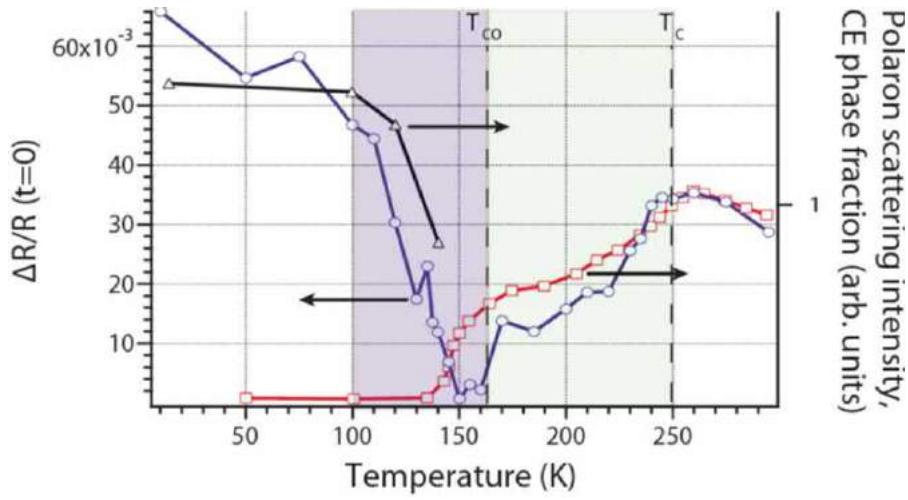

Fig. 10 Amplitude of the polaronic response at a probe wavelength of 13 μm at zero delay (blue line/circles) compared to single-polaron diffuse scattering measured by X-ray diffraction (red line/squares) and the antiferromagnetic volume phase fraction (black line/triangles) measured by neutron scattering. The shaded areas indicate temperature ranges in which minority phases compete with the FM (green) and COO (blue) majority phases. Reproduced from Prasankumar, R.P., Zvyagin, S., Kamenev, K.V., *et al.*, 2007. Physical Review B 76, 020402(R).

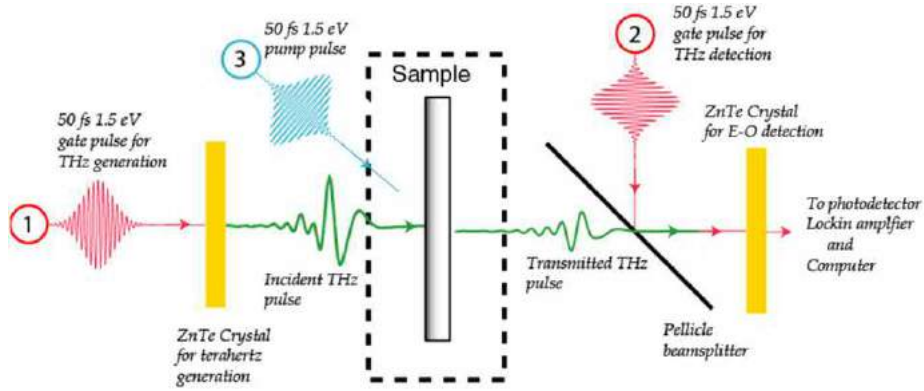

Fig. 11 Schematic of the typical experimental arrangement for THz-TDS and OPTP spectroscopies. Beam 1 generates THz generation via difference frequency mixing in nonlinear crystals (e.g., ZnTe) or photoconductive antennas. The THz radiation transmitted through the sample is measured through the electro-optic effect by overlapping it with beam 2 in the second crystal or antenna. By adding beam 3, the sample can be photoexcited and the photoinduced changes in the THz transmission measured as a function of the pump–probe delay.

emergence of CMR. Nevertheless, the use of ultrafast mid-IR spectroscopy to track photoexcited charge dynamics in $Tl_2Mn_2O_7$ demonstrated that charge localization in the insulating phase is caused by spin, rather than lattice, fluctuations. This implied that the CMR response in this material is determined primarily by the nanoscale inhomogeneities in spin ordering on Mn sites, unlike the double exchange and JT polaron mechanisms believed to govern the CMR manganites.

These studies of low-energy properties and materials dynamics can be further extended into the far-IR using sub-ps, broadband pulses with frequencies of $\sim$0.1–20 THz that are generated and measured through photoconductive or electro-optic techniques (Fig. 11). Importantly, both the amplitude and phase of the THz electric field can be measured in the time domain, providing direct access to both the real and imaginary parts of the dielectric function or conductivity at THz frequencies through relatively simple Fourier transform routines. These THz time-domain spectroscopy (THz-TDS) experiments can be performed with extremely high signal-to-noise ratio without the need for cooled detectors, such as bolometers, constituting a major advantage of THz spectroscopy over conventional optical conductivity measurements using FTIR (see Section "Time-Integrated Optical Spectroscopy of Correlated Electron Materials"), which require complex Kramers–Kronig transforms to extract $\sigma(\omega)$. In addition, the same temporally coherent femtosecond pulses that generate and detect THz bursts can also be used for photoexcitation, enabling optical-pump THz-probe (OPTP) studies of photoinduced changes in $\sigma(\omega)$ with sub-ps time resolution.

THz spectroscopy is particularly well suited for investigations of metal-insulator phase transitions, where multiple charge scattering and localization processes compete to determine the ultimate conductivity. For example, OPTP has been used to unveil the relative contributions of spin and lattice fluctuations to the increasing resistivity of the FM metallic phase in CMR manganites. Systematic studies performed on optimally doped manganites using OPOP spectroscopy have shown that the photoinduced dynamics were dominated by two components: (1) a fast (few ps) temperature-independent decay associated with e-ph relaxation, similar to metals (see Section "All-Optical Pump–Probe Spectroscopy of Correlated Electron Materials"), and (2) a slow (hundreds of ps) component, with the time constant peaking at the ferromagnetic T_c. The latter process corresponds to gradual spin disordering caused by energy transfer from the hot electrons to the lattice to the spin subsystem, and can be described by the three-temperature model (including electrons, phonons, and spins). This model was used to extract the electron–lattice, electron–spin, and spin–lattice energy transfer rates in various manganites, but could not quantify the role of particular subsystems in the degradation of the conductivity with increasing temperature.

OPTP provided a means for differentiating the effects of spin and lattice fluctuations on charge transport by following the photoinduced conductivity changes in the time domain. Fig. 12 shows the results of OPTP spectroscopy performed on films of the CMR manganites LCMO and LSMO. These measurements revealed the ability of THz experiments to obtain conductivity values both with and without optical excitation in a contactless manner. In addition, Fig. 12(a) demonstrates that time-resolved changes in the conductivity at various temperatures following photoexcitation with a 100 fs pulse at 1.55 eV evolve on two characteristic timescales. The fast component is related to electron-phonon relaxation, while the slow component is related to spin-lattice thermalization, as in the OPOP results discussed above. An analysis of the relative amplitudes of the fast and slow components observed in the OPTP response enabled the authors to determine the time-dependent spin and lattice temperatures T_{spin} and T_{phonon}, and quantify the relative importance of spin disorder ($\partial\sigma/\partial T_{spin}$) versus thermally disordered phonons ($\partial\sigma/\partial T_{phonon}$) in limiting the conductivity below T_c. Fig. 12(b) shows that the total $\partial\sigma/\partial T$ is primarily determined by thermally disordered phonons, while spin fluctuations dominate near T_c, in accordance with double-exchange mechanisms governing charge transport in CMR manganites.

OPTP has also been used to reveal the dynamics of phase separation in VO_2, caused by the complex interplay between electron correlations and lattice re-arrangements. In this material, the MIT, as well as the accompanying structural dynamics, can be induced non-thermally using impulsive photoexcitation above the $\sim$7 mJ/cm^2 fluence threshold. For this reason, the MIT in VO_2 has been extensively investigated by ultrafast techniques in an attempt to reveal whether electron localization derives from lattice distortions (band insulator) or on-site Coulomb repulsion (Mott insulator). OPOP measurements while varying the laser pulse duration

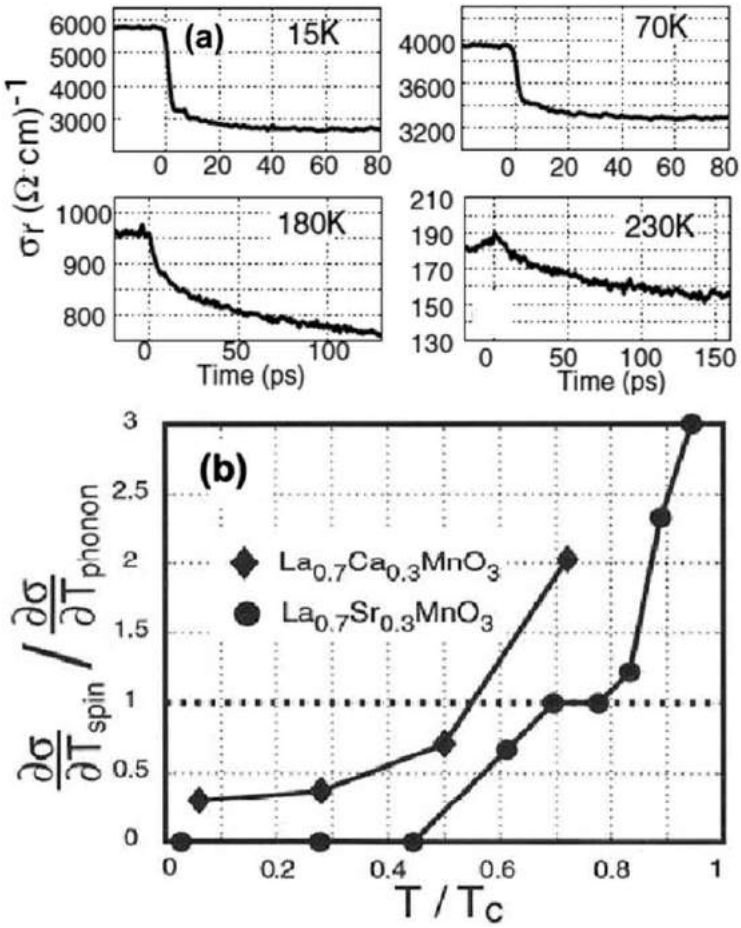

Fig. 12 (a) Two-component decrease in the real conductivity of photoexcited epitaxial LCMO films. (b) Relative contributions of the spin and phonon fluctuations to the conductivity changes as a function of temperature. Reprinted with permission from Averitt, R.D., Lobad, A.I., Kwon, C., *et al.*, 2001. Physical Review Letters 87, 017401.

showed that the intrinsic MIT timescale saturates at ~80 fs. This is too fast for thermally driven transitions that occur within e-ph thermalization times of a few ps, but is too slow for the almost instantaneous electronic dynamics expected to occur within the excitation pulsewidth. Therefore, the observed transition rates and concomitant optical phonon dynamics implied that the metallic phase in both the photoinduced and thermally driven MIT emerges primarily from atomic structural motion, rather than the time-varying electron correlations. Moreover, the strong dependence of the MIT rate on the electron excitation density also suggested that the metallic phase gradually grows from the confines of the optical absorption depth into the bulk. The dynamics of phase separation and domain formation near the MIT threshold are reflected in the time-dependent conductivity, which can be directly accessed by OPTP. As illustrated in **Fig. 13(a)**, the observed photoinduced conductivity changes for a given excitation fluence are significantly enhanced (up to $\sigma = 1000\ \Omega^{-1}\ \text{cm}^{-1}$ in the metallic phase) as the temperature is increased toward $T_c = 340$ K. This response indicates a "softening" of the insulating state that leads to a reduction of the deposited energy required to drive the MIT in VO_2 (**Fig. 13(b)**). A simple effective medium model attributed this reduction to an increasing volume fraction of metallic precursors in the insulating bulk of the material as it approaches the MIT. These metallic seeds facilitate the growth and coalescence of a metallic conducting phase, initiated by absorption of the optical pulse. Inhomogeneity and phase coexistence are ubiquitous in CEM, and OPTP provides new insight into the mechanisms governing the delicate balance between multiple competing interactions that leads to this behavior.

OPTP has also been particularly useful for studying superconductors, since the superconducting condensate density is proportional to a $1/\omega$ component in the imaginary conductivity; these measurements therefore enable direct tracking of the condensate recovery dynamics. The first OPTP measurements on superconductors were performed on underdoped and near optimally doped $Y_{1-x}Ba_2Cu_3O_{7-x}$ and $Y_{1-x}Pr_xBa_2Cu_3O_7$ films. The optimally doped film exhibited a temperature-dependent recovery increasing from ~1.7 ps at 10 K to >4 ps below the superconducting transition temperature (T_c), indicative of dynamics following the appearance of the superconducting gap. The lifetime decreased to ~2 ps above T_c, due to e-ph thermalization in the normal state. The underdoped films revealed a temperature-independent recovery time of ~3.5 ps even above T_c, indicative of dynamics influenced by a pseudogap.

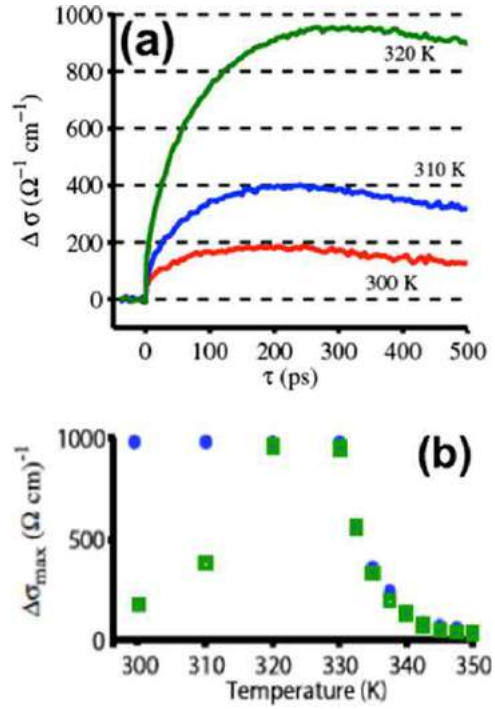

Fig. 13 (a) Induced conductivity change as a function of time for various initial temperatures. (b) Magnitudes of the photoinduced conductivity changes (green squares) measured with the same fluence at different temperatures and the maximum possible conductivity changes that can bring its magnitude to that of the metallic phase (blue circles). Reproduced from Hilton, D.J., Prasankumar, R.P., Fourmaux, S., *et al.*, 2007. Physical Review Letters 99, 226401.

Detailed OPTP studies of the recovery dynamics in the BCS superconductor MgB_2 as a function of photoexcitation fluence and temperature were analyzed using the RT equations (see Section "All-Optical Pump–Probe Spectroscopy of Correlated Electron Materials"), enabling the determination of β, τ_γ, and R. Remarkably, the pair breaking process was found to be quite slow, occurring on a timescale of ~10 ps, in contrast to the ~1 ps timescale observed for Pb, another BCS superconductor, as well as for the cuprate superconductors. This "anomalous" pair breaking dynamics is attributed to energy relaxation to high frequency phonons following photoexcitation instead of, as commonly assumed, electron-electron thermalization.

Apart from photoinduced changes in the conductivity, changes in the resonant absorption of THz radiation by low energy excitations (e.g., magnons, phonons, and polarons) manifest the emergence and disappearance of relevant order parameters as a CEM undergoes multiple phase transitions. **Fig. 14(a)** shows an example of this in the AFM multiferroic $HoMnO_3$, which has a spin wave (magnon) resonance in the THz transmission spectrum when the material becomes magnetically ordered below $T_N = 42$ K. In general, AFM order is difficult to detect with optical techniques due to its net zero magnetization, which nullifies the magneto-optical Kerr effect. THz spectroscopy provides an alternative way of probing AFM order using the associated magnon resonances, and OPTP can unveil the coupling of spins to other degrees of freedom by separating them in the time domain.

The dynamics of spin-lattice relaxation in $HoMnO_3$ following photoexcitation is shown in **Fig. 14(b)**. Importantly, not only the peak amplitude, but the entire spectrum of the THz pulse can be detected as a function of pump–probe delay, making the temporal evolution of the resonance lineshape readily available for analysis (insets in **Fig. 14(b)**). Photoinduced changes in the spectra of the transmitted THz E-field measured at different sample temperatures and pump fluences match well with how the magnon mode spectral shape would change with increasing sample temperature, as shown in **Fig. 14(a)**. This implies that the ~4–12 ps rise time dynamics observed in the OPTP signal corresponds to the transfer of energy from the lattice (heated by much faster e-ph thermalization) to the spins, similar to that observed in FM CMR manganites. However, unlike in CMR manganites, spin-lattice relaxation in $HoMnO_3$ is much faster, indicating more efficient magnon-phonon scattering in AFMs. This represents a fundamental difference in spin dynamics in AFM and FM materials – the requirement for total spin conservation allows lattice vibrations to directly heat (or disorder) AFM spins, while spin-lattice thermalization in FMs relies on smaller, less efficient terms, like spin-orbit coupling.

Finally, THz emission spectroscopy, where THz fields are generated directly from the sample by optical photoexcitation, has emerged as a promising tool for investigating dynamic phenomena in CEM and other materials. For example, photoexcitation of an iron thin film with an intense optical pulse resulted in the emission of coherent far-IR radiation. This is surprising, since metals are generally good reflectors. Subsequent investigations of the emission efficiency as a function of the sample azimuthal angle (i.e., crystal symmetry) demonstrated that THz radiation is produced by ultrafast demagnetization on a ~2 ps timescale. The changing magnetic field necessarily results in a time-varying electric field, yielding a radiated signal, but these timescales are too short to be

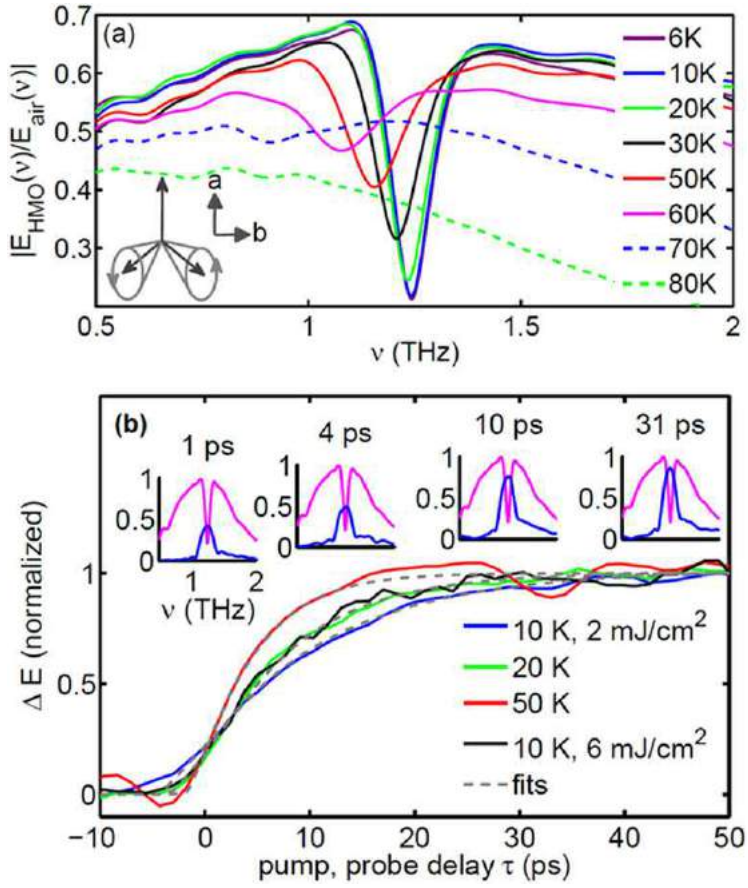

Fig. 14 (a) Temperature dependence of the magnon resonance in the THz transmission spectrum of $HoMnO_3$. The magnon mode is illustrated at the bottom left. (b) Normalized OPTP signals taken at different temperatures and excitation fluences. The top insets show photoinduced changes in the THz transmission spectra for several different pump–probe delays (blue), overlapped with the THz transmission spectrum without photoexcitation. Reproduced from Bowlan, P., Trugman, S.A., Bowlan, J., *et al.*, 21016. Physical Review B (Rapid Communications) 94, 100404(R).

consistent with a scenario where spin waves are emitted following electron-lattice-spin energy transfer. It is instead likely that ultrafast demagnetization is a non-thermal process, caused either by spin currents or localized spin-orbit interactions. The strong dependence of the emitted THz signal on the magnetization and crystal symmetry might find numerous applications in studies of thermally and photoinduced phase transitions, with strong synergy between structural motion and spin ordering dynamics.

Photoinduced Phase Transitions and Mode-Selective Excitation

The subtle balance between competing interactions in CEMs leads to the coexistence of multiple, almost degenerate states that can serve as an ultimate ground state under proper perturbation. Doping, temperature variation or application of pressure and fields may alter the balance in favor of a desired order parameter and steer the material to assume the corresponding ground state. This multistability makes CEM an attractive choice for both technological applications and studies of complex phase transformation phenomena. However, the energy barriers separating the ground states can be rather large for conventional thermal or pressure drivers to initiate the phase transition (Fig. 15). Ultrafast spectroscopy offers an alternative way to steer the system between phase stability regions on the multidimensional potential energy surface by using (i) electronic excited states (e.g., photodoping) or (ii) coherent excitation of low-energy modes corresponding to relevant DOF (e.g., phonons or magnons), as illustrated by the arrows (i, ii) in Fig. 15. Although the original balance between orders will eventually be re-established by thermal fluctuations, important information can be obtained from these measurements about short-lived metastable states and recovery of the associated competing interactions.

Many investigations of photoinduced phase transitions (PIPT) in CEM used visible or near-IR photons to photodope the system (process (i) in Fig. 15) and broadband probes to follow the ensuing non-equilibrium dynamics. The most extensively studied PIPT is in VO_2, where femtosecond photodoping was shown to induce a structural shift from monoclinic to rutile, accompanied by collapse of the gap and charge delocalization within ~100 fs. In addition, ultrafast electron diffraction was used to reveal a new metastable state of unidentified nature in a cuprate superconductor driven by 1.5 eV photons (see Section "Ultrafast Electron Diffraction"). Finally, ultrafast X-ray probes allowed deciphering of the dynamic coupling among charge order,

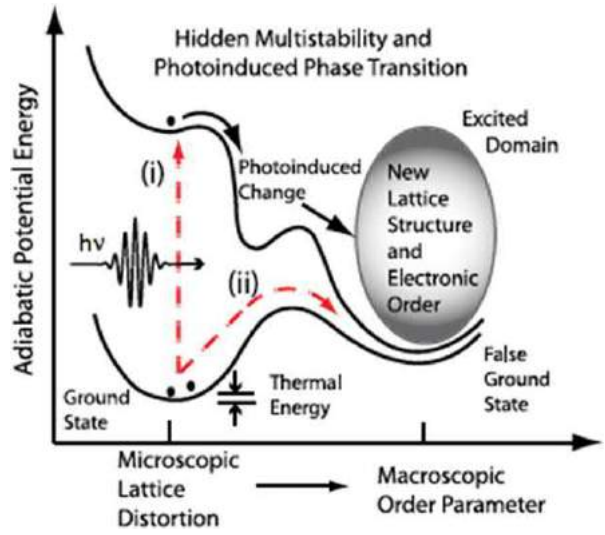

Fig. 15 Schematic representation of the potential energy dependence on a structural or other parameter. Nearly degenerate ground states are separated by an energy barrier, which can be overcome either through thermal fluctuations or by driving the system to the new phase through the excited (i) or ground (ii) states using photons of appropriate energy. Reproduced from Nasu, K., Ping, H., Mizouchi, H., 2001. Journal of Physics: Condensed Matter 13, R693.

orbital order and atomic structure during light-induced melting in a charge-and-orbital ordered manganite using a single time-dependent order parameter (see Section "Ultrafast Extreme UV and X-Ray Diffraction and Spectroscopy").

These experiments provided much insight into the mechanisms governing PIPT, but have shown that interpretation of the time-resolved signals is very complicated and requires simultaneous observation of the dynamics of more than one DOF (e.g., electronic, lattice and spin). A powerful approach to resolving this issue would be to directly excite specific low-energy modes and observe their effect on a given order parameter as it evolves across multiple stability regions without leaving the system ground state (process (ii) in **Fig. 15**). Coherent lattice motion can be induced in several ways at optical frequencies. Rapid thermal expansion, produced by electron-phonon thermalization of photoexcited carriers, can launch an acoustic phonon wavepacket and modulate the local strain state within the pulse duration. Ionic rearrangement to accommodate photoinduced charge density changes can generate optical phonon oscillations. Alternatively, impulsive stimulated Raman scattering of a tailored train of non-resonant femtosecond pulses can drive lattice oscillations to the anharmonic regime. On the other hand, coherent induction of collective spin oscillations (magnons) can be accomplished through the inverse Faraday effect with circularly polarized femtosecond pulses. Nevertheless, all-optical excitation of low-energy DOF (and PIPT) often introduces a large thermal load on the sample, due to the dissipation of the excess photon energy through electron-phonon scattering, which interferes with non-thermal PIPT mechanisms and steers the order parameter in an undesired direction. Moreover, the number of optically accessible modes is limited by the symmetry of the material, leaving many relevant DOF inaccessible through optical processes.

The above shortcomings of *optically* exciting low energy modes are remedied by the use of intense mid-IR and THz pulses to *directly* excite IR-active modes. Modern optical parametric amplifiers and difference frequency generators supply broadly tunable, high energy mid-IR pulses, while the frequency mixing and tilted pulse-front techniques produce single-cycle THz pulses with electric and magnetic field amplitudes of $\sim$1 MV/cm and 0.3 T or more, respectively. These pulses can now be used to directly photoexcite and probe specific low energy modes, improving upon previous all-optical techniques that only indirectly coupled to these excitations. Here, we describe several representative applications of these techniques to mode-selective excitation of correlated materials, while more detail on intense THz generation methods and investigations of a broader class of materials can be found in the reading list.

In the first application to correlated materials, mid-IR mode-selective excitation converted a manganite from insulating to metallic on sub-ps timescales. $Pr_{1-x}Ca_xMnO_3$ is the only perovskite manganite that remains insulating for all doping levels x. However, it was shown to have a thermally inaccessible metallic phase, which could be induced by a magnetic field or optical excitation. In perovskite manganites, the O-Mn-O bond angle is closely related to the inter-site electron hopping amplitude (t) and determines the ratio between electron bandwidth and on-site Coulomb repulsion (U) (see Section "Introduction"), hence determining the conductivity. Modulation of the O-Mn-O angle by resonant excitation of an IR-active Mn-O stretching mode (**Fig. 16(a)**) at $\sim$71 meV was shown to change the reflectivity in the same manner as interband absorption of optical pulses (**Fig. 16(b)**). Simultaneous transport measurements exhibited a five order of magnitude increase in the DC conductivity, indicating a vibrationally induced MIT. The dominant role of Mn-O oscillations as a PIPT driver was further confirmed by matching the mid-IR excitation spectrum to the respective phonon resonance in optical conductivity (**Fig. 16(c)**). In subsequent studies of other manganites, resonant Mn-O excitations were demonstrated to induce ultrafast demagnetization and efficient orbital order melting,

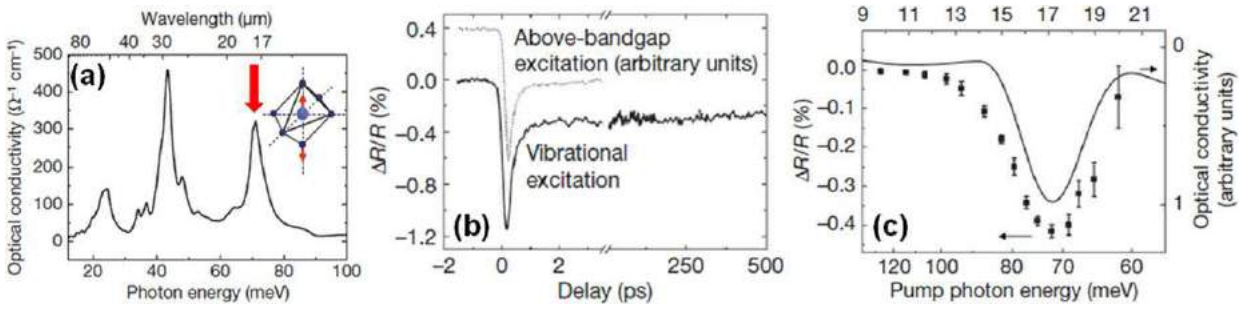

Fig. 16 Mode-selective excitation of the MIT in $Pr_{0.7}Ca_{0.3}MnO_3$. (a) Mid-IR absorption spectrum showing the Mg-O stretching mode (inset) at ~70 meV (red arrow). (b) Comparison between the transient reflectivity changes at an 800 nm probe wavelength produced by mid-IR (solid line) and 800 nm (dotted line) excitation. The similarity between these dynamics implies that mid-IR pulses can drive the MIT. (c) Mid-IR static optical conductivity (solid line) and the amplitude of transient reflectivity changes as a function of pump wavelength (squares). Reproduced from Rini, M., Tobey, R., Dean, N., *et al.*, 2007. Nature 449, 72.

with absorbed energies of less than 1% of that necessary for thermally induced transitions. Interestingly, both electronic and magnetic orders in manganite heterostructures are amenable to external coherent strain modulation, caused by resonant vibrational excitation of stretching modes in the adjacent substrate.

Mode-selective excitation has yielded even more intriguing results in high-T_c superconductors. In these materials, the search for room temperature superconductivity has inspired intensive inquiry into the mechanisms responsible for converting parent AFM Mott insulators into cuprate HTSC by increasing the hole or electron doping. Across a large portion of the phase diagram, superconductivity in these compounds is suppressed by a rich variety of different order parameters. Mode-selective excitation can unveil the competing states that are hidden from other probes, and help in understanding the HTSC suppression and emergence mechanisms. For example, $La_{1.675}Eu_{0.2}Sr_{0.125}CuO_4$ belongs to the family of underdoped compounds, where HTSC is suppressed by a striped charge ordered state. Resonant mid-IR excitation of an in-plane Cu-O stretching mode at 80 meV was able to alter the balance between these competing orders and produce a broadband THz response, reminiscent of the equilibrium superconducting phase at higher doping. This transient superconducting state lasted for nanoseconds and persisted to temperatures well above the normal T_c for this doping level. This behavior was attributed to possible formation of an energy barrier that stabilized the new superconducting state against thermal fluctuations.

Intense THz pulses were also used to drive ionic motion into an anharmonic regime and map the potential energy of a resonant phonon mode in $SrTiO_3$. The large ferroelectric polarization of $SrTiO_3$ is associated with a soft mode phonon, in which Ti^{4+} ions oscillate between O^{2-} ions on the opposite face centers. This mode resonantly absorbs 1.5 THz radiation when a low amplitude THz field is used. However, when the field was increased to 80 kV/cm, the phonon resonance shifted to higher frequencies due to dynamic stiffening of the confinement potential. This was an indication that the vibrational amplitude reached the anharmonic (quartic) part of the potential, with the photoinduced ion displacements exceeding 1 picometer (pm) – a magnitude comparable to the static ionic displacements in the FE state. Thus, intense coherent THz excitation can potentially be used to switch the macroscopic FE polarization on picosecond timescales.

Ultrafast control of magnetization dynamics promises numerous applications in magnetic switching devices. Manipulation of FM order can be achieved with polarized optical pulses or the magnetic field of intense THz pulses, but control over potentially faster AFM switching is much harder to implement. However, the large magnetic fields available with intense THz pulses enable resonant excitation and coherent manipulation of spin dynamics in AFM materials. In the AFM oxide NiO, the broadband magnetic field of a single-cycle THz pulse triggered long-lived precession of antiferromagnetically aligned spins, with a resonant magnon frequency of 1 THz. A sequence of THz pulses, spaced by 6.5 precession periods, was then used to coherently quench coherent spin oscillations. Similarly, as described in Section "Magnetoelectric multiferroics," multiferroics feature new elementary excitations, electromagnons, which are expected to mediate and enable ultrafast switching of AFM order in multiferroics using transient electric fields. Kubacka *et al.* used resonant THz fields to excite electromagnon oscillations in the archetypical multiferroic $TbMnO_3$. Resonant X-ray scattering provided direct insight into the induced spin dynamics, demonstrating that the THz field caused synchronous rotation of the spin cycloid plane, with a 4° amplitude. Even larger rotational motion can be achieved by increasing the THz field, reaching ~ 90° for 1–2 MV/cm.

Apart from driving specific modes and observing their effect on various order parameters, the utility of intense THz pulses lies in their ability to induce and monitor dynamic phase transitions, similar to the mid-IR radiation discussed above. As with many other techniques, the MIT in VO_2 provides a fertile ground for exploring the capabilities of coherent THz excitation. Liu *et al.* demonstrated that intense THz excitation with peak fields of 300 kV/cm was insufficient to induce the metallic phase in VO_2. In order to determine the feasibility and threshold of such an MIT, they created resonant metal structures on the sample surface, where the electric field of the THz pulse could be enhanced by a factor of ~30 in the gap between closely spaced metal stripes. This enhancement enabled them to break the MIT threshold, transforming VO_2 into a metallic state within a few ps of excitation. Moreover, they found that fields in excess of ~4 MV/cm caused pronounced damage to the material due to electric breakdown.

Ultrafast Extreme UV and X-Ray Diffraction and Spectroscopy

The exotic properties of correlated electron materials emerge from a labyrinthine pattern of complex interactions among charge, lattice, spin and orbital degrees of freedom. Therefore, our understanding of the microscopic mechanisms underlying both static and transient material behavior would greatly benefit from the ability to selectively probe the dynamics and coupling between relevant excitations and order parameters. However, UOS in the visible/UV range often relies on the indirect coupling of microscopic processes to observable variations in the macroscopic dielectric function, and is limited to a material response averaged over large spatial areas and multiple concurrent dynamics. In Sections "Ultrafast Infrared Spectroscopy: THz to Mid-IR" and "Photoinduced Phase Transitions and Mode-Selective Excitation," we described how the expansion of UOS to mid-IR/THz frequencies enabled specificity in probing and driving low-energy modes. Similar opportunities exist on the other side of the spectrum, where a wealth of ultrafast X-ray techniques provide sub-nanometer spatial resolution, selectivity between charge, spin and lattice dynamics, and high element specificity.

Several different schemes have been implemented for the generation of femtosecond X-ray pulses. Tabletop plasma sources can produce hard (6–10 keV) X-ray bursts by focusing terawatt 1.5 eV femtosecond laser pulses onto a moving metal wire. Alternatively, high harmonic generation (HHG) can convert near and mid-IR radiation to coherent soft-X-ray pulses of variable polarization, with up to keV photon energies. Furthermore, at large facilities, laser slicing of pulsed synchrotron radiation can reduce the X-ray pulse duration to femtoseconds while still preserving the wide energy tunability of the emitted photons. Finally, in the past decade, accelerator-based X-ray free electron lasers (XFEL), such as the Linac Coherent Light Source (LCLS) at Stanford, have appeared on the scene, with the promise to open new areas in ultrafast materials science by producing femtosecond pulses with very high brightness over a wide photon energy range.

One of the earliest applications of ultrafast X-ray spectroscopies focused on correlating the lattice dynamics and electronic structure changes during the photoinduced MIT in VO_2. As illustrated in **Fig. 17(a)**, X-ray diffraction (XRD) measurements on VO_2 below the transition temperature at negative delays (i.e., on an unperturbed sample) show a broad peak around 13.9° that corresponds to the low-temperature insulating monoclinic crystalline structure. Following above-threshold photoexcitation, the structure changes to rutile, as indicated by the appearance of the corresponding diffraction peak around 13.8°. This transition occurs within 100 fs and correlates well with the dynamics of rutile metallic phase formation observed in OPOP experiments. The direct insight into atomic motion given by XRD was complemented by X-ray absorption measurements, which revealed concomitant dynamics in the electronic structure during the MIT. Near-edge X-ray absorption spectroscopy (XAS) probes the transitions from core atomic levels to the valence or conduction band, and is thus very sensitive to variations in the electronic structure, level population and ionic environments. **Fig. 17(b)** and **(c)** show time-resolved XAS at the V_{2p}–V_{3d} and O_{1s}–V_{3d} absorption edges, respectively. The photoinduced absorption increase at the vanadium V_{2p} resonance can be explained by rapid band gap collapse,

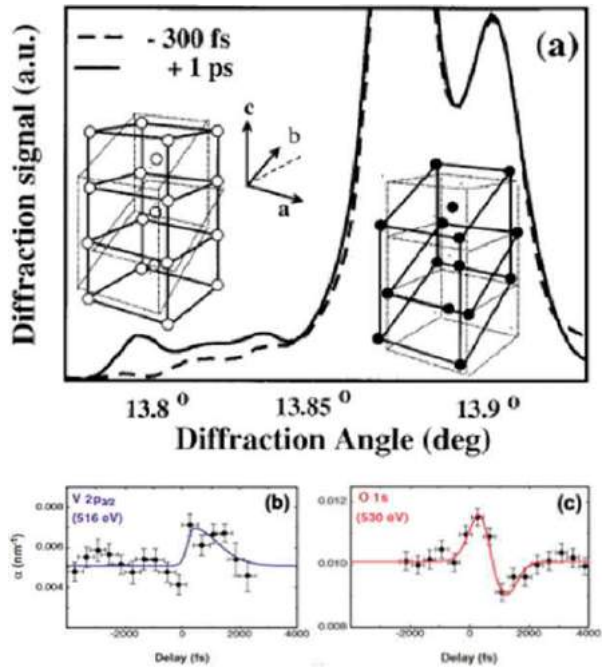

Fig. 17 (a) X-ray diffraction signal measured before (dashed) and after (solid) excitation of VO_2. for negative and positive time delays. The increase in diffraction around the 13.8° rutile peak at positive delay corresponds to the growing volume of the metallic phase during the photoinduced MIT (Reproduced from Cavalleri, A., Toth, Cs., Siders, C.W., *et al.*, 2001. Physical Review Letters 87, 237401.). (b, c) X-ray absorption dynamics, measured at the V_{2p} and O_{1s} resonances, is indicative of a gap opening in the electronic structure as the MIT develops in time. Reproduced from Cavalleri, A., Rini, M., Chong, H.H.W., *et al.*, 2005. Physical Review Letters 95, 067405.

with subsequent rapid (~1 ps) electron-phonon thermalization within the conduction band. This confirms the non-thermal nature of the photoinduced MIT in VO_2, because the energy gap disappears before the lattice is heated above T_c. However, the positive to negative XAS dynamics measured at the O_{1s} absorption edge (Fig. 17(c)) cannot be explained by gap collapse alone, and indicates core level shifts induced by photodoping and dynamic valence changes of the O^{2-} and V^{4+} ions.

These studies demonstrated the ability of ultrafast X-ray spectroscopy to separate atomic motion and electronic dynamics within bands with different orbital symmetries. In other experiments, sensitivity of the hard XRD efficiency to the lattice, orbital, and charge periodicities, and resonant enhancement of the soft X-ray scattering amplitude at the ionic absorption edges were exploited to selectively probe the dynamics of co-existing charge, orbital and magnetic order in correlated materials. Studies of phase transitions induced by coherent lattice, spin and charge excitations in bulk and heterostructured materials are particularly interesting, even essential, for both fundamental understanding of materials behavior and technological applications.

It is worth mentioning an important recent development in ultrafast X-ray science – the generation of attosecond (as) pulses. Because of the large frequency bandwidth (~18 eV for 100 as) required to support such a short pulsewidth, HHG or other nonlinear techniques that produce radiation in the extreme UV or soft X-ray portion of the electromagnetic spectrum have to be used. Attosecond spectroscopy should provide unprecedented insight into electron motion, which happens on an extremely fast timescale. For instance, electron-electron thermalization occurs within 10–100 fs, while other fundamental processes such as screening, charge imaging, and non-thermal electron generation occur on an attosecond timescale. Initial experiments in atomic gases and molecules thus show the great promise of attosecond spectroscopy for resolving electron dynamics; however, very few experiments have been performed in solid state systems, and thus far only to time the delay (~100 as) between electron emission from the core and delocalized energy states. Further advances in source reliability and detection techniques are thus necessary to harness attosecond pulses for materials science.

Ultrafast Electron Diffraction

Ultrafast electron diffraction (UED) offers an attractive alternative to X-rays for exploring ultrafast structural dynamics in correlated electron materials, especially low dimensional ones, for example, films or monolayers. The wavelength of electrons is shorter than typical X-ray photons (~20 pm for 30 keV electrons versus ~120 pm for 10 keV photons), which reduces the diffraction angles and allows more Bragg peaks to appear within the same detector area. This significantly increases the amount of structural information that can be extracted from UED experiments. Furthermore, electrons interact with matter much more strongly than X-rays and are more sensitive to small crystals or microscale sample regions. This is also a disadvantage due to the importance of multiple scattering events, which restricts the electron penetration depth (~tens of nm for 10 keV electrons) and might complicate structural determination from the analysis of ED patterns from bulk crystals. Nevertheless, UED is gaining increasing attention, with a growing number of applications to CEM.

UED is a pump–probe technique in which femtosecond UV probe pulses are used to generate electron packets by photoemission from a cathode material. The electrons are then accelerated towards the sample to produce diffraction patterns upon transmission or grazing reflection from the photoexcited material. In table-top systems, electrostatic fields are usually used to reach electron energies up to 100 keV. In order to probe a larger region of reciprocal space or increase the penetration depth, 2–6 MeV electrons can be generated with radio-frequency driven linear accelerators. The temporal resolution of UED is primarily determined by the electron bunch duration, which is limited by Coulomb repulsion to 200–300 fs. Different concepts have been developed to address this issue, extending all the way to producing single-electron pulses with high repetition rates to reach the 100 fs limit. Geometrical broadening due to different pump and probe incidence angles and other propagation effects can also be corrected by tilting the front of the optical excitation pulse. Finally, similar to ultrafast X-ray techniques, efforts have been made to combine UED with spectroscopic measurements in order to correlate structural dynamics to the evolution of electronic and other properties.

The majority of previous UED studies on CEM focused on photoinduced phase transitions, with the goal of separating structural dynamics from other DOF. In the high-temperature superconductor $La_2CuO_{4+\delta}$, the emergence of a non-equilibrium structural shift within 27 ps after photon absorption was observed with UED. As previously discussed, undoped HTSC are antiferromagnetic Mott insulators, and hole or electron doping induces superconductivity below the critical temperature (T_c) with a metallic state above it. A photoinduced structural change occurred only when the photodoping reached ~0.12 photons per copper site, which was very close to the optimal chemical doping level of 0.16 holes per copper site. The nature of the PIPT in this material was not identified, but possible light-induced charge redistribution might be responsible for driving the system through the inverse Mott transition.

CDW materials feature strong coupling between periodic charge and lattice modulations. The trajectories and cooperative character of charge and atomic motion can be independently followed by simultaneously observing the dynamics of lattice Bragg peaks and CDW satellite peaks using UED. In these experiments, melting of the charge order within hundreds of fs was identified, accompanied by collective lattice oscillations. The ensuing charge and lattice motions demonstrate well-defined separation of the coupled electron–lattice and CDW order parameters, leading to a commensurate-to-incommensurate CDW transition.

Finally, VO_2 was a natural candidate for UED studies, due to the unresolved question of whether the photoinduced MIT is structurally or charge driven. The ability to simultaneously detect multiple diffraction peaks (hence, crystallographic directions) as a function of time provided unprecedented insight into atomic motion during the MIT. Fig. 18(a) shows three characteristic timescales at which the diffraction intensity of Bragg peaks evolves following excitation with a 1.5 eV pulse. Immediately after

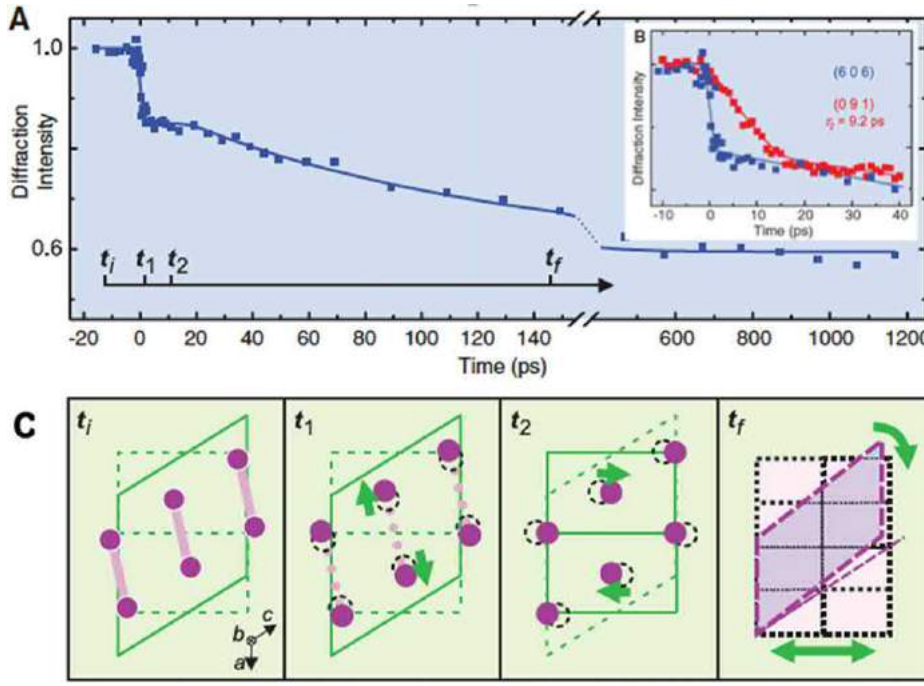

Fig. 18 (a) Intensity change of the (606) diffraction peak, associated with ionic motion along the V–V bond. Three characteristic timescales are also shown. (b) The blue curve describes the dynamics of V–V bond elongation, as measured by the (606) peak, and the red curve shows the translation of V ions normal to the V–V bond (091 peak) dynamics. (c) Schematics of the step-wise structural changes inside the unit cell and, on larger length scales, of shear motion following photoexcitation of VO_2. Reproduced from Baum, P., Yang, D.S., Zewail, A.H., 2007. Science 318, 788.

photon absorption, the peak intensity drops within t_1 and remains almost constant till t_2. Interestingly, t_1 had different values of 0.3 ps and 10 ps for two groups of UED peaks, which were associated with the atomic motion of vanadium ions along and across the V–V bond, respectively (**Fig. 18(b)**). The eventual saturation at t_2 was succeeded by much slower decay dynamics, on a timescale of $t_3 \sim 100$ ps. This kinetic behavior (and especially the discreteness of t_1) implied that the structural phase transition proceeded in the stepwise manner depicted in **Fig. 18(c)**. At a time t_i, photon absorption transfers the charge into a vanadium *d*-band with antibonding character, which induces repulsion between V atoms and their motion away from each other along the bond direction. This motion occurs at $t_1 \sim 0.3$ ps and affects only the corresponding Bragg peak intensities. Subsequent rotation of the V–V bond and unit cell transformation toward the metallic rutile configuration within $t_2 \sim 10$ ps can only be observed in another set of Bragg peaks. Finally, continued change of the peak intensity at $t_3 \sim 100$ ps indicates a shear that propagates into the bulk of the material and converts the crystalline structure to rutile. This process agrees well with the X-ray diffraction measurements in **Fig. 17(a)**. Moreover, comparison with OPOP and OPTP results indicated that metallic states appear while the material is still undergoing the monoclinic to rutile transition, i.e., V–V motion at t_1. This confirmed the presence of an intermediate monoclinic metal phase, similar to the strongly correlated metal state observed by IR microscopy during the thermally-driven MIT.

Time and Angle-Resolved Photoemission Spectroscopy

Knowledge of the electronic energy structure in momentum space and its evolution under external stimuli are essential for understanding the (non)equilibrium properties of correlated electron materials. While powerful, most UOS techniques are intrinsically momentum independent and cannot be used to track carrier relaxation dynamics across momentum space. On the other hand, static angle-resolved photoemission spectroscopy (ARPES) is capable of mapping occupied electronic states in momentum space. ARPES significantly advanced our understanding of correlated systems by revealing the dispersion of quasi-particles and signatures of many-body correlations in reciprocal space. Adding ultrafast time resolution to ARPES (trARPES) gives completely new insight into electronic structure dynamics, both below and above the Fermi level, and provides a complete set of information on the momentum-, energy-, and time-dependent excitations responsible for the properties of CEM.

ARPES is based on electron ejection from a crystalline material after illumination by UV photons with an energy exceeding the material work function (4–5 eV). The energy and momentum of the electrons before and after photoemission are related by conservation laws and can be measured with a proper spectrometer to create energy-momentum dispersion maps. In trARPES, ionizing radiation is usually produced either by upconverting the energy of 1.5 eV laser photons to 6–9 eV in a nonlinear crystal or by generating 10–60 eV photons through HHG in gases. HHG-based trARPES allows probing a larger volume within momentum space due to its higher photon energies, albeit at the cost of lower energy resolution (70–200 meV versus ~20 meV obtained with

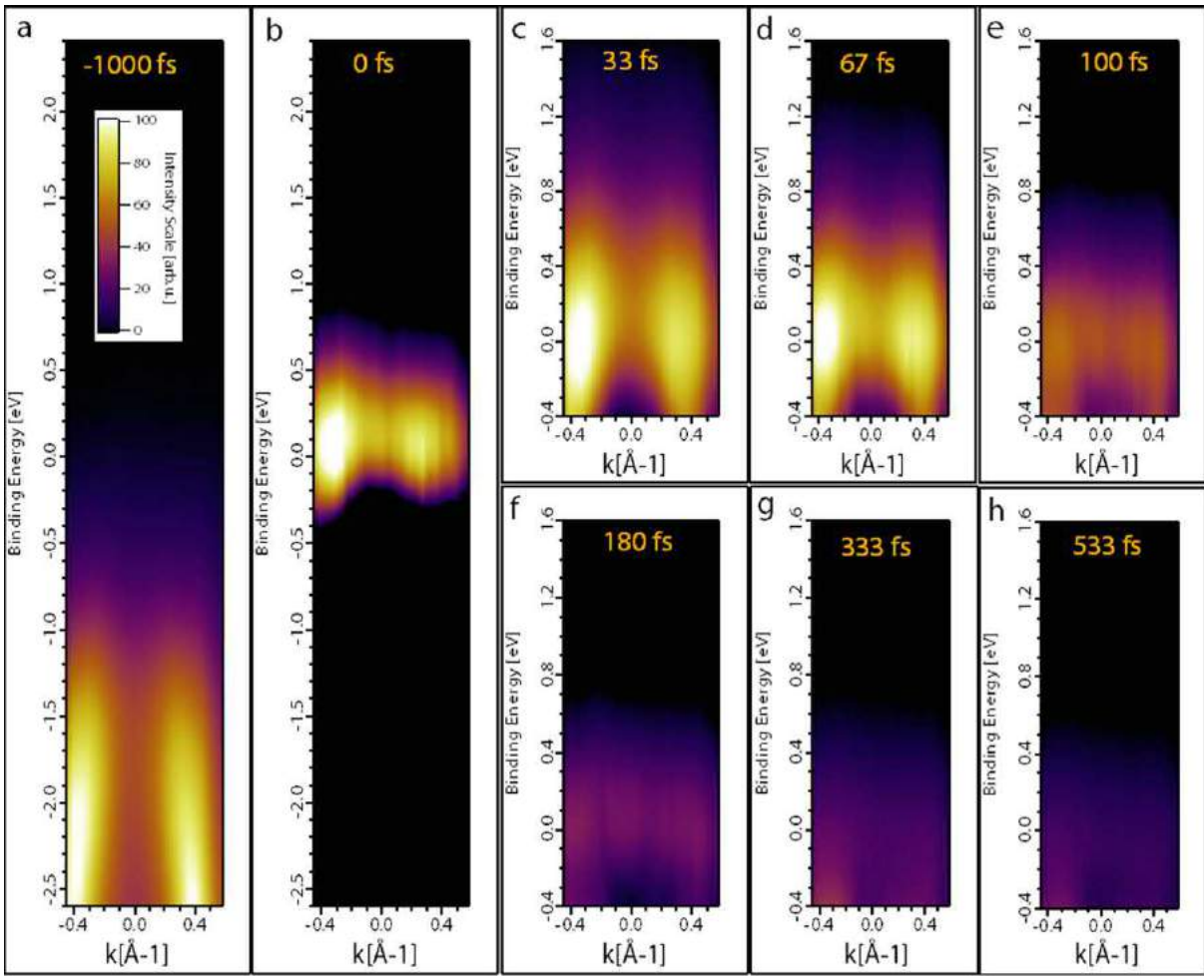

Fig. 19 Time-resolved ARPES study of URu_2Si_2 in the hidden order state. (a) Energy structure before the pump pulse arrives. (b) The double structure above the Fermi level appears after photoexcitation and corresponds to the new, long-lived quasiparticles, decaying in density as shown in (c–h) with a characteristic lifetime of 213 fs. Reproduced from Dakovski, G.L., Li, Y., Gilbertson, S.M., *et al.*, 2011. Physical Review B 84, 161103(R).

upconversion techniques). In the past decade, ultrafast trARPES has been extensively used to explore the non-equilibrium behavior of a wide range of materials. Among correlated electron materials, dynamics in HTSC, charge density wave compounds and heavy fermion actinides have been closely investigated.

An excellent example of trARPES capabilities is the investigation of band renormalization processes during the phase transition in URu_2Si_2, one of the most heavily studied heavy fermion compounds due to the peculiar hidden order (HO) state occurring below 17.5 K. Numerous theories has been proposed over the past two decades to describe this state, but the exact nature of the HO order parameter still remains unclear. As described in Section "All-Optical Pump–Probe Spectroscopy of Correlated Electron Materials," ultrafast OPOP experiments identified a pseudogap region (17.5 K$<T<$25 K) separating the normal Kondo-lattice state, with a 10 meV hybridization gap, from the HO phase, with a 5 meV gap. trARPES was then applied to determine the relative positions of both gaps in energy and momentum space, because optical excitation can populate the states above E_F, enabling their observation with a temporally delayed ARPES probe (compare the static and pumped ARPES results in **Fig. 19(a)** and **(b)**, respectively). Time-resolved measurements of the quasiparticle population and decay dynamics revealed a multi-peak structure at the upper edge of the HO gap (**Fig. 19(b–h)**), where quasiparticles accumulate for hundreds of fs due to the phonon bottleneck (see the discussion of the RT modeling of OPOP results in Section "All-Optical Pump–Probe Spectroscopy of Correlated Electron Materials,"). These results suggested that the HO gap forms as a momentum-dependent modification of the momentum-independent hybridization gap, which might explain the contradictory measurements of the transport properties and electronic structure in this compound by other time-averaged techniques.

Ultrafast Second Harmonic Generation Spectroscopy

Nonlinear frequency conversion methods play an important role in optics and laser physics. The conversion efficiency is very sensitive to the properties of the nonlinear medium, since it involves mixing of multiple electromagnetic waves with different

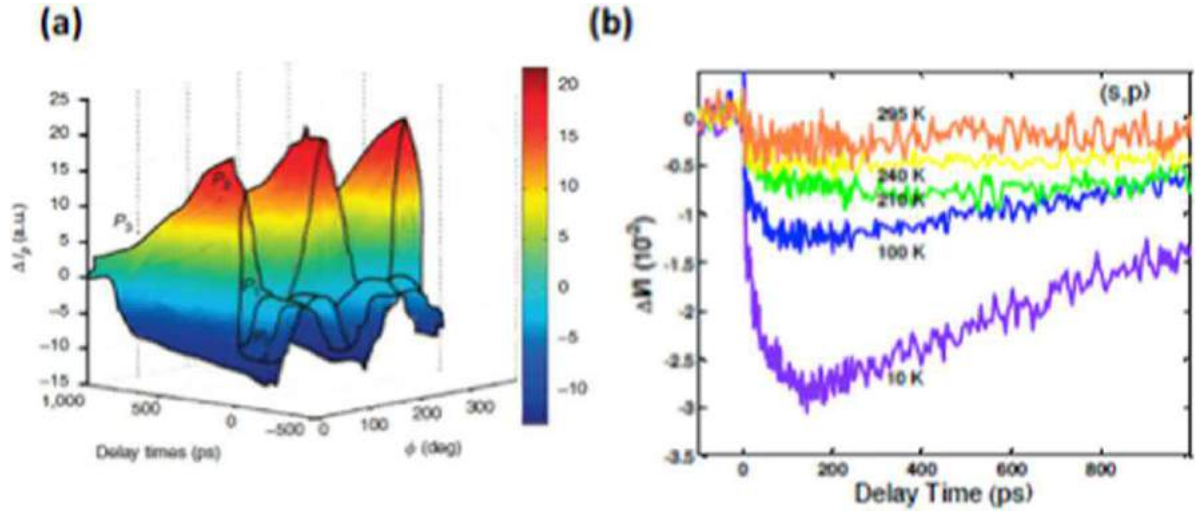

Fig. 20 trSHG signals from a BSTO/LCMO heterostructure as a function of (a) the incident light polarization at 10 K; and (b) temperature, demonstrating ultrafast interfacial ME coupling. Reproduced from Sheu, Y.M., Trugman, S.A., Yan, L., *et al.*, 2014. Nature Communications 5, 5832.

polarizations and propagation directions. In second harmonic generation (SHG), two light fields E_j and E_k of the same fundamental frequency ω are coupled through the nonlinear susceptibility $\chi^{(2)}_{ijk}$, producing a nonlinear polarization $P_i(2\omega)=\chi^{(2)}_{ijk}\,E_j(\omega)E_k(\omega)$ with the ensuing electric fields oscillating at 2ω. The tensor $\chi^{(2)}_{ijk}$ is extremely sensitive to the lattice, electronic and magnetic symmetries of a crystals and disappears when spatial inversion symmetry is present, making SHG a powerful probe of broken symmetry in a wide range of systems.

SHG spectroscopy has been widely used for the characterization of ferroelectric phase transitions, where inversion symmetry breaking by the FE polarization leads to an enhancement of the SHG signal (typically as the temperature is reduced). Moreover, SHG may also depend on magnetic symmetries through higher order nonlinearities or multipole contributions. Extension of SHG methods to the time domain is straightforward and can be done by photoexciting the sample before probing the SHG response, thus providing access to ultrafast dynamics of the electronic, vibrational, magnetic, and crystallographic structure of materials. Time-resolved SHG (trSHG) has been used to characterize non-equilibrium magnetic, orbital, and magnetoelectric responses of CEM. Arguably, the most promising applications of trSHG are as a simple and readily accessible alternative to X-ray scattering techniques for probing structural material dynamics during PIPT.

One example is shown in **Fig. 20**, which illustrates the ability of trSHG to track ferroelectric polarization dynamics caused by concurrent thermal expansion and magnetostrictive coupling across the interface between FM and FE layers in a multiferroic $Ba_{0.1}Sr_{0.9}TiO_3$ (BSTO)/LCMO heterostructure. In these experiments, femtosecond optical pulses were used to selectively excite the LCMO layer, which then underwent e-ph and spin-lattice relaxation within tens of ps (as in the OPTP results described in **Fig. 12** of Section "Ultrafast Infrared Spectroscopy: THz to Mid-IR"). trSHG from the top BSTO layer (**Fig. 20(a)**) exhibited an enhancement of the FE polarization on fast ($\sim$1 ps) and slow ($\sim$10–100 ps) timescales (**Fig. 20(b)**). Interestingly, the amplitude of the large slow component decreased with increasing spin disorder in LCMO as the temperature increased towards the FM T_c. These dynamics were explained as a strain-induced enhancement of the BSTO polarization mediated by (1) rapid thermal expansion of the photoexcited LCMO layer, caused by e-ph energy transfer and (2) slower expansion of the LCMO layer through changes in the magnetostriction that were induced by spin-lattice relaxation and the ensuing spin disorder. This demonstrates that ultrafast interfacial ME coupling in such heterostructures is dominated by elastic coupling across the interface, not the slower heat diffusion from the LCMO to the BSTO layer, and more generally, provides a nice example of the power of trSHG for shedding light on the electronic and structural dynamics of correlated electron materials.

Ultrafast Microscopy

Strong interactions between various degrees of freedom in CEM produce a delicate balance between competing states with very different natures, such as charge or orbital ordered, magnetic, ferroelectric and superconducting. This balance is easily perturbed by intrinsic fluctuations or extrinsic stimuli, leading to electronic, magnetic and structural inhomogeneities at the nano-to-micrometer scales. Often, such non-uniformity serves as a marker of competing interactions or phase coexistence, for example, near phase transitions, and indicates emergent functionality in materials systems. Over the years, many microscopic techniques have been developed to reveal the intimate connection between these nano-to-microscale spatial effects and the materials properties. However, in correlated materials spatial inhomogeneities also lead to nontrivial temporal dynamics (and vice versa), thus demanding tools that directly measure the properties at both relevant length and time scales to identify the role of intrinsic versus extrinsic effects and the extent to which such inhomogeneities determine functionality.

Here we give a brief, and by no means exhaustive, overview of the microscopic techniques that have already been demonstrated to combine ultrafast time-resolved imaging and spectroscopy. Achieving simultaneous sub-picosecond and sub-micron spatial resolution is a serious technical challenge: these 4D (space and time) experiments are usually performed by averaging the signals of a large number of pump–probe pulse pairs to improve the sensitivity, which requires high reproducibility of the photoinduced spatial phase separation from pulse to pulse. The same mechanisms that are responsible for the intrinsic inhomogeneity of correlated materials also make phase separation a very random process, and most of the previous effort focused on 4D methodology development using rather simple test systems, such as metal or semiconductor nanostructures. Nevertheless, initial encouraging applications to CEM (e.g., the canonical MIT in VO_2) have just begun to appear, and the field is bound for rapid growth as more 4D instruments become available to the research community.

The simplest spatially resolved optical technique of all is conventional optical microscopy, which can be extended to both the UV and IR spectral ranges with appropriate focusing optics. The spatial resolution of this technique is restricted by the diffraction limit, thus constraining its applicability to length scales exceeding the probe wavelength ($\sim$1 μm in near-IR region). Broadband reflection microscopy was used to investigate the peculiar striped dimerization patterns in a Mott-Hubbard insulator induced by a local current-driven MIT and metal-insulator phase separation in a correlated superconductor. Optical microscopy is also capable of other spectroscopic modalities besides absorption or reflection measurements. Collection of the SHG signal revealed the ferroelectric domain structure and its correlation with antiferromagnetic order in multiferroic samples. Integration with a DC magnetic field and polarization diagnostics provided a way to measure the effects of large-scale inhomogeneities in the crystalline structure on the ferromagnetic properties of manganites using magneto-optical Kerr effect microscopy (MOKE). Most of these modalities are compatible with time-resolved pump–probe geometries; for example, OPOP microscopy of VO_2 revealed that the timescales governing the photoinduced MIT decrease at higher excitation fluences, as measured from the peak of the focused beam profile to its tail. Ultrafast MOKE microscopy was also used to track the photoinduced magnetization changes in TbFeCo with better than 200 fs/200 nm combined resolution. In this work, it was shown that dipole-dipole interactions at the boundary of the illuminated region enhance the effective magnetic field and lead to local magnetization reversal on sub-nanosecond timescales.

Tuning the probe wavelength to the hard X-ray region brings the spatial resolution down to the sub-micron level when performing X-ray diffraction or spectroscopic measurements. In one approach to X-ray microscopy, the beam is focused to a 10–30 nm spot by various methods, for example, using a zone plate or Kirkpatrick–Baez mirror pair, and raster scanned across the area of interest while the diffraction patterns or scattering spectra are detected at each position. For example, scanning hard X-ray diffraction with 30 nm resolution was exploited to follow the monoclinic to rutile crystalline structure conversion during the MIT in VO_2. Comparison between the temperature variation of the monoclinic phase fraction and insulating phase fraction measured by scanning near-field infrared microscopy (see below) showed that the electronic MIT proceeds monotonically, while the structure tends to oscillate around T_c. In the soft X-ray region, raster imaging was combined with magnetic spectroscopy to visualize switching of AFM order with electric fields in multiferroic $BiFeO_3$. These results demonstrated that AFM order only exists in FE polarized regions of the sample, whereas unpoled areas remain magnetically disordered. Extending these X-ray measurements to the time domain might be challenging, due to photon flux constraints and shot-to-shot instability of the photoinduced crystallographic and spectroscopic inhomogeneities. However, with the development of brighter ultrafast XFEL sources and novel, resonant coherent diffractive (lensless) imaging approaches, it might be possible to perform wide field, rather than raster scanned, measurements of structural and magnetic dynamics in correlated materials, as was recently shown for a non-correlated ferrimagnetic GdFeCo alloy. In that study, coherent X-ray spectromicroscopy of photoexcited GdFeCo revealed an initially inhomogeneous transient distribution of the angular spin momentum, after which the subsequent flow of angular momentum proceeded from the Fe-rich nanoregions to the Gd-rich ones on sub-ps timescales.

Another promising ultrafast wide field microscopy approach is photoemission electron microscopy (PEEM). PEEM records electrons emitted from a sample following the absorption of one X-ray or two visible photons. The electrons are accelerated towards the lens and projected to create a magnified image on the detector, similar to electron microscopes. A spatial resolution of $\sim$10 nm can be routinely achieved, and spin sensitivity added by using linearly or circularly polarized photons. PEEM has been extensively exploited to study the structural, magnetic and electronic properties of CEM. For example, in multiferroic $BiFeO_3$, the combination of PEEM and linear X-ray magnetic absorption dichroism (XMLD) spectroscopy was used to image the AFM domain distribution as a function of electric field applied to the sample. The variable electric field was found to modify the FE polarization and FE domain structure, which was measured with piezoelectric force microscopy (PFM). Comparison between results from XMLD-PEEM and PFM on the same sample area verified the strong coupling between FE and AFM orders in BFO and demonstrated AFM switching with an electric field. Temporal resolution can be added to PEEM by incorporating femtosecond pulses from either femtosecond XFEL or tabletop visible/near-IR laser systems. Ultrafast PEEM has been widely used to image the dynamics within plasmonic nanostructures with better than 10 fs temporal and 50 nm spatial resolution, but has never been applied to characterize CEM, to the best of our knowledge.

In the three decades following their invention, scanning probe microscopies have also provided much insight into the nanoscale properties of correlated materials. The highest spatial resolution can be achieved with scanning tunneling microscopy (STM), which is based on quantum tunneling of electrons through the energy barrier formed by the sharp metallic tip and a sample placed in the tip vicinity. The strong dependence of the current on the tip-sample spacing enables surface topography variations to be mapped with sub-angstrom resolution. The major virtue of STM lies in its ability to probe the local density of states of the sample in a narrow energy band around the Fermi level. For instance, the combination of high-resolution topographic and spectroscopic imaging revealed the non-uniform distribution of the superconducting gap size on the surface of a cuprate

HTSC. The gap inhomogeneity was attributed to oxygen impurity distributions, which suppressed the superconducting phase around them. STM has also demonstrated the percolative nature of the MIT in manganites, where a metallic phase gradually grows within the insulating matrix as the temperature is lowered across the MIT. To add time resolution, one can take advantage of the fact that the tunneling junction is a nonlinear region, where femtosecond electromagnetic pulses can be mixed to produce ultrafast transients in the tunneling current. The amplitude and dynamics of these transients depends strongly on the local material properties, making them a good probe of charge transfer and capture processes in metal and semiconductor nanostructures with sub-ps resolution. Modulation of the voltage bias with single-cycle THz pulses also provides sensitivity to the local electronic structure and molecular resonances while allowing tracking of molecular motion and charge dynamics in semiconductors on sub-nm/ps scales. However, ultrafast STM can only be used on conducting surfaces, due to its reliance on detection of the tunneling current.

Scanning near-field optical microscopies (SNOM) have recently risen to prominence for time-averaged and time-resolved exploration of nanoscale properties of CEM. Sub-wavelength imaging in SNOM is achieved either by using a small waveguide aperture for illumination/collection or by scattering the light from a sharp metallic tip. In the former, both the spatial resolution and spectral bandwidth are severely limited by the waveguide characteristics. On the other hand, the scattering-based technique (sSNOM) relies on propagation of the scattered light through free space, and its resolution is determined by the tip curvature, which can be below 5 nm. Furthermore, sSNOM can take full advantage of the recent breakthroughs in ultrafast broadband source development and perform nanoscale spectroscopy in a spectral range spanning visible to THz frequencies. As described earlier (see Section "Ultrafast Infrared Spectroscopy: THz to Mid-IR"), THz and mid-IR radiation encompass the energies of low-energy modes, energy gaps and transport phenomena in correlated materials and are thus of most interest for sSNOM applications. In VO_2, collapse of the Mott gap during the MIT causes a significant spectral weight shift in the optical conductivity below 3000 cm^{-1}, due to the emergence of the Drude response in the metallic state. This shift was used as a contrast mechanism for sSNOM imaging of

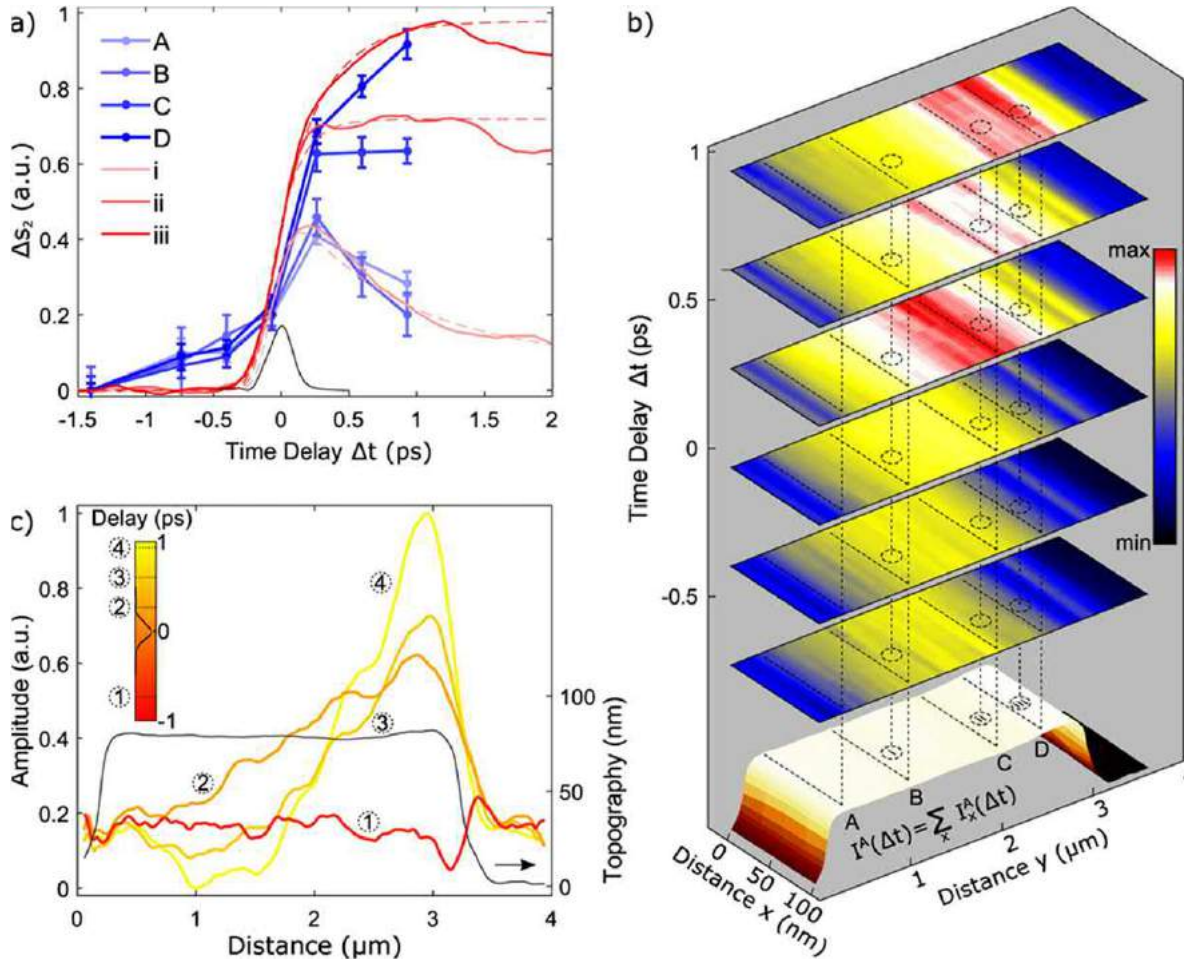

Fig. 21 (a) Time-resolved sSNOM of the photoinduced MIT in VO_2, acquired with a sub-threshold fluence of 2.9 mJ/cm^2 and 1032 nm pump and 4.7 μm probe wavelengths. The red curves were taken at positions (i–iii) marked by circles in (b). The blue time points were extracted from the image series in (b) at the points (A–D). (b) Spatio-temporal images of the metallic phase (red) growth dynamics in a VO_2 microcrystal. (c) Cross-sections of the photoinduced scattering changes in (b) for four time delays across the microcrystal. Reproduced from Donges, S.A., Khatib, O., O'Callahan, B.T., et al., 2016. Nano Letters 16, 3029.

metal–insulator phase inhomogeneity during the MIT. Similar to STM on manganites, sSNOM measurements at 930 cm^{-1} revealed phase coexistence well below and above T_c. Moreover, sSNOM enabled local spectroscopy within an individual metallic island and demonstrated that the emerging metal phase close to T_c is different from the rutile one at higher temperatures, instead representing a strongly correlated metal with a pseudogap-like response in the IR conductivity.

sSNOM is an optical technique and is, therefore, very amenable to incorporating pump–probe measurements with ultrafast lasers. Indeed, the first time-resolved near-field optical microscopic measurements on CEM were recently demonstrated using sSNOM on nanostructured VO_2. VO_2 was excited above its bandgap with 1032 nm pulses of 190 fs duration, and the amplitude of a 4.7 μm, 200 fs probe pulse scattered from the tip was measured as a function of position and time (**Fig. 21(b)**). Although the excitation fluence of 2.9 mJ/cm^2 was below the threshold for the photoinduced MIT, in some areas on the sample the amplitude of the scattering signal did not decay right away, as expected for sub-threshold carrier relaxation. On the contrary, the scattering signal continued to rise until saturating within a few ps, indicating the formation of metallic regions with enhanced scattering (curves (ii-iii) and red regions in **Fig. 21(a)** and **(b)**, respectively). Interestingly, this incredible variety of dynamics, ranging from sub-threshold carrier excitation to a saturated MIT phase transition, was observed just a few hundreds of nm apart. Previous OPTP results (see Section "Ultrafast Infrared Spectroscopy: THz to Mid-IR") revealed softening of the photoinduced MIT, due to an increasing number of metallic nucleation sites as the temperature approached T_c. Similarly, the localized sub-threshold induction of MIT revealed by sSNOM could be attributed to the presence of a metal-rich phase in these regions, caused by inhomogeneous doping or strain in nominally homogeneous VO_2 crystals. These results demonstrate the feasibility of time-resolved microscopic studies of CEM, even in a multi-shot-averaging geometry, thus encouraging further advances in the field.

Further Reading

Time-integrated and time-resolved optical experiments on correlated electron materials:

Averitt, R.D., Taylor, A.J., 2002. Ultrafast optical and far-infrared quasiparticle dynamics in correlated electron materials. J. Phys. Condens. Matter 14, R1357.
Basov, D.N., Averitt, R.D., van der Marel, D., Dressel, M., Haule, K., 2011. Electrodynamics of correlated electron materials. Rev. Mod. Phys. 83, 471.
Demsar, J., Sarrao, J.L., Taylor, A.J., 2006. Dynamics of photoexcited quasiparticles in heavy electron compounds. J. Phys. Condens. Matter 18, R281.
Giannetti, C., Capone, M., Fausti, D., *et al*., 2016. Ultrafast optical spectroscopy of strongly correlated materials and high-temperature superconductors: A non-equilibrium approach. Adv. Phys 65 (2), 58.
Hilton, D.J., Prasankumar, R.P., Trugman, S.A., Taylor, A.J., Averitt, R.D., 2006. On photo-induced phenomena in complex materials: Probing quasiparticle dynamics using infrared and far-infrared pulses. J. Phys. Soc. Jpn. 75, 011006.
Mankowsky, R., Forst, M., Cavalleri, A., 2016. Non-equilibrium control of complex solids by nonlinear phononics. Rep. Prog. Phys. 79, 064503.
Nasu, K., Ping, H., Mizouchi, H., 2001. Photoinduced structural phase transitions and their dynamics. J. Phys. Condens. Matter 13, R693.
Orenstein, J., 2012. Ultrafast spectroscopy of quantum materials. Physics Today 65, 44.
Prasankumar, R.P., Taylor, A.J. (Eds.), 2011. Optical Techniques for Solid-State Materials Characterization. San Francisco, CA: Taylor & Francis.
Wegkamp, D., Stahler, J., 2015. Ultrafast dynamics during the photoinduced phase transition in VO_2. Prog. Surf. Sci. 90, 464.
Zhang, J., Averitt, R.D., 2014. Dynamics and control in complex transition metal oxides. Annu. Rev. Mater. Res. 44, 19.

General background on correlated electron materials:

Chakhalian, J., Freeland, J.W., Millis, A.J., Panagopoulos, C., Rondinelli, J.M., 2014. Emergent properties in plane view: Strong correlations at oxide interfaces. Rev. Mod. Phys. 86, 1189.
Dagotto, E., 2003. Nanoscale Phase Separation and Colossal Magnetoresistance. Berlin: Springer.
Degiorgi, L., 1999. The electrodynamic response of heavy-electron compounds. Rev. Mod. Phys. 71, 687.
Imada, M., Fujimori, A., Tokura, Y., 1998. Metal–insulator transitions. Rev. Mod. Phys. 70, 1039.
Keimer, B., Kivelson, S.A., Norman, M.R., Uchida, S., Zaanen, J., 2015. From quantum matter to high-temperature superconductivity in copper oxides. Nature 518, 179.
Khomskii, D., 2001. Electronic structure, exchange, and magnetism in oxides. In: Thornton, M.J., Ziese, M. (Eds.), Lecture Notes in Physics, vol. 569. Berlin: Springer-Verlag, p. 89.
Scott, J.F., 2007. Applications of modern ferroelectrics. Science 315, 954.
Shuvaev, A.M., Mukhin, A.A., Pimenov, A., 2011. Magnetic and magnetoelectric excitations in multiferroic manganites. J. Phys. Condens. Matter 23, 113201.
Spaldin, N.A., Cheong, S.-W., Ramesh, R., 2010. Multiferroics: Past, present, and future. Physics Today 63, 38.
Tokura, Y., 2003. Correlated-electron physics in transition-metal oxides. Physics Today 56, 50.
Tokura, Y., 2000. Colossal Magnetoresistive Oxides. CRC Press.

Tutorial on Multidimensional Coherent Spectroscopy

Mark Siemens, University of Denver, Denver, CO, United States

Introduction: The Power of 2DCS

Multidimensional spectroscopy measures the coherent response of a system and displays it as a function of two (or more) frequencies, corresponding to resonances such as electron absorption or vibrational excitation in the system. This multidimensional map of the frequency response is extremely helpful for disentangling congested spectra, revealing transport pathways, and measuring quantum coherences in heterogeneous systems. Even better, the fundamental understanding enabled by this multidimensional approach is not limited to a specific physical system, but can provide a comparison with, or in some cases a measurement of, the Hamiltonian for interaction dynamics for very different resonances. For example, in the last 25 years coherent multidimensional spectroscopy has spread across the electromagnetic spectrum:

- ~1970–*present*: At radio frequencies, multidimensional nuclear magnetic resonance (NMR) measures nuclear spin couplings in molecules to reveal the structure of proteins as they perform their physiological functions;
- 1993–*present*: In the infrared, two-dimensional infrared (2DIR) is sensitive to molecular vibrations, allowing direct measurement of hydrogen-bond dynamics in water (Fecko *et al.*, 2003) and vibrational couplings in DNA (Krummel *et al.*, 2003);
- 2003–*present*: Visible and near-infrared light enables multidimensional coherent spectroscopy of electronic resonances, revealing quantum transport pathways in photosynthetic complexes (Collini and Scholes, 2009; Brixner *et al.*, 2005) and many-body dynamics in semiconductor nanostructures (Li *et al.*, 2006; Singh *et al.*, 2016); and
- 2011–*present*: Terahertz two-dimensional spectroscopy directly measures solid-state lattice vibrations (phonons) and couplings to electrons.

While three-dimensional or even higher-order spectroscopies have been demonstrated, two-dimensional coherent spectroscopy (2DCS) provides enormous new physical insight compared to one-dimensional spectroscopies, and it is by far the most common implementation of coherent multidimensional spectroscopy. Higher-order spectroscopies have the same theoretical underpinnings and use similar, though more complicated, experimental setups, so this discussion of 2DCS is relevant to those as well.

Fig. 1 provides a schematic comparison between 2D and one-dimensional (1D) spectroscopies that highlights why 2DCS is such a powerful tool. A traditional 1D absorption spectrum measurement, as shown schematically in **Fig. 1(a)**, can resolve the number and center frequencies of resonances, but cannot determine the nature of coupling, broadening, or higher-order states within the resonances. In contrast, the 2D spectra in **Fig. 1(b)–(e)** show significantly more information and provide quick identification of the physics and active transport pathways in a system.

A 2D spectrum provides a 2D response function, showing how the likelihood that excitation at a certain input/absorption frequency (labeled by the vertical axis) will lead to emission at a particular frequency. For isolated resonances, as shown in **Fig. 1(b)**, all peaks lie along the diagonal line (highlighted in green) that specifies equal input and output frequency. If local environmental fluctuations affect the resonance frequencies, inhomogeneous broadening leads to a significant elongation of the signal along the diagonal of the 2D spectrum, as shown in **Fig. 1(c)**. Off-diagonal peaks indicate coupling, as shown in **Fig. 1(d)**, and careful analysis can determine the coherence and strength of the coupling. **Fig. 1(e)** shows that 2DCS is also well-suited to identifying and measuring multiply-excited state dynamics such as resonance anharmonicities, which show up as a negative peak (negative because of the different quantum pathways sampled) shifted to the left of the primary absorption resonance.

2D coherent spectra provide a direct identification and visualization of quantum pathways, providing quick identification of the physics in a system and even quantitative measurement of essential system parameters. A detailed guide to interpreting the features in 2D spectra is provided in Section Interpreting 2D Spectra.

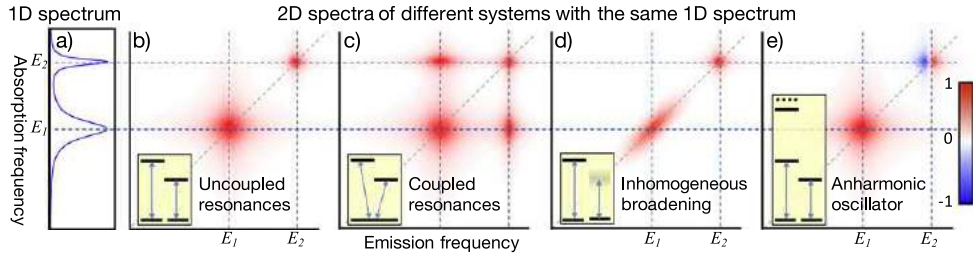

Fig. 1 Schematic comparison of one-dimensional (1D) and 2D spectroscopy. The 1D absorption spectrum shown in panel (a) could correspond to many different physical situations; 2D spectra on the right can resolve these. Yellow insets show the energy level diagrams corresponding to the unique 2D spectrum shown: (b) Uncoupled resonances, (c) inhomogeneous broadening, (d) coupled resonances, and (e) anharmonic oscillator. Interpretation of 2D spectra and implications for each of these physical situations will be discussed in detail in Section Interpreting 2D Spectra. Real part of 2D absorption spectra are shown, scaled by the colorbar on the right.

 doi:10.1016/B978-0-12-803581-8.09526-6

Theory of Coherent Multidimensional Spectroscopy

The development of 2D spectroscopy draws not only from nonlinear optics and coherent spectroscopy basics, but also from the historical context of the different physical systems studied. Chemists studying molecular vibrations and conformations, physicists studying electron dynamics in semiconductor nanostructures, and biologists studying exciton transport in photosynthetic complexes have differences in presentation and notation, and the relative importance of different features in a 2D spectrum can vary based on the nature of the resonance (electron, molecular vibration, spin, etc.) being studied. However, the underlying theoretical understanding of 2DCS experiments is based on the same physical principles.

Fourier Transform Spectroscopy

The most obvious way to collect a 2D coherent spectrum would probably be a "double-resonance experiment" performed by sweeping the frequency of two excitation beams and recording the signal amplitude as a function of the frequency of both beams: $S(\omega_1,\omega_2)$. However, the sensitivity of this approach is very low because of the small field amplitudes involved in a nonlinear optical process, so it has been superseded by time-domain methods using a series of intense laser pulses. If these pulses have controllably-stepped delays between them, then the time-domain signal can be Fourier-transformed to yield a spectrum. This "Fourier transform spectroscopy" approach for multidimensional spectroscopy was first proposed for NMR in 1971 by Jeener before being implemented by Ernst (1992), and is now widely used for 2DCS experiments throughout the electromagnetic spectrum.

The principle of Fourier transform spectroscopy is that a spectrum can be obtained by Fourier transforming a coherent signal measured in the time domain. The simplest example of an experiment that does this is a time-resolved double-pump-probe measurement, as shown in **Fig. 2(a)**, in which three laser pulses are overlapped on a sample with a variable time delay τ between them. If the complex (both amplitude and phase) coherent signal is measured as a function of the delay time τ between pump pulses (**Fig. 2(b)**), then the discrete time-domain response can be Fourier transformed to yield the 1D spectrum (**Fig. 2(c)**) – which is identical to what a spectrometer would measure.

One-dimensional Fourier transform spectroscopy can be used to measure spectra with very high resolution, at wavelengths for which conventional spectrometers are unavailable, and with higher signal-to-noise through time-domain noise suppression techniques such as modulation and lock-in amplification. However, the exciting advance that Ernst pushed forward in the 1970s for NMR (Ernst *et al.*, 1988) is that the method is not limited to one dimension; extended pulse sequences can measure coherent signals from two or even more time dimensions that can be independently Fourier transformed to yield multidimensional spectral maps.

Double-Sided Feynman Diagrams

One of the strengths of 2DCS is that different pulse sequences can be used to study different quantum pathways in a system. In fact, the measured third-order signal from a particular pulse sequence usually results from multiple quantum pathways, and so accurate interpretation of 2D spectra relies on understanding exactly which quantum pathways are contributing. There are various bookkeeping tools available to write down and survey the possible quantum pathways (including ladder diagrams, which are often used in describing Raman spectra), but one that has gained wide usage across the range of 2DCS experiments is the double-sided Feynman diagram. Double-sided Feynman diagrams are particularly powerful because they allow the state coherence to be tracked over time, including changes resulting from photon absorption and emission (Boyd, 2003).

Here is a brief guide to interpreting double-sided Feynman diagrams (FDs). For the following discussion, we will assume a system described by a three-level energy ladder - perhaps corresponding to the first three states of an anharmonic oscillator – with ground-state $|0\rangle$ and excited states $|1\rangle$ and $|2\rangle$, as shown in the gray box in **Fig. 3**. The following discussion points match the panels in **Fig. 3** that illustrate the basics of double-sided Feynman diagrams graphically.

1. The time evolution of the system is tracked by the parallel vertical lines, with the left side indicating the "ket" side of the density matrix ρ (Eq. 1) and the right side representing the "bra." The initial state of the system is at the bottom, and time increases upward. As is standard in the "bra-ket" notation, a population is indicated when the "bra" and "ket" are the same state (e.g., $|a\rangle\langle a|$), otherwise a coherent superposition between states is specified (such as $|a\rangle\langle b|$). Both populations and coherences can evolve with time under the system Hamiltonian, until another optical pulse arrives and changes the state.

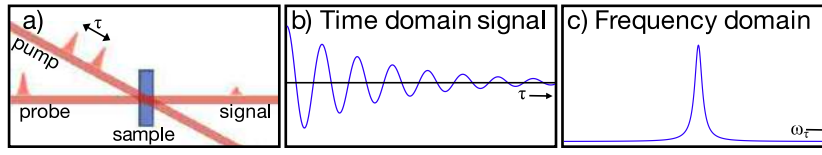

Fig. 2 (a) Schematic depiction of 1D pump-probe Fourier transform spectroscopy. (b) Real part of the coherent signal in the time domain. (c) Real part of the signal in the frequency domain, as measured by a spectrometer or a Fourier transform of the time-domain signal.

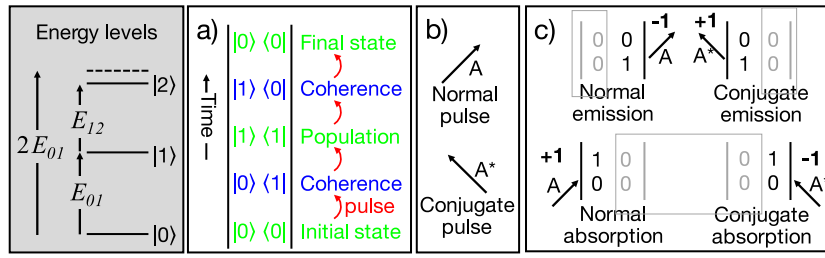

Fig. 3 An introduction to double-sided Feynman diagrams, see details in the text. Gray panel: Schematic 3-level energy ladder. Depending on the system, $E_{01} \neq E_{12}$. (a) Example double-sided Feynman diagram, showing that time increases upward. (b) The direction of normal and conjugate-acting pulsed electro-magnetic fields. (c) How to track absorption and emission of normal and conjugate pulses in a double-sided Feynman diagram. The bold $+1$ or -1 beside each figure tracks the sign of the interaction, which is determined by whether the field interacts with the "bra" or the "ket." Reproduced from Boyd, R.W., 2003. Nonlinear Optics. Academic Press.

2. Photon interactions are represented by diagonal arrows whose orientation indicates the action of the "jth" radiation field on the system: tipped-left indicates a "conjugate" field interaction contributing as $E_j^* = \varepsilon_j^* \exp\left[i\omega_j t - i\mathbf{k_j} \cdot \mathbf{r}\right]$, while tipped-right means a "normal" field contribution of $E_j = \varepsilon_j \exp[-i\omega_j t + i\mathbf{k_j} \cdot \mathbf{r}]$. This distinction is necessary because the light-matter interaction Hamiltonian is $\hat{H}(t) - \hat{\mu}E$, where $E = E_j + E_j^*$, and either the conjugate or non-conjugate portion of the field could act on the system. In the literature, these arrows are labeled in a variety of ways: by photon momentum, photon energy, frequency, or just pulse number.
3. Since an arrow on the left (right) means that the field interacts with the ket (bra) of the density operator, an arrow's pointing towards or away from the FD shows how a photon is transferring energy to the system: the system absorbs a photon if the arrow points towards the diagram, and an arrow pointing away from the diagram describes photon emission. Photon absorption acts as a raising operator and photon emission acts as a lowering operator on the corresponding side of the density matrix, so the ket or bra state of the system changes accordingly in response to a photon interaction, and is recorded on the next line up in the diagram.

Careful use of Feynman diagrams allows for all possible quantum pathways to be accounted for and for relevant coherences to be tracked, enabling a direct way to determine expected peak locations in a 2D spectrum. This is essential for proper interpretation of the physical meaning of measured coherent spectra.

Theoretical Signal Origin

2DCS experiments measure the third-order response of a system of dipoles to applied electromagnetic fields, and an understanding of the theoretical origin of the signal is important for interpreting 2D spectra. Why is third-order coherence studied so heavily? The first-order (linear) response describes traditional (linear) spectroscopies like absorption and photoluminescence and the second-order response function is zero in an isotropic medium due to symmetry (Boyd, 2003), so third-order is the dominant term in a nonlinear power-series expansion of the polarization with the incident field. A third-order polarization will result from the system response to three excitation fields (Boyd, 2003), which could all be from the same pulse, but a three-pulse experiment will provide the most control over the state dynamics of a system.

While quantum mechanical descriptions of light interacting with a pure system wavefunction ψ can be sufficient to describe gas-phase coherence, condensed-phase systems involve statistical ensembles of molecules rather than pure states, and so the density matrix of an ensemble ρ is used instead of ψ:

$$\rho = \sum_s p_s |\psi_s\rangle\langle\psi_s| = \begin{pmatrix} \rho_{00} & \rho_{01} \\ \rho_{10} & \rho_{11} \end{pmatrix} \tag{1}$$

with p_s being the probability of a system being in state $|\psi_s\rangle$. The second half of Eq. (1) assumes a two-level system with states $|0\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $|1\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, in which case ρ_{00} is the average population in state $|0\rangle$, ρ_1 is average the population in state $|1\rangle$, and $\rho_{10} = \rho_{01}{}^*$ is the coherence between states $|0\rangle$ and $|1\rangle$.

The evolution of a system $\rho(t)$ during and after optical excitation can be calculated with the standard optical Bloch equations (OBEs). At this point, different approaches are pursued depending on the system and type of excitation. For example, strong "π-pulse" excitation in NMR experiments generates pure populations or coherences, while electronic and vibrational spectroscopies in the visible and infrared are safely in the "weak-excitation" limit and so perturbation theory should be applied (Boyd, 2003). Using perturbation theory and assuming the short-pulse (delta-function assumption) limit, the system evolution can be tracked over time as a time-ordered series of independent steps consisting of three different types:

- optically-induced transitions between population/coherence states in the system, which raise or lower the "bra" ("ket") index in the density operator according to the Feynman diagram rules, and multiply the signal by $+(-)i\mu_{01}$;

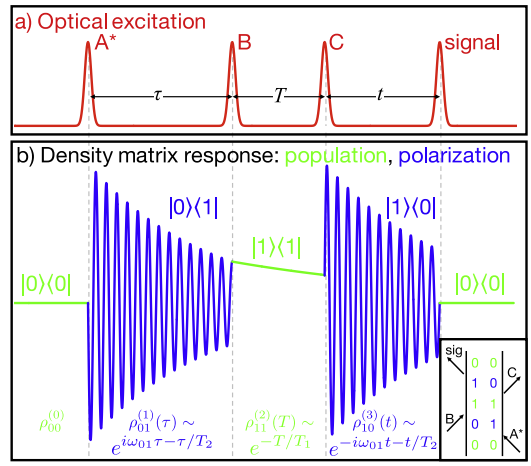

Fig. 4 Tracking population and coherence for a particular quantum pathway. (a) Optical excitation sequence. (b) Coherent system response, with each order of perturbation labeled and the resulting signals tracked. This quantum pathway corresponds to the double-sided Feynman diagram in the inset at lower-right.

- a state $|i\rangle\langle i|$ will evolve with time as an exponentially decaying population e^{-T/T_1}, where T_1 is the population decay time;
- the coherence $|i\rangle\langle j|$ leads to oscillation at a frequency of ω_{ij} and decay at the dephasing rate γ_{ij}, yielding a signal contribution of $e^{-i\omega_{ij}t}e^{-\gamma_{ij}t}$. The inverse of the dephasing rate is the characteristic homogeneous decay time $T_2 \equiv 1/\gamma$, which is frequently used in the literature.

For each "order" of perturbation theory, there will be an additional field interaction introduced in the process, and the constituent step responses can be multiplied to yield the total signal, as we will see shortly.

Double-sided Feynman diagrams can be incorporated with the results from perturbation theory, to calculate the expected signal for each quantum pathway. This is illustrated in **Fig. 4**, which illustrates the populations and polarizations involved in the quantum pathway shown in **Fig. 3(a)**. This quantum pathway leads to a third-order density matrix $\rho^{(3)}_{\text{signal}}(t)$:

$$\rho^{(3)}_{\text{signal}}(t) = \underbrace{\underbrace{\underbrace{(-i\mu_{01})_{A^*}\left(e^{-i\omega_{01}\tau}e^{-\gamma_{01}\tau}\right)}_{\rho^{(1)}_{01}(\tau)}(i\mu_{01})_B\left(e^{-T/T_1}\right)}_{\rho^{(2)}_{11}(\tau+T)}(-i\mu_{01})_C\left(e^{-i\omega_{01}t}e^{-\gamma_{01}t}\right)}_{\rho^{(3)}_{01}(\tau+T+t)} \tag{2}$$

In this equation and in **Fig. 4**, optical interactions from pulses A, B, and C are colored red, coherent oscillations and decays are blue, population dynamics are in green, and the first, second, and third-order perturbation theory steps building to this result are specified with superscript numbers in parenthesis.

The density matrix ρ is a powerful tool for tracking the coherence and population states in a system; and it can be connected to the macroscopic polarization of the system $P(t)$, which radiates to generate the measured signal. Specifically, the macroscopic polarization can be calculated by taking the expectation value of the density matrix multiplied by the dipole operator $\hat{\mu}$:

$$P(t) = \langle \rho_{10}(t) \rangle = \text{Tr}[\hat{\mu}\rho(t)] \underbrace{\rightarrow}_{\text{3rd-order}} P^{(3)}(t) = \langle \rho^{(3)}_{10}(t) \rangle \tag{3}$$

where the right-hand side is specialized the third-order polarization resulting from perturbation theory.

While three-pulse experiments provide tremendous control over the quantum pathways probed by providing the three fields that will be applied, quantum pathways corresponding to a given pulse acting more than once are also possible. However, these alternative quantum pathways can be eliminated from the measured signal through quantum pathway selection.

Quantum Pathway Selection in 2D Coherent Spectra

There is inevitably a huge number of possible quantum pathways available to a multi-level system, and measuring all of them at once would only confuse and congest the measured spectrum. Polarization and pulse sequence manipulation provide enhanced control over which pathways are allowed, which is one of the strongest aspects of 2DCS. Selection is possible because each quantum pathway specifies a particular ordering of quantum states that is only possible with the appropriate fields. In addition to this quantum pathway control, experimental techniques such as phase-matching, frequency-tagging, and phase-cycling can

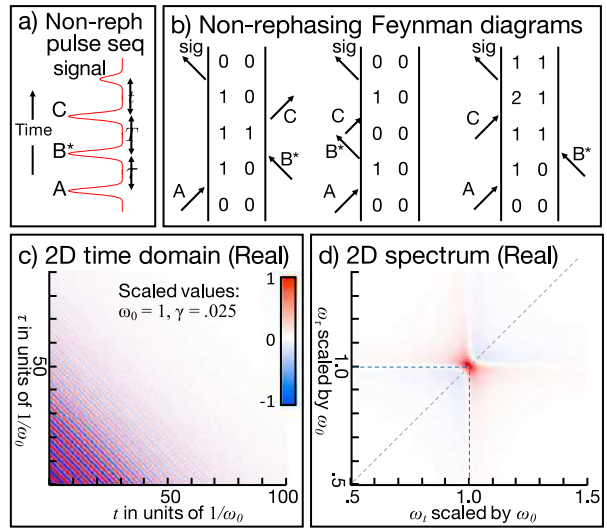

Fig. 5 Summary of non-rephasing 2D spectroscopy: (a) pulse sequence, (b) double-sided Feynman diagrams for active quantum pathways, (c) real part of the 2D time domain signal, (d) 2D spectrum.

provide even more pathway discrimination by further narrowing which pathways are detected (even though all allowed pathways are active).

Polarization Control of Quantum Pathways in 2DCS Measurements

Control over the polarization of each of the incident pulses in a 2DCS experiment is often used to enhance or suppress certain quantum pathways. Polarization affects different kinds of resonant systems in different ways – for example, it couples with the relative angles of transition dipoles in molecules (Zanni *et al.*, 2001b), but it is related to optical selection rules in electron spectroscopy (Singh *et al.*, 2016). Controlling the polarization of excitation pulses can enhance or diminish the prevalence of each quantum pathway, and it can even be used to fully eliminate specific pathways; this technique is most often applied to eliminate the strong diagonal peaks from a 2D spectrum, which allows for weaker cross-peaks that are related to interesting transport and coherence dynamics to be more clearly resolved (Zanni *et al.*, 2001b).

Pulse Sequencing to Control Quantum Pathways in 2DCS Measurements

Different pulse sequences can also excite fundamentally-different quantum pathways in systems, and allow for studying different sorts of dynamics. Here we discuss the three most common types of 2D spectra in this article: Rephasing, Non-rephasing, and Double quantum.

Pulse sequence I: Rephasing 2D spectra

Rephasing 2D spectroscopy, also known as photon-echo spectroscopy, is the most common form of 2DCS. In this configuration, the pulses arrive at the sample with the conjugate pulse acting first: A^*–B–C. This is the well-known "photon echo" scheme from transient four-wave mixing, so-called because an inhomogeneously-broadened system will emit a signal at a time t equal to the A^*–B delay time τ, as will be discussed in more detail in Section Interpreting 2D Spectra of Inhomogeneously-Broadened Systems. We will see that rephasing spectroscopy is very powerful because the inhomogeneous and homogeneous broadening act along the $(t+\tau)$ and $(t-\tau)$ dimensions respectively, which are orthogonal directions in the 2D time domain. This orthogonality of the homogeneous and inhomogeneous broadening is also evident in the 2D spectra: the linewidth along the diagonal is dominated by inhomogeneous broadening, while the cross-diagonal linewidth indicates the homogeneous broadening.

Some authors plot rephasing spectra with the ω_τ axis negative and pointing down (toward larger negative numbers). This convention is used to emphasize that the vertical spectral dimension corresponds to a conjugate pulse. In this case, the "diagonal" direction on the 2D spectrum is down and to the right.

Pulse sequence II: Nonrephasing 2D spectra

In a non-rephasing 2D spectroscopy experiment, the second pulse B is designated as the conjugate pulse, so the pulse sequence on the sample is A–B^*–C. As shown in **Fig. 5(b)**, this excites fundamentally-different quantum pathways than the rephasing scheme. Scanning A before B^* is equivalent to a "$-\tau$" scan, i.e., continuing a rephasing scan to "negative times" with the non-conjugate pulse now first. Homogeneous and inhomogeneous broadening act along the same $(t+\tau)$ direction in the non-rephasing

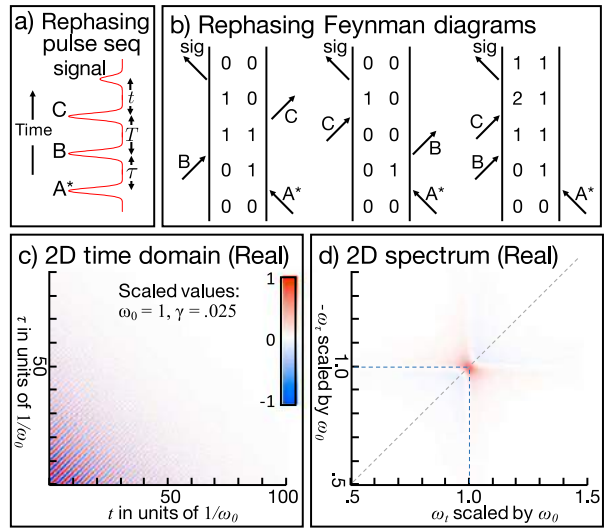

Fig. 6 Summary of rephasing 2D spectroscopy: (a) pulse sequence, (b) double-sided Feynman diagrams for active quantum pathways, (c) real part of the 2D time domain signal, (d) 2D spectrum.

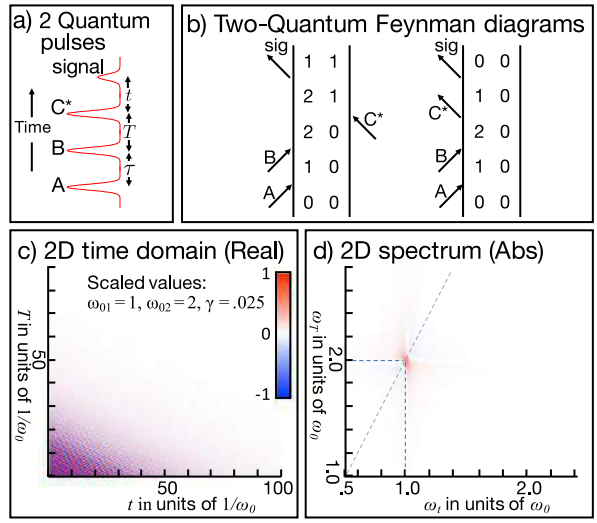

Fig. 7 Summary of double quantum 2D spectroscopy: (a) pulse sequence, (b) double-sided Feynman diagrams for active quantum.

configuration, so the two causes of broadening are not separable in a 2D non-rephasing spectrum. For this reason, stand-alone non-rephasing 2D coherent spectroscopy is much less common than rephasing.

While non-rephasing spectra are not common on their own, they are very powerful when combined with rephasing spectra. Rephasing and non-rephasing results can be collected together by scanning τ through both positive and negative values, which results in a combined 2D time picture, which could be seen here by flipping the 2D time plot in **Fig. 5(c)** and placing it below (**Fig. 6(c)**). Fourier transforming the sum of rephasing and non-rephasing spectra leads to "absorptive lineshapes," which are important for two reasons: first, the resonances are narrower and so are easier to resolve in congested spectra, and second, the real part is positive for a single resonance lineshape (Khalil *et al.*, 2003).

Pulse sequence III: Double quantum 2D spectra

Assuming that the system starts in the ground state, rephasing and non-rephasing spectra measure dephasing dynamics between the ground and first excited states, as well as the coherence between the first and second excited states. These spectroscopies are therefore measuring the dynamics of transitions excited by a single photon, i.e., "single-quantum." Coherences between states that require multiple excitations are accessed and probed if the rephasing pulse is the last of the three excitation pulses to arrive, i.e., the pulse sequence is A–B–C^*. In this case, shown in **Fig. 7(a)**, if the time T between pulses B and C is scanned and Fourier transformed, the coherence studied is $|c\rangle\langle a|$. This is the non-radiative coherence between the ground state and the doubly-excited state, and so these 2D spectra are often referred to as "double quantum."

Quantum Pathways in 3DCS and Other Higher-Order Spectroscopies

While two-dimensional coherent spectroscopy provides significant advantages over one-dimensional spectra, recent work has shown that going to three-dimensional coherent spectroscopy (3DCS) or even higher dimensions provides even more separation of quantum pathways (Li *et al.*, 2013; Cundiff, 2014; Hamm, 2006). By analogy with 2DCS, which requires measurements of a third-order coherent signal over *two* time delays in which the system is in a coherent superposition state, 3DCS usually requires *three* time delays over coherence times (When there is a Raman coherence between states (Li *et al.*, 2013), the coherent oscillation between states during T between pulses A and B can be measured, scanned in time, and subsequently Fourier transformed. In this case, 3DCS can be performed with a third-order signal.). A glance at the Feynman diagrams shows that this cannot occur for three-pulse excitation, so a fifth-order process must be used. However, 2D experimental setups equipped with three delay-controlled pulses can still perform 3DCS measurements, as long as the signal is detected in a direction that accounts for fifth-order quantum pathways that are allowed when each pulse acts act more than once.

Experimental Implementation: 3-Pulse Transient Four-Wave Mixing

While 2D spectra are reported in the frequency domain, they are usually recorded by excitation of a sample with a carefully-controlled sequence of pulses, measurement of a nonlinear coherent signal, and Fourier transformation of the signal with respect to pulse delay times to obtain the frequency response. Therefore, the usual techniques of nonlinear optics and coherent spectroscopy are very applicable.

2DCS is an extension of a time-resolved four-wave mixing (FWM) experiment, in which excitation wave pulses A, B, and C act on a sample to produce a signal wave pulse. The direction of the outgoing signal is set by momentum conservation and any of these pulses can also act as a conjugate pulse (indicated with a star *) on the opposite side of the density matrix, so that $k_{\text{signal}} = \pm k_A \pm k_B \pm k_C$. While quantum pathways corresponding to all possible combinations of normal and conjugate pulse sequences occur, having the pulses incident from different directions allows measurement selection of a particular set of quantum pathways by looking at the signal in a particular direction. A similar result can be obtained by frequency-tagging each excitation pulse and extracting a signal at the beat note corresponding to the desired quantum pathways (Tekavec *et al.*, 2007).

Usually, one pulse (say, pulse "A," but it could be any pulse) is designated as the conjugate pulse, leading to a signal generated in the direction $k_{\text{signal}} = -k_A^* + k_B + k_C$ (There is nothing unique about the light in the conjugate-designated pulse, but by selecting only the signal in the momentum-matched direction for that pulse being conjugated allows selection of just the quantum pathways in which the designated pulse (and only that pulse) acts as a conjugate.). Incidence directions of pulses A, B, and C can be chosen such that the signal is generated in a direction that selects quantum pathways of interest and is easy to align to. One common scheme is to use a "box geometry" as shown in **Fig. 8**, in which A, B, C, and a reference beam are aligned to propagate parallel to each other at the corners of a box. If these beams are then focused onto the sample with a centered lens, the signal will be generated in the same direction as the "reference," which provides a guide for alignment (Bristow *et al.*, 2009). Another common technique uses the "pump-probe" geometry, in which the first two pulses are incident from the same direction, i.e., $k_A = k_B$ (Shim and Zanni, 2009), so the signal is generated along the direction of the probe pulse C: $k_{\text{signal}} = -k_A^* + k_B + k_C = k_C$. While this "pump-probe" implementation of 2DCS is much simpler to set up (especially if adaptive optics are used to control the time delay τ between pulses A and B with intrinsic phase stability), this wavevector degeneracy for pulses A and B provides less selectivity over quantum pathways. On the other hand this non-selection can actually be an advantage: $k_A = k_B$ means that rephasing and non-rephasing spectra are both phase-matched along the k_C direction and so the coherent signal in this direction is the sum of rephasing and non-rephasing pathways; this means that absorption spectra are automatically collected (Shim and Zanni, 2009).

In a time-resolved four-wave mixing experiment, the signal is measured as a function of the delay between two of the pulses, usually the first two pulses A and B. One could imagine that time-resolving an additional time between pulses and then

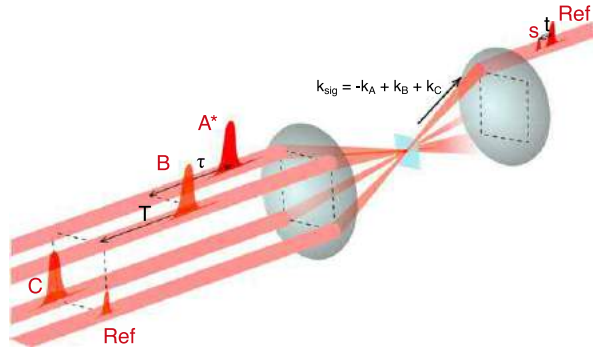

Fig. 8 Common setup for 2DCS that allows full control over all three time delays τ (between pulses A and B), T (between pulses B and C) and t (between pulse C and the output signal). The output can be sent to a spectrometer to directly resolve the output frequency ω_t while τ or T are scanned, or it can be measured with a photodiode while both t and τ are scanned.

performing a 2D Fourier transform would enable 2D coherent spectroscopy, but two additional enhancements are required to enable robust Fourier transformation of pulse delays:

1. Full phase information of the signal is obtained by enforcing interferometric stability between the pulses. Small thermal drifts or acoustic vibrations in optic mounts lead to changes in the signal phase during the course of a time-scanning measurement that cause a distortion of the measured 2D signal after Fourier transformation. A number of phase stabilization methods have been demonstrated, which are suitable for different cases based on various advantages and disadvantages. One of the first methods used common optics for each pulse, and delays were provided by changing the amount of glass in one beampath via translating glass prisms, but this technique was limited to pulse delays ≤ 5 picoseconds (Brixner *et al.*, 2005). Slightly longer delays are possible with intrinsically-phase stable pulsetrains generated by a computer-controlled phase ramp across the spectrum of an incident pulse, and propagated through common optics (Shim and Zanni, 2009). Active methods use mirrors that can be moved by piezoelectric stages to lock each pulse delay to the desired phase; these systems have the advantage of allowing very long delays (few nanoseconds), at the expense of a more complicated setup (Bristow *et al.*, 2009). In a recently-developed technique, the phase is not actively locked, but phase fluctuations are measured and removed from the signal by frequency-tagging each of the excitation pulses (Tekavec *et al.*, 2007).

2. The full complex signal must be recorded as a function of time delays between the excitation pulses. For a time-resolved output signal, this can be done with a lock-in amplifier that can resolve the real and imaginary parts of the signal. Alternatively, the complex signal can be spectrally-resolved in a spectrometer by interfering the signal with a known-phase reference beam with a technique called spectral interferometry (Lepetit *et al.*, 1995); the resulting interference fringes as a function of frequency allow the complex signal to be recovered. Either of these techniques provides the correct relative phase for the signal at all delay times and is sufficient for measuring amplitude 2D spectra, but obtaining the correct absolute phase to determine the real and imaginary parts of a 2D spectrum requires a separate measurement of the absolute phase.

Under these conditions, the phase evolution during different times can be correlated by taking a multidimensional Fourier transform.

Resolution in 2D Spectroscopy

As with any spectroscopic technique, resolution is an important experimental consideration in 2DCS. For a 2D coherent spectrum, different factors can determine the resolution along each spectral dimension. If a spectrometer is used to spectrally-resolve the emission energies (ω_t), then the resolution is determined by the spectrometers slit width, grating constant, and length as usual for spectrometer-resolved measurements. When frequency information is obtained by time-resolving and Fourier-transforming the signal, as is usual for obtaining the absorption spectrum axis (ω_τ) but can be used for resolving emission energies as well, the spectral resolution is determined by the range of the time delay τ_{max}:$\Delta\omega_{\tau,res} = 1/\tau_{max}$.

Running long scans is both technically-challenging and time-consuming, so the optimal length of the time delay is determined by the resolution required by the goals of the experiment. For example, for signals with a long-lived coherence such as a tightly-confined exciton in a quantum dot, a short and fast scan would suffice for resolving resonance energies or observing relaxation or quantum coherence between energetically-separated states. On the other hand, a scan over very long delays will be required to enable lineshape analysis and coherence linewidth measurements that are not resolution-limited. This, combined with the limited delay range of most 2DCS experimental techniques, limits the spectral resolution; for example, a 1 nanosecond delay yields a spectral resolution of 1 GHz and is already near the limit of what is possible with macroscopic delay-stage movement.

Interpreting 2D Spectra

Not only do 2DCS experiments provide a powerfully-controlled mapping and visualization of specific quantum pathways, they also provide measurements of fundamental system properties such as dephasing rates and inhomogeneity. Here we will provide a guide to understanding and interpreting 2D spectra. As discussed previously, the choice of pulse ordering determines the quantum pathways probed, and will result in different physical quantities being measured. Unless otherwise stated, we will consider the case of rephasing 2D spectra below.

Interpreting Homogeneous Lineshapes

Peaks located along the diagonal of a 2D spectrum are "degenerate," that is, they indicate that the absorption and emission energies are the same. This indicates that the same coherence is probed during times τ and t, and thus the lineshape of a diagonal peak can be used to characterize the coherent dynamics of that state.

The homogeneous dephasing of a system is the natural linewidth of a pure and isolated resonance, determined by the dephasing dynamics in the system (e.g., for electronic excitations, radiative recombination and scattering with other carriers such as phonons can dephase the electron coherence). Nearly-pure homogeneous responses can be observed in warm, sparse atomic vapors because of the limited interactions between atoms (Li *et al.*, 2013). Polaritonic systems also exhibit homogeneous responses because of the strong coupling to a single photon mode in a cavity. In these cases, the lineshapes have a distinctive "star" or "cross" shape as expected by a direct Fourier transform of the homogeneous signal, and as seen in **Fig. 5**.

Interpreting 2D Spectra of Coupled Systems

Perhaps the most exciting application of 2DCS is the way in which couplings between resonance can be revealed, characterized, and quantified. This is enabled because a 2D spectrum is effectively a transfer matrix, directly showing the output resonances that can be excited for a given absorption frequency. This reveals powerful transient structural information, such as enabling direct observation of which electronic transfer pathways are active and dominant in the complex quantum networks involved in photosynthetic light-harvesting proteins (Brixner *et al.*, 2005) and confirming surprising coherent transfer in synthetic polymers (Collini and Scholes, 2009).

More than revealing that coupling exists, a dynamic series of 2D coherent spectra can also clearly define the incoherent vs coherent nature of the coupling and can even measure important coupling parameters. This is illustrated in **Figs. 9** and **10**, and discussed below.

Fig. 9 shows the case of coherent coupling between states. The energy diagram for this case is an upside-down "Λ" system, in which two excited states $|1\rangle$ and $|1'\rangle$ have a common ground state $|0\rangle$. The possible quantum pathways for a rephasing pulse sequence are shown with double-sided Feynman diagrams in **Fig. 9(b)**. These pathways can be grouped by their input and output resonance energy: diagonal peaks in the 2D spectrum are formed by diagrams R_1 and R_2 at $(\omega_{01},\omega_{01})$ and R_5 and R_6 at $(\omega_{01'},\omega_{01'})$, but the other diagrams lead to cross-peaks as shown in **Fig. 9(c)**. Taking the upper-left cross-peak (light blue) as an example, both R_7 and R_8 contribute on the 2D spectrum at $(\omega_\tau=\omega_{01'}, \omega_t=\omega_{01})$, but the dynamics during the time T between pulses B and C are very different for these two pathways: in R_7, there is a coherence between states $|1\rangle$ and $|1'\rangle$, but in R_8 there is only a population in the ground state. This coherence between excited states is non-radiative, and can be directly measured by measuring the cross-peak amplitude in a series of (ω_τ,ω_t) spectra for different values of T. This is illustrated in **Fig. 9(d)**. The contributions of each pathway are illustrated schematically in this figure, but a real 2DCS experiment could not resolve them. However, the difference in oscillation frequency with T means that a 3DCS experiment could completely separate the individual quantum pathways that make up the cross-peaks in a coherent coupling case.

If there is no coherence between states, there could still be population relaxation between states, as shown in **Fig. 10**. In this case, the possible quantum pathways for a rephasing pulse sequence are simpler: diagonal peaks in the 2D spectrum are formed by the same diagrams as in the coherent coupling case (because diagonal resonances in a 2D spectrum indicate no coupling), but there is only one coupling diagram that describes the relaxation. This diagram, labeled R_9, is different from the double-sided Feynman diagrams we have drawn so far because it has an extra line during time T, between pulses B and C. When pulse B generates a population in the $|1'\rangle$ state, diagram R_9 shows that it is possible for that population to relax to the $|1\rangle$ state without optical excitation. This relaxation process leads to an off-diagonal cross-peak just as in the coherent case, but this peak only appears above the diagonal and it has very different time dependence. **Fig. 10(d)** shows that as a function of T, the relaxation cross-peak grows for time T_{relax} before decaying with the rest of the system's population with the population relaxation time T_2.

If a system of uncoupled resonances is at a high enough temperature such that the thermal energy is on the order of or greater than the difference between the excited state energies, then activation from low-to-high energy excited states can also be achieved.

Spectra are shown here for pure homogeneous resonances, but this analysis works equally-well for coupled resonances with inhomogeneous broadening and/or anharmonicity.

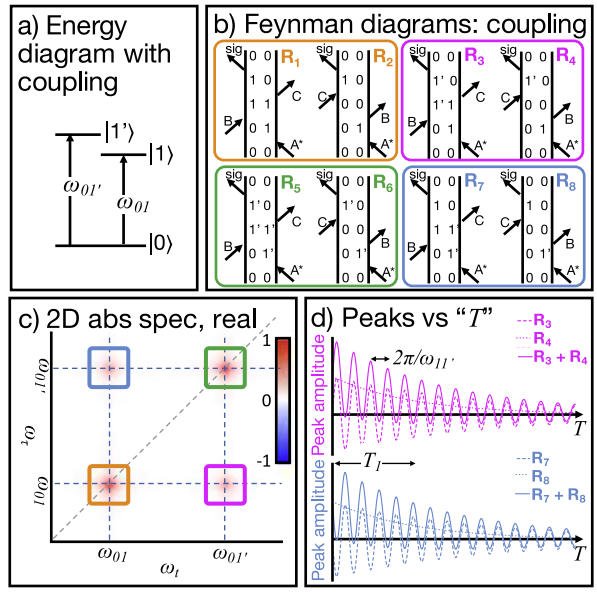

Fig. 9 2D spectra with cross-peaks due to coherent coupling between resonances. (a) energy diagram with coupling, (b) Feynman diagrams: coupling, (c) 2D abs spec, real, (d) Peaks vs. "T"

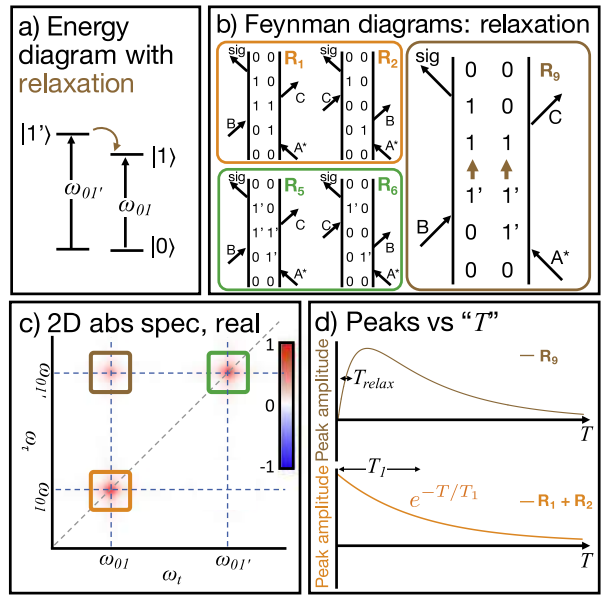

Fig. 10 2D spectra with cross-peaks due to incoherent population relaxation between states. (a) energy diagram with relaxation, (b) Feynman diagrams: relaxation, (c) 2D abs spec, real, (d) Peaks vs. "T"

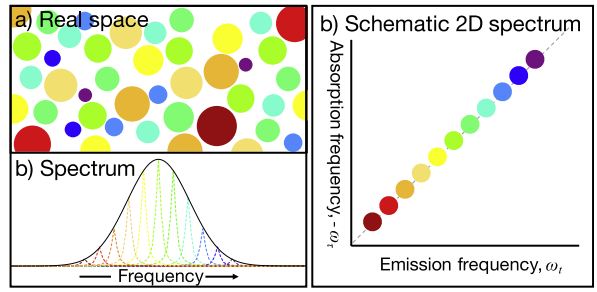

Fig. 11 Schematic depiction of how heterogeneity in a system affects 1D and 2D spectra. (a) Real-space representation of a heterogeneous mixture of nanostructures whose resonance energy depends strongly on the nanostructure size. Color indicates resonance energy. (b) Linear 1D spectrum (absorption or photoluminescence) for the system shown in (a). The fundamental resonance linewidth is obscured by inhomogeneous broadening. (c) Schematic 2D spectrum, illustrating the effect of inhomogeneity to stretch the signal along the diagonal. Color in this schematic 2D spectrum is labeling the shifted resonance frequency and is not related to the amplitude of the response.

Interpreting 2D Spectra of Inhomogeneously-Broadened Systems

Inhomogeneity is unavoidable in condensed matter and molecular systems because of local non-uniformities in the environment, such as fluctuations in the electronic confinement potential that are intrinsic to heterogeneous solid-state systems and conformational variations in molecular systems. In 1D spectra, this leads to a broadening of the fundamental resonance lineshape, as illustrated in **Fig. 11(a)** and **(b)** for a cartoon quantum-confined system such as quantum dots where the resonance energy scales inversely with the size, which varies. This inhomogeneous broadening often dominates 1D absorption or photoluminescence spectra, and there is no way to recover the homogeneous linewidth with linear 1D spectroscopies.

In 2D spectra, inhomogeneous broadening is manifest as an elongation of the signal along the diagonal. Inhomogeneous broadening causes this along-the-diagonal broadening because individual resonators experience different shifts in resonance energy $\omega_{0,i}=\omega_0+\Delta\omega_i$, the effect of which can be seen from the Fourier shift theorem (Ernst *et al.*, 1988):

$$s(\omega_0\tau')\exp[i\Delta\omega\tau'] = S((\omega_0+\Delta\omega)\tau') \tag{4}$$

where $s(\tau')$ and $S(\omega)$ are the Fourier-pair signals in the time and frequency domains, respectively. In other words, a change of the resonance frequency in the time domain (left side of Eq. (4)), is equivalent to a shift of the signal in the Fourier (spectral) domain. In the 2D time domain shown in **Fig. 11(c)**, the signal is resonant along the diagonal ($\tau=t-\tau$) direction; application of the Fourier shift theorem implies a shift of $\Delta\omega$ along the diagonal ($\omega_\tau=\omega_t-\omega$ direction) of a 2D spectrum. Inhomogeneous broadening can

be thought of as simply a distribution of homogeneous systems with slightly shifted resonance energies, leading to a signal spread along the diagonal in a 2D spectrum, as shown in Fig. 11(c).

Measuring homogeneous and inhomogeneous linewidths

In the midst of inhomogeneous broadening, 2DCS allows for clear separation and quantitative measurement of homogeneous and inhomogeneous linewidths. One of the great strengths of rephasing 2DCS is shown in Fig. 12(b), which clearly shows that the inhomogeneous response is isolated along the diagonal ($\omega_{\tau'}=\omega_t+\omega_\tau$ direction) of a 2D spectrum, and the homogeneous response along the cross-diagonal ($\omega_{t'}=\omega_t-\omega_\tau$). Quantitative measurement of both homogeneous and inhomogeneous lineshapes can be performed by fitting cross-diagonal and diagonal slices from a 2D spectrum (Siemens *et al.*, 2010), but accurate measurements require careful consideration of the quantum pathways involved and the effects of homogeneous and inhomogeneous broadening on the lineshapes.

Spectral diffusion

Because 2D spectroscopy can resolve resonance lineshapes along two orthogonal axes, it is an excellent tool for observing and measuring lineshape dynamics. One example in which lineshape dynamics are extremely important is spectral diffusion. Spectral diffusion describes the dynamic and random redistribution of populated states, such that, given sufficient time, the observed resonance energy is equally-likely to be in any possible state within an inhomogeneous distribution, regardless of absorption frequency. For the schematic system shown in Fig. 11(a), spectral diffusion corresponds to site-to-site hopping such as an excited electron hopping between nanostructures or polymers. The physics governing this hopping depends on the system being studied, and possibilities include including quantum tunneling and phonon scattering. In contrast to this interpretation for spectral diffusion of electrons, spectral diffusion of molecular vibrational resonances in a fluid is usually understood directly in the spectral domain as spectral changes of individual resonators. This is because molecules in a fluid are free to move, so vibrational excitations do not "hop" between molecules; rather, the resonance energy of a specific molecular vibration evolves over time as it moves and so experiences different environments and conformations.

In 2DCS measurements of spectral diffusion, a series of 2D scans is performed for steps of T, and the lineshapes are compared. This is illustrated schematically in Fig. 13, where the population diffusion with increasing T is clearly observable as a spreading of the originally-tight lineshape along the cross-diagonal direction. This long-T "spectrally-diffuse" limit in 2D spectra arises because for any absorbed excitation ω_{τ_0}, there is sufficient time for population mixing between local states that the signal could be emitted at any possible ω_t. The emission frequency of the signal will therefore be equally likely to be emitted across the distribution of inhomogeneously-broadened states, leading to no correlation between absorption and emission frequencies. 2DCS is the ideal tool for measuring the timescale of this process, which is an important parameter for understanding the physics governing inter-site population transfer in various systems.

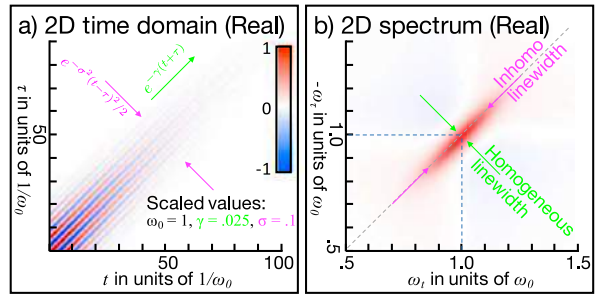

Fig. 12 2D coherent spectroscopy of strong inhomogeneous broadening. (a) Coherent signal in the 2D time domain. (b) 2D spectrum, showing separation of the inhomogeneous and homogeneous linewidths along the diagonal and cross-diagonal directions.

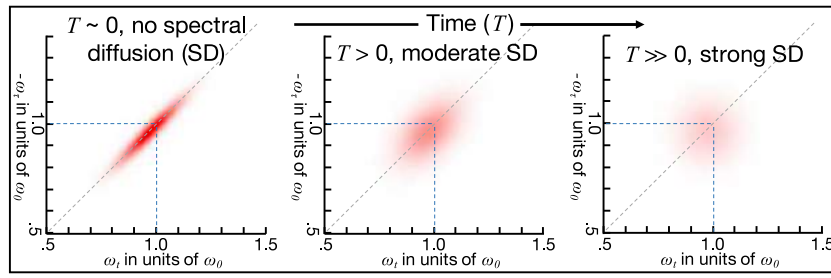

Fig. 13 Spectral diffusion (SD) in 2D spectra. Schematic 2D spectra for increasing waiting time (T, fixed delay between pulses B and C). The cross-diagonal linewidth is only equal to the homogeneous (T_2) linewidth for short waiting time; the cross-diagonal broadening at longer waiting time is due to spectral diffusion, and the lineshape approaches a 2D Gaussian.

Interpreting Anharmonic Oscillator Spectra

The harmonic oscillator is a common theoretical starting point in describing resonant systems near equilibrium, such as vibrational modes in a molecule or solid. However, the perfectly-parabolic potential well required for a pure harmonic response is rare, and so higher-order corrections are usually required. This anharmonicity of resonances can be difficult to identify, much less measure, in congested or inhomogeneous 1D spectra, but 2DCS provides a clear visualization and means of direct measurement of the important anharmonic oscillator parameters. These measurements can be combined with theoretical models to provide information about bond strength, length, and orientation in molecular systems, and many-body interactions in electronic systems.

Consider the anharmonic potential well shown in **Fig. 14(a)**. Assuming that the system starts in the ground state, the double-sided Feynman diagrams for rephasing spectra are shown in **Fig. 14(b)**. Diagrams R_1 and R_2 both oscillate in both τ and t at a frequency of ω_{01}. However, pathway R_3 oscillates in t at frequency ω_{12}, and for an anharmonic oscillator $\omega_{12} \neq \omega_{01}$, so this quantum pathway will lead to quantum beating in the 2D time domain and it will be shifted in the 2D spectral domain, as shown in **Fig. 14(c)** and **(d)**.

Absorption 2D spectroscopy of an anharmonic oscillator provides two important advantages over rephasing 2D spectra: enhanced resolution because of tighter peak shapes and a clear presentation since the real part of the spectrum is absorptive. From this we can clearly see that pathways R_1 and R_2 have a positive contribution to the absorption spectrum, while the R_3 pathway contributes negatively. This sign flip in R_3 arises from having an odd number of pulse interactions on the right-hand side of the Feynman diagram (Boyd, 2003).

The anharmonicity $\Delta = \omega_{01} - \omega_{12}$ can be approximated directly from a 2D spectrum of an anharmonic oscillator by measuring the frequency separating the positive and negative peaks. When the anharmonicity is larger than the linewidths, these peaks are well-resolved and the peak separation is a good measurement of anharmonicity (Zanni *et al.*, 2001a). But more often, the linewidths are larger than the anharmonicity and so the peaks interfere and partially cancel. In this case, careful fitting and lineshape recovery is needed in order to accurately measure the anharmonicity (Zanni *et al.*, 2001a).

The example shown here was for a single, homogeneous resonance, but the same signals interpretation is valid for ensembles with inhomogeneity, making 2DCS a powerful tool for fully characterizing anharmonicity of vibrations in molecular systems (Zanni *et al.*, 2001a). Additionally, this anharmonic approach has recently been applied to reveal the bosonic nature of excitons and the importance of many-body interactions in a semiconductor quantum well (Singh *et al.*, 2016).

Connecting 2D and 1D Spectra

Along with the interpretation of various 2D spectra, it is important to recognize the connection between 2D and 1D spectra. This important connection can be seen with the "projection-slice theorem" of Fourier Transforms (Ernst *et al.*, 1988):

$$\mathcal{F}\left[\int s(x,y)dx\right] = S(\omega_y) \tag{5}$$

which says that Fourier transformation of a projection (integration of all data perpendicular to an axis) onto an axis is equivalent to a slice in the Fourier-pair domain. This is illustrated by the insets of **Fig. 15**, each row shows the projection-slice theorem implemented for the time domain and the Fourier-pair frequency domain.

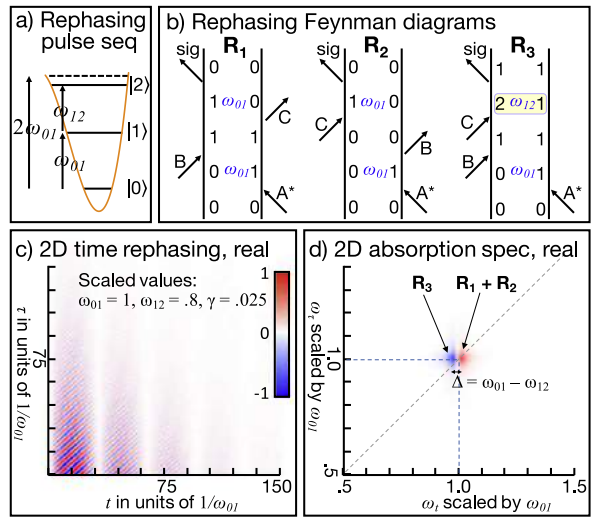

Fig. 14 2DCS of an anharmonic oscillator. (a) Energy level structure. (b) Rephasing Feynman diagrams, with coherence frequencies marked in blue; note the third diagram which has a different oscillation frequency during time t between pulses C and *sig*. (c) 2D time domain signal (real part of rephasing signal). (d) 2D coherent spectrum of an anharmonic oscillator (real part of absorption spectrum, i.e., rephasing + nonrephasing).

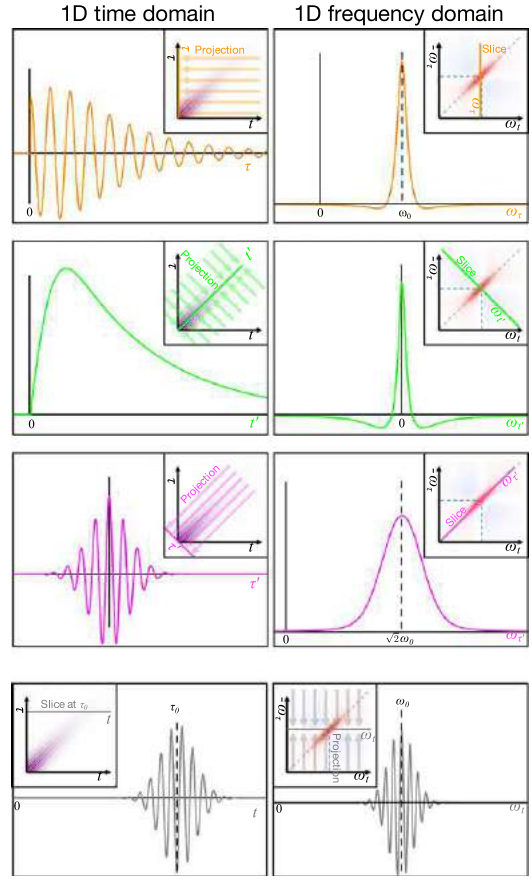

Fig. 15 How to convert from 2D to 1D spectroscopy with the projection-slice theorem. The first three rows show 1D time domain (left column) and frequency domain (right column) signals corresponding to slices along different directions of a 2D spectrum. Insets show the corresponding projection (slice) in the time (frequency) domain. The first row is equivalent to a time-integrated four-wave mixing experiment, but second and third rows do not have simple 1D analogs. In the last row, a slice (projection) in the time (frequency) domain is shown, clearly showing the "photon echo" in which the time-domain signal occurs when the time delays t and τ are equal.

In the case of 2D spectra, projecting onto a given axis in the 2D time domain and then Fourier transforming yields the equivalent 1D spectrum slice along the same direction, in accordance with the projection-slice theorem. Conversely, projection onto a particular axis of the 2D spectrum is equivalent to a slice along the same axis direction in the 2D time domain, as shown for a "photon echo" experiment in the last row of **Fig. 15**. In this way, 2D scans can be easily compared to their 1D analogs, even along directions different from the acquisition such as the diagonal and cross-diagonal. For example, lineshapes along diagonal and cross-diagonal directions of can be immediately recovered from a 2D spectrum, allowing for new insight into homogeneous and inhomogeneous linewidths (Siemens *et al.*, 2010).

One important caveat should be noted: because 2DCS measures the third-order response of a system, a corollary 1D spectrum obtained with the projection-slice theorem will be a 1D coherent spectrum of the third-order response. This 1D spectrum will be equal to that obtained from a time-resolved pump-probe experiment (which measures the third-order response), but will *not* be the same as a linear absorption or photoluminescence spectrum measuring the linear response of the system.

Conclusion and Outlook

2DCS is an active field of research, and the future is very bright. Higher-order spectroscopies, such as 3D, are one area of active growth in multidimensional coherent spectroscopy research, and show promise for further-enhanced resolution of quantum transport pathways, including the possibility of fully-characterizing the Hamiltonian of a system. The recent development of widely-tunable pulsed laser sources and the expansion of 2DCS from the infrared into both the visible and the terahertz spectral regions is a powerful combination: as new materials or molecules are developed for various applications, 2DCS can be quickly deployed to provide deep insight into fundamental coherence dynamics and quantum transport pathways available to the electronic, vibrational, and phononic energy carriers in the system. In the not-so-distant future, it is possible that 2DCS will be a standard analysis tool that can be applied not just for end-of-cycle characterization, but as an essential part of the development cycle of new materials.

While the future is bright, the present is also an exciting time for 2DCS research. The ability to directly supersede one-dimensional linear and coherent spectroscopies with a multidimensional frequency response map is allowing direct measurements of physical properties and tracking of dynamics with a resolution and precision that would have been unimaginable only 25 years ago. In that time, 2DCS has moved from a niche test-bed experiment to become an essential tool in spectroscopy, which promises further excitement for the next quarter-century and beyond.

References

Boyd, R.W., 2003. Nonlinear Optics. Academic Press.

Bristow, A.D., Karaiskaj, D., Dai, X., *et al.*, 2009. A versatile ultrastable platform for optical multidimensional fourier-transform spectroscopy. Review of Scientific Instruments 80, 073108.

Brixner, T., Stenger, J., Vaswani, H., *et al.*, 2005. Two-dimensional spectroscopy of electronic couplings in photosynthesis. Nature 434, 625–628.

Collini, E., Scholes, G.D., 2009. Coherent intrachain energy migration in a conjugated polymer at room temperature. Science 323, 369–373.

Cundiff, S.T., 2014. Optical three dimensional coherent spectroscopy. Physical Chemistry Chemical Physics 16, 8193–8200.

Ernst, R.R., 1992. Nuclear magnetic resonance Fourier transform spectroscopy. In: Malmström, B.G. (Ed.), Nobel Lectures. Singapore: World Scientific Publishing Co.

Ernst, R.R., Bodenhausen, G., Wokaun, A., 1988. Principles of Nuclear Magnetic Resonance in One and Two Dimensions. Oxford: Clarendon Press.

Fecko, C., Eaves, J., Loparo, J., Tokmakoff, A., Geissler, P., 2003. Ultrafast hydrogen-bond dynamics in the infrared spectroscopy of water. Science 301, 1698–1702.

Hamm, P., 2006. Three-dimensional-ir spectroscopy: Beyond the two-point frequency fluctuation correlation function. The Journal of chemical physics 124, 124506.

Khalil, M., Demirdöven, N., Tokmakoff, A., 2003. Obtaining absorptive line shapes in two-dimensional infrared vibrational correlation spectra. Physical Review Letters 90, 047401.

Krummel, A.T., Mukherjee, P., Zanni, M.T., 2003. Inter and intrastrand vibrational coupling in dna studied with heterodyned 2D-IR spectroscopy. The Journal of Physical Chemistry B 107, 9165–9169.

Lepetit, L., Cheriaux, G., Joffre, M., 1995. Linear techniques of phase measurement by femtosecond spectral interferometry for applications in spectroscopy. Journal of the Optical Society of America B: Optical Physics 12, 2467–2474.

Li, H., Bristow, A.D., Siemens, M.E., Moody, G., Cundiff, S.T., 2013. Unraveling quantum pathways using optical 3d fourier-transform spectroscopy. Nature Communications 4, 1390.

Li, X., Zhang, T., Borca, C.N., Cundiff, S.T., 2006. Many-body interactions in semiconductors probed by optical two-dimensional fourier transform spectroscopy. Physical Review Letters 96, 1–4.

Shim, S.-H., Zanni, M.T., 2009. How to turn your pump–probe instrument into a multidimensional spectrometer: 2D IR and vis spectroscopies via pulse shaping. Physical Chemistry Chemical Physics 11, 748–761.

Siemens, M.E., Moody, G., Li, H., Bristow, A.D., Cundiff, S.T., 2010. Resonance lineshapes in two-dimensional fourier transform spectroscopy. Optics Express 18, 17699–17708.

Singh, R., Suzuki, T., Autry, T.M., *et al.*, 2016. Polarization-dependent exciton linewidth in semiconductor quantum wells: a consequence of bosonic nature of excitons. Physical Review B 94, 081304.

Tekavec, P., Lott, G., Marcus, A., 2007. Fluorescence-detected two-dimensional electronic coherence spectroscopy by acousto-optic phase modulation. The Journal of Chemical Physics 127, 214307.

Zanni, M.T., Asplund, M.C., Hochstrasser, R.M., 2001a. Two-dimensional heterodyned and stimulated infrared photon echoes of N-methylacetamide-D. The Journal of Chemical Physics 114, 4579–4590.

Zanni, M.T., Ge, N.-H., Kim, Y.S., Hochstrasser, R.M., 2001b. Two-dimensional ir spectroscopy can be designed to eliminate the diagonal peaks and expose only the crosspeaks needed for structure determination. Proceedings of the National Academy of Sciences 98, 11265–11270.

Two-Dimensional Infrared (2D IR) Spectroscopy

Lauren E Buchanan, Vanderbilt University, Nashville, TN, United States
Wei Xiong, University of California, San Diego, CA, United States

Introduction

Coherent two-dimensional infrared (2D IR) spectroscopy has flourished in its application to problems in chemistry, biophysics and materials science since its first demonstration in 1998 (Hamm *et al.*, 1998). As an IR analogue of 2D nuclear magnetic resonance (NMR) spectroscopy, 2D IR spectroscopy (Hamm and Zanni, 2011; Cho, 2009; Fayer, 2013, 2009; Zheng *et al.*, 2007; Kurochkin *et al.*, 2007; Wright, 2011; Hunt, 2009; Buchanan *et al.*, 2012; Sueur *et al.*, 2015; Mukamel, 2000; Bakulin *et al.*, 2009) probes the time-dependent third-order nonlinear response of molecular vibrational fingerprints, which gives it unique advantages over traditional vibrational spectroscopies, such as linear IR and Raman spectroscopy. These advantages include its ability to: better resolve congested spectral peaks, which enables the determination of molecular structures that are otherwise difficult to obtain; distinguish homogenous and inhomogeneous spectral lineshapes, which allows measurements of local environments; measure spectral diffusions that are related to picosecond molecular dynamics; and observe cross peaks, which reveals structural rearrangements, couplings between vibrational groups, and energy transfer from one site to another. In this article, we will briefly introduce the basic theory of 2D IR spectroscopy and discuss advances in experimental design. Then, we will highlight a few scientific fields to which 2D IR spectroscopy has made seminal contributions. Last, we will present some new developments that are on the horizon which could further bring in transformative capabilities in the fields of non-linear laser spectroscopy.

Brief Theory

2D IR spectroscopy is a coherent non-linear optical process which measures time-dependent vibrational coherences. All 2D IR spectrometers implement the pulse sequence illustrated in **Fig. 1(a)**. IR_1 generates vibrational coherences, which are converted into population states by IR_2. After t_2, IR_3 creates a second set of vibrational coherences that could be from the same or different vibrational modes as the first set; subsequently, an IR signal is emitted and heterodyne detected by a reference IR beam known as local oscillator (LO). The two vibrational coherences are characterized by scanning t_1 and t_3 in time domain. Typically the spectra are visualized in the frequency domain by Fourier transforming the time domain data along both the t_1 and t_3 axes into the ω_1 and ω_3 axes. The measurement of the two molecular vibrational coherences created by IR_1 and IR_3 allows 2D IR to follow molecular dynamics that occur during the measurement time frame ($t_1 + t_2 + t_3$) and also differentiate homogenous and inhomogeneous lineshape broadenings, as we explain next.

A cartoon of a 2D IR spectrum is shown in **Fig. 1(b)**. Here, the plot follows the convention in 2D NMR, which plots the ω_1 frequency along the vertical axis and ω_3 frequency along horizontal axis. Another plot convention in which the horizontal and vertical axes are flipped is also widely used. The spectral peak pairs (1, 2 and 3) along the diagonal line are referred as diagonal peaks and the off-diagonal peaks (highlighted with dashed boxes) are called cross peaks. The diagonal peaks result from the same vibrational coherences being excited before and after waiting time t_2. Cross peaks appear if two different vibrational coherences can be coherently excited by IR_1 and IR_3, which results from vibrational coupling (Khalil *et al.*, 2003a), energy transfer (Xiong *et al.*, 2009; Rubtsova and Rubtsov, 2013, 2015; Lin and Rubtsov, 2012; Rubtsova *et al.*, 2015), or chemical exchange (Kim and Hochstrasser, 2005; Anna *et al.*, 2009; Zheng *et al.*, 2005). Typically, a pair of peaks with opposite phase appear together along the same ω_1 frequency. This is because 2D IR spectroscopy measures both the fundamental and the overtone, or combination bands of vibrational modes. The frequency separation between fundamental and overtone peaks along ω_3 is the anharmonic shift and can be used to determine the anharmonicity, and the peak shifts of cross peaks can be used to extract coupling strength of molecular potential energy surfaces (Golonzka *et al.*, 2015). Besides the anharmonic shift, there are many unique molecular insights that can be extracted from 2D IR spectra.

One such property is the local environment and dynamics, which are encoded in the lineshape of the diagonal peaks. In liquid and solid state systems, the local environments can cause both fast pure vibrational dephasing and vibrational energy relaxation, which is referred to as homogenous broadening, and slow but heterogeneous dynamics, known as inhomogeneous broadening. In linear IR spectroscopy, the homogeneous broadening is characterized by a Lorentzian lineshape while inhomogeneous broadening is often Gaussian. It is straightforward to differentiate Lorentzian from Gaussian lineshapes, but more often the local environments create both fast and slow dynamics, which causes ambiguities in lineshape analysis of linear IR spectra. This ambiguity is overcome in 2D IR spectra. As shown in **Fig. 1(b)**, different local environments can lead to distinct lineshapes. For homogeneous dynamics, the local environments around the probed molecules fluctuate on the time scale of molecular vibrations ($\sim$100 fs) and cause the molecules to lose their vibrational memory. As a result, the t_1 and t_3 vibrational coherences remain uncorrelated and the 2D IR peaks have round lineshapes (peak 2). For inhomogeneous dynamics, the local molecular environments remain static but heterogeneous on the time scale of molecular vibrations. Thus, there is little fluctuation in the vibrational frequency of the probed

 doi:10.1016/B978-0-12-803581-8.09527-8

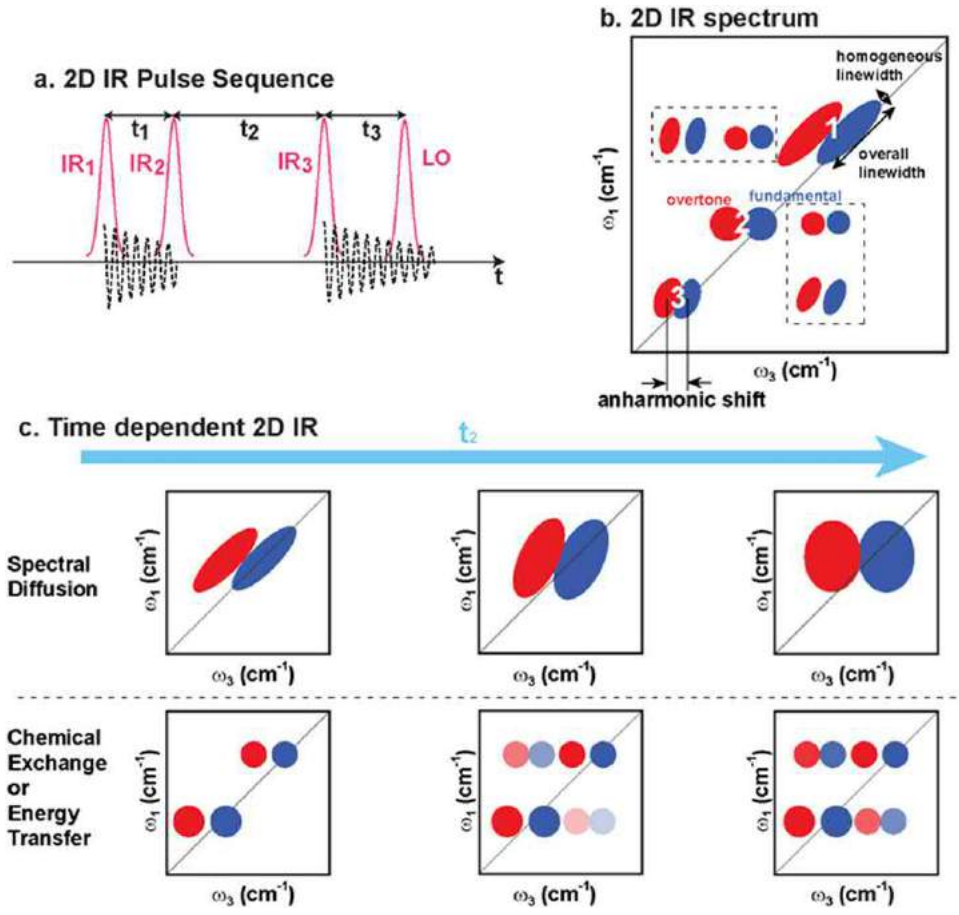

Fig. 1 Schematic 2D IR pulse sequence and spectra. (a) Pulse sequence used to generate two vibrational coherences. IR_1 creates an initial set of vibrational coherences, which are converted into population states by IR_2. IR_3 relaunches the coherences, which emit an IR signal that is heterodyned by LO. The two coherences are characterized by scanning t_1 and t_3. The time domain scan is Fourier transformed to obtain a 2D IR spectrum in the frequency domain. (b) A schematic 2D IR spectrum highlighting spectral features such as cross peaks (marked with dashed boxes), lineshapes and the anharmonic shifts. (c) Schematic 2D IR spectra demonstrating time-dependent features. Spectral diffusion occurs as vibrational modes lose vibrational memory. As a result, tilted and elongated 2D IR peaks become round and symmetric. Cross peaks can grow in as a result of chemical exchange or energy transfer. The rise time of the two sets of cross peaks can differ, which reflects the rates of forward and backward reactions at equilibrium.

molecules – every vibrational mode maintains its vibrational memory and collectively forms a peak elongated along the diagonal (peak 1). For more complicated scenarios in which a mix of homogeneous and inhomogeneous environments exist around probed molecules, the lineshape is neither round nor elongated, but a convolution of these two (peak 3). Still it is easy to distinguish this lineshape from purely homogeneous or inhomogeneous lineshapes and thus qualitatively determine the nature of the local environments.

To obtain quantitative information, the homogeneous linewidth can be determined from the antidiagonal linewidth, whereas the diagonal linewidth corresponds to the overall linewidth (**Fig. 1(b)**; Hamm and Zanni, 2011; Cho, 2009; Khalil *et al.*, 2003b). For molecular systems that involve dynamics in the intermediate range – dynamics that are slower than homogenous broadening, but not as slow as inhomogeneous broadening – spectral diffusion measurements are necessary. Spectral diffusion is measured by scanning t_2 and monitoring the change of spectral lineshape. As t_2 increases and the vibrational modes lose their vibrational memory, the lineshapes evolve from elongated and tilted shapes to round shapes (**Fig. 1(c)**, top). Various parameters have been developed to quantify lineshape changes during spectral diffusion measurements (Hybl *et al.*, 2002; Eaves *et al.*, 2005; Demirdöven *et al.*, 2002; Asbury *et al.*, 2004; Kwac and Cho, 2003b; Hamm, 2006; Fayer, 2009), making it a powerful method to characterize dynamics on the picosecond time scale. Overall, lineshape analysis of 2D IR spectra can provide many previously unavailable insights on local molecular environments and dynamics.

A second unique and useful feature of 2D IR spectra is cross peaks. Since cross peaks involve two different vibrational coherences, studying cross peaks can reveal coupling and dynamics between vibrational modes. The intensity of the cross peaks changes as a function of waiting time t_2 and can indicate chemical exchange or energy transfer (**Fig. 1(c)**, bottom). For instance, in a molecular system with two molecular conformations in equilibrium, there will be two vibrational peaks corresponding to these

conformations along the diagonal. If chemical exchange or energy transfer occurs, cross peaks corresponding to the two conformations involved in the exchange or transfer processes should appear (Zheng *et al.*, 2005; Kim and Hochstrasser, 2005; Rubtsova and Rubtsov, 2013, 2015; Lin and Rubtsov, 2012; Rubtsova *et al.*, 2015). The rise time of the cross peaks reflects the rate of chemical exchange or energy transfer – the faster the cross peaks grow, the quicker the chemical exchange or energy transfer occurs. Furthermore, relative angles between vibrational modes can be calculated from the cross peak intensity ratio between 2D IR spectra obtained using parallel versus perpendicular pump/probe polarization schemes (Hochstrasser, 2001).

By measuring third-order vibrational response functions, 2D IR spectroscopy opens up many novel pathways to reveal molecular structure and dynamics in the liquid and solid phase. This section is only a brief introduction for the theoretical frame work that prepares the audience for the discussion of scientific applications of 2D IR. There have been many comprehensive theoretical developments and computer simulations in 2D IR research by Skinner, Mukamel, Tanimura, Cho, Jansen, and others, to which we refer the audience for further reading (Auer *et al.*, 2007; Woys *et al.*, 2010; Paarmann *et al.*, 2009; Sanda and Mukamel, 2006; Ishizaki and Tanimura, 2006, 2007; Hasegawa and Tanimura, 2008; Park and Cho, 1998; Kwac *et al.*, 2004; Kwac and Cho, 2003a; Roy *et al.*, 2011).

Experimental Approaches

A 2D IR spectrometer requires an ultrafast IR light source, delay lines, a sample chamber, and a detection system. Various implementations exist for each stage of the 2D IR spectrometer, each of which has its own advantages and limitations.

Ultrafast IR Light Source

Ultrafast IR light sources that can create femtosecond pulses at μJ pulse energies and kHz repetition rates form the technical foundation for 2D IR spectroscopy. Because 2D IR signals are generated via a third-order nonlinear optical process, high pulse energies are required. The combination of solid state Ti:sapphire regenerative amplifier lasers and optical parametric amplification (OPA) provides a reliable approach to deliver femtosecond tunable mid-IR pulses for 2D IR spectroscopy. Here we give only a brief introduction on how this setup works; for detailed knowledge about OPA, please refer to an excellent review article (Cerullo and De, 2003). The light source setup typically begins with a commercial oscillator-seeded regenerative amplifier which outputs 800 nm pulses with <100 fs pulse durations and mJ pulse energies at 1 kHz. The 800 nm pulse is sent into a multi-stage OPA system, which converts a single 800 nm photon into a pair of photons at near IR wavelength. The OPA conversion is coherent and therefore requires energy and momentum conservation, as summarized in Eq. (1a,b):

$$\omega_{800\ \mathrm{nm}} = \omega_{Sig} + \omega_{Idl} \tag{1a}$$

$$k_{800\ \mathrm{nm}} = k_{Sig} + k_{Idl} \tag{1b}$$

Here, ω is the frequency of the pulses and k is the corresponding wavevector. *Sig* represent the emitted photon that has higher frequency, which is called signal, and *Idl* represents the one with lower frequency, which is called idler. The condition for conservation of momentum (Eq. 1b) is called phase matching and is achieved using birefringent non-linear optical crystals such as β-barium borate (BBO). To tune the frequency of the signal and idler, the birefringent crystal is rotated to the appropriate phase-matching tilt angle. To generate mid-IR pulses, the signal and idler pulses need to be overlapped spatially and temporally on a second non-linear crystal to undergo a difference frequency generation (DFG) process, as summarized in Eq. (2a,b):

$$\omega_{sig} - \omega_{idl} = \omega_{IR} \tag{2a}$$

$$k_{sig} - k_{idl} = k_{IR} \tag{2b}$$

The mid-IR pulse is generated at a frequency that is the difference between frequencies of signal and idler pulses. Birefringent crystals such as $AgGaS_2$ are widely used in DFG processes. To tune the frequency of mid-IR pulse, the tilt angle of the crystals for both OPA and DFG need to be adjusted in order to have the optimal signal and idler frequencies for the DFG process.

Although the combination of regenerative amplification and OPA is the most popular method to generate intense mid-IR pulses for 2D IR spectroscopy, the majority of these laser systems operate at 1 kHz. To improve sensitivity and broaden applications to samples with weak vibrational transitions, it is necessary to maintain pulse energies while increasing the repetition rate. The Krummel group has recently developed a home-built optical parametric chirped-pulse amplification (OPCPA) setup to implement 2D IR spectroscopy at 100 kHz (Tracy *et al.*, 2016; Luther *et al.*, 2016). They showed that a 100 kHz system is able to collect a 2D IR spectrum with high signal-to-noise in less than 1 s. This development can open new opportunities for resolving fast chemical and biological process that occur at the sub-second time scale.

Delay-Lines and Sample Chamber

After mid-IR is generated, the 2D IR pulse sequence needs to be generated and the IR pulses overlapped spatially and temporally at the sample. Here we discuss the two most popular approaches and their advantages and limitations.

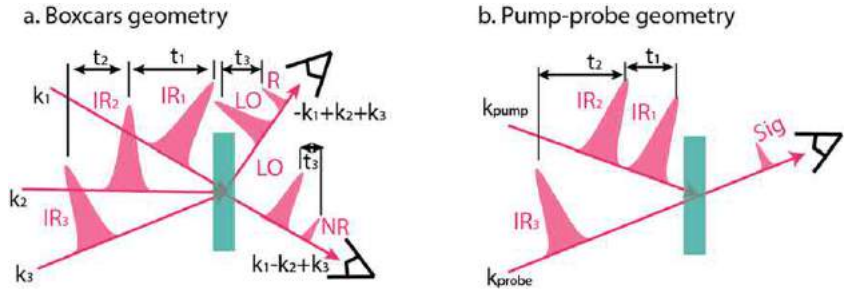

Fig. 2 Geometry of sample chamber. (a) Boxcars geometry. All pulses approach the sample at different angles, and the resphasing (R) and nonrephasing (NR) signals emit at different phase matching angles. Signals are heterodyned by LO. (b) Pump probe geometry. IR_1 and IR_2 together form the pump beam and IR_3 is the probe beam. The phase matching directions of both rephrasing and nonrephasing signals are the same as IR_3. Therefore, IR_3 also acts as LO to heterodyne the signal, a process known as self-heterodyning.

Fig. 2(a) shows the traditional "boxcars" geometry in which three IR beams are aligned so that they approach the sample from three different angles (Khalil *et al.*, 2003b; Rock *et al.*, 2013; Asplund *et al.*, 2000). The phase-matching conditions dictate that the 2D IR signal is emitted at a fourth, unique angle. The benefit of this setup is that rephrasing and nonrephasing 2D IR spectra can be measured separately because they have different phase-matching directions. In addition, the local oscillator can be tuned independently from the other pulses to optimize the signal-to-noise ratio. The limitation of the "boxcars" geometry is that it requires long acquisition times since both t_1 and t_3 need to be scanned using mechanical stages. Thus, it usually takes several hours to complete a spectrum, which makes the setup vulnerable to phase drift and limits it from studying processes such as protein folding that take place on the minute time scale. In addition, if purely absorptive 2D IR spectra are desired, rephasing and nonrephasing spectra must be collected separately and their phase drift must be corrected manually before adding them together (Khalil *et al.*, 2003b).

One way to overcome the above-mentioned limitations is to use a pump-probe geometry for acquiring 2D IR spectra (Shim and Zanni, 2009; Shim *et al.*, 2009; Hamm *et al.*, 2000; Cervetto *et al.*, 2004; Rock *et al.*, 2013). In the pump-probe geometry, IR_1 and IR_2 are arranged collinearly, whereas IR_3 is sent to the sample at a different angle. Because IR_1 and IR_2 are collinear, phase matching dictates that the signal is emitted in the same direction as IR_3. This arrangement not only simplifies the instrumental setup, but also allows self-heterodyne detection; because IR_3 and the 2D IR signal propagate in the same direction, residual IR_3 can serve as the local oscillator to heterodyne the signal. Thus, no additional local oscillator is needed and the phase between signal and local oscillator does not drift. Self-heterodyned signals are often detected by sending IR_3 and signal into a monochromator and projecting them onto a detector. In this way, the time domain signal is experimentally Fourier transformed along t_3 axis into ω_3. Thus, in the pump-probe geometry only t_1 needs to be scanned and the data acquisition time is shorter than "boxcars" geometry.

There are many ways of realizing pump-probe geometry, but the most popular and convenient method is using a mid-IR pulse shaper to construct a double mid-IR pulse pair that serves as IR_1 and IR_2 (Shim and Zanni, 2009; Shim *et al.*, 2009). This shaper-based pump-probe geometry is preferred because other experimental tricks, such as phase cycling and rotating frame, can be implemented easily and because it can scan the t_1 time delay on a shot-to-shot basis, which not only reduces noise from long term fluctuations but also enables rapid data acquisition on the time scale of seconds. Thus, it is possible to use pulse-shaper-based 2D IR setups to study fast protein folding kinetics. Another popular implementation of the pump-probe geometry is the "hole-burning" experiment, which uses a narrow band pump pulse and broad band probe pulse. Instead of taking data in the time domain, "hole-burning" experiments scan the center frequency of the narrow band pump and then plots a series of "hole-burning" pump-probe spectra at different pump frequencies to obtain a 2D IR spectrum. This method is easy to maintain and operate, but it lacks temporal resolution and the spectral lineshapes are often distorted (Hamm *et al.*, 2000; Cervetto *et al.*, 2004).

2D IR Signal Detection Systems

The detection of 2D IR signals is the final part of the 2D IR spectrometer. While the fast development of Si-based photon detectors has benefited detection in the visible regime, mid-IR detectors are both rare and expensive. The most common mid-IR detectors are HgCdTe (MCT) detectors. HgCdTe is a low band gap semiconductor material that can be excited by mid-IR radiation. To reduce charge carrier noise due to thermal fluctuations, MCT detectors often require cryogenic cooling. Both single element and array MCT detectors are used for 2D IR spectroscopy.

Single element MCT detectors are used in the traditional boxcars geometry, in which both t_1 and t_3 are scanned in the time domain. To eliminate fluctuations from the local oscillator, balanced heterodyne detection is often implemented (Mukherjee *et al.*, 2005; Fulmer *et al.*, 2004). The benefit of using a single element MCT detector is that it is economical and its fast read-out time enables shot-to-shot averaging and lock-in-detection, which improve signal to noise. However, as mentioned in the previous section, the data acquisition is slow.

Alternatively, 64- or 128-element MCT array detectors, in combination with a spectrograph, can be used for signal detection in pump-probe geometries, as discussed in the Section Delay-Lines and Sample Chamber. MCT array detectors maintain most of the

advantages of the single element detector, while the pump-probe geometry significantly reduces data acquisition times since only one time delay needs to be scanned. The only caveat is that MCT arrays are relatively expensive. Most recently, two-dimensional focal plane array (FPA) IR detectors have become available for scientific research. Since an FPA has more pixels than a traditional array detector, it allows a reference beam to be monitored simultaneously with data collection, which improves signal-to-noise. Furthermore, this detector makes new capabilities such as 2D IR imaging possible (Serrano *et al.*, 2015; Ostrander *et al.*, 2016).

Besides directly detecting mid-IR signals, an alternative approach is to upconvert IR photons into the visible regime, where sensitive visible detectors can detect the signal. Kubarych and co-workers pioneered this approach by mixing a chirped narrowband 800 nm pulse with the emitted IR signal on a non-linear crystal to upconvert the emitted IR signal into a visible signal via sum-frequency generation (Nee *et al.*, 2008, 2007; Anna and Kubarych, 2010). Because the upconversion process is non-resonant, the temporal profile of the original 2D IR signal is well-preserved. The upconversion process enables the use of visible light detectors that are typically cheaper and more sensitive to be used for 2D IR spectroscopy. So far, both charge-coupled device (CCD) and complementary metal–oxide–semiconductor (CMOS) cameras have been implemented in 2D IR spectrometers (Rock *et al.*, 2013).

Applications

The experimental methods discussed above represent the most popular approaches to obtain 2D IR spectra. Each has its own advantages and limitations, but all have allowed 2D IR spectroscopy to provide unprecedented insights about molecular structure and dynamics. Below, we discuss a few examples in which 2D IR spectroscopy has been applied to problems across chemistry, materials, and biology.

Material Sciences

Solid state materials

2D IR spectroscopy has been implemented to gain unique insights into the structures and dynamics of materials, which has implications on designing and controlling materials at molecular level. 2D IR spectroscopy can better resolve molecular inhomogeneities than FTIR (Xiong *et al.*, 2009; Oudenhoven *et al.*, 2015; Barbour *et al.*, 2006), which is very useful in studying heterogeneous solid state materials. Furthermore, the cross peaks of 2D IR spectra are often used to follow dynamic processes in materials, such as ion-exchange and charge (Kwon and Park, 2015; Fulfer and Kuroda, 2016; Xiong *et al.*, 2009). Lastly, by measuring 2D IR spectra at different t_2, it is feasible to probe the dynamics of molecular ligands or solvent molecules near the material surfaces (Yan *et al.*, 2015, 2016; Nishida *et al.*, 2014; Cui *et al.*, 2016; Huber *et al.*, 2015; Huber and Massari, 2014; Ren *et al.*, 2015, 2014; Dutta *et al.*, 2015).

Charge dynamics of dye-sensitized solar cells

2D IR spectroscopy reveals that charge transfer dynamics in dye-sensitized solar cells (DSSCs) are conformation dependent (Xiong *et al.*, 2009; Oudenhoven *et al.*, 2015). In DSSCs, multiple timescale charge transfer dynamics have been observed by transient absorption spectroscopy. It is speculated that dye molecules can form different conformations on the surface of nanoparticles, each of which has its own charge transfer pathway which therefore lead to the multiple time scale dynamics. However, it is difficult to resolve conformations using FTIR spectroscopy, in which spectral peaks are congested. It is also difficult to distinguish conformations in traditional visible-pump IR-probe experiments because the vibrational peaks are on top of a broad free-electron absorption background, which makes resolving vibrational features difficult (Anderson and Lian, 2005). Using 2D IR spectroscopy, Xiong, Zanni and co-workers studied the surface conformation and charge dynamics of $Re(CO)_3$(Bi-pyridine)COOHCl sensitized TiO_2 nanoparticles. Three vibrational peaks of the A′(1) mode of the CO stretch of the dye molecules are clearly resolved (**Fig. 3(a)**). In solution phase, the A′(1) mode only results in one vibrational peak. Thus, the three distinct peaks reflect that three different conformations are formed at the TiO_2 surface. The reason that 2D IR can differentiate vibrational peaks better than FTIR is that 2D IR signal is proportional to transition dipole moment, μ, to the fourth order (μ^4) whereas FTIR depends on μ^2.

To understand how different conformations contribute to the charge transfer dynamics, a UV pulse is inserted between IR_2 and IR_3 during waiting time t_2. This allows 2D IR to correlate conformations before and after charge transfer has occurred (Bredenbeck *et al.*, 2004). When a new cross peak appears in the 2D IR spectrum, it reveals that charge transfer has occurred and the corresponding conformation is in an electronic excited state. At the same time, diagonal peaks of all three conformations are bleached, demonstrating that the ground state populations of all three conformations are decreased. Thus, all conformations are excited by the UV pulse and involved in charge transfer. However, the conformation at $\omega_{pump}=2040\ cm^{-1}$ does not have the excited state cross peak that the lower frequency conformations do (**Fig. 3(b)**). This difference indicates that at the time that the dye molecules are probed, the lower frequency conformations are still in electronic excited states, but the conformation at $\omega_{pump}=2040\ cm^{-1}$ is in neither a ground nor excited electronic state. Therefore, it is highly likely that the electron has transferred from this conformation to the TiO_2 (Xiong *et al.*, 2009). Further investigations using 2D IR spectroscopy show that these three conformations are three different aggregates of dye at the surface of TiO_2 (Oudenhoven *et al.*, 2015). Thus, 2D IR has a unique capability to resolve multiple molecular conformations in materials and further resolve charge transfer dynamics with conformational specificity when combined with UV or visible excitation pulses.

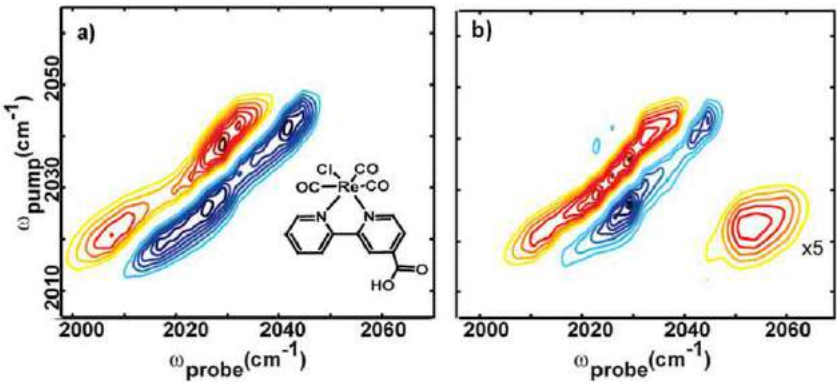

Fig. 3 2D IR spectra of dye -sensitized solar cells. (a) A static 2DIR spectrum shows three vibrational peaks, indicating three different conformations when dyes attach to the surface of TiO_2. (b) A transient 2D IR spectrum after charge transfer is initiated by UV pulse. Since all three diagonal peaks appear, all three conformations are involved in the charge transfer. However, a large cross peak only appears at $\omega_{pump} = 2020$ cm^{-1}, whereas no cross peak appears at higher frequency. This result indicates that different conformations have different charge transfer dynamics. Details refer to main text. Adapted from Xiong, W., Laaser, J.E., Paoprasert, P., *et al.*, 2009. Transient 2D IR spectroscopy of charge injection in dye-sensitized nanocrystalline thin films. Journal of the American Chemical Society 131, 18040–18041.

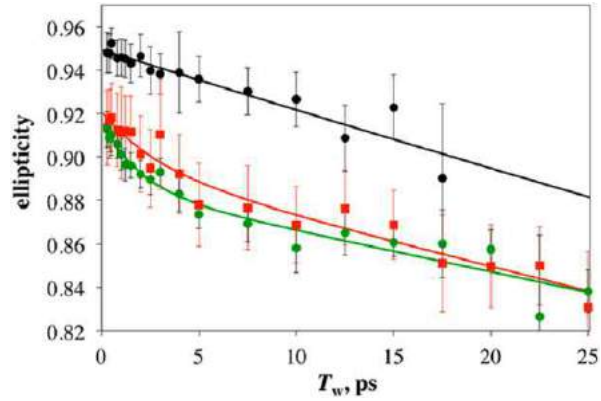

Fig. 4 Spectral diffusion of Si-H on the surface of nanoscale pores, infiltrated with pentane (black circles), isopropanol (green circles), and isopropanol replaced with pentane (red squares). Overlaid solid lines are multiexponential fits to the data. After pentane replaces isopropanol as the solvent in nanoscale pore, spectral diffusion dynamics remain the same as isopropanol-infiltrated pores. This suggests that isopropanol strongly binds to the silica surface and it cannot be displaced by pentane. Adapted from Huber, C.J., Massari, A.M., 2014. Characterizing solvent dynamics in nanoscopic silica sol – gel glass pores by 2D-IR spectroscopy of an intrinsic vibrational probe. Journal of Physical Chemistry C 118, 25567–25578.

Nanoscale materials

Nanoscale materials are attractive to scientific and technological developments due to their large surface area and tunable mechanical, optical and electromagnetic properties. In particular, nanoscale porous materials have been developed for chemical and energy storage purposes. Molecules inside these materials are spatially confined and experience molecule-pore surface interactions which can change the structure and dynamics of confined molecular ensembles (Nishida *et al.*, 2014; Huber *et al.*, 2015; Huber and Massari, 2014). Understanding molecule-pore surface interactions and how they affect confined molecular ensembles is critical to optimizing the capacity and selectivity of nanoscale porous materials. To understand these effects, different vibrational probes have been developed for 2D IR spectroscopy to investigate the structure and dynamics of molecules confined in nanoscale pores.

Si-H vibrational modes on silica surfaces have been used to probe local solvent dynamics in porous silica nanoparticles. It is shown that spectral diffusion measurements of the Si-H modes are sensitive to solvent dynamics up to 5 nm away (Huber *et al.*, 2015). Using the same vibrational probes, solvent-pore surface interactions are studied for nanoscale silica sol-gel glass pores. Huber and Massari observe different spectral diffusion dynamics when the pores are infiltrated with pentane versus isopropanol. More interestingly, when isopropanol in the nanopores is replaced by pentane, the spectral diffusion dynamics remain similar to the isopropanol-infiltrated nanopores (**Fig. 4**). This result suggests that although pentane could replace the majority of solvents molecules inside the pores, isopropanol strongly bound to silica surface and the binding process is irreversible (Huber and Massari, 2014).

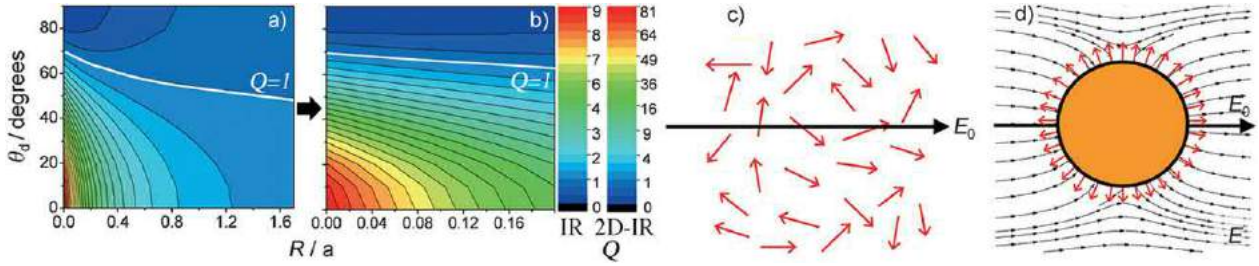

Fig. 5 Signal enhancement of 2D IR spectroscopy on metal nanoparticles. Plots of the IR (a) and 2D-IR (b) signal enhancement factors as a function of dipole tilt angle, θ_d, and distance from the surface, R, for an orientationally-averaged ensemble of dipoles located around a spherical metal nanoparticle of radius Illustration of relative angles between electric field and transition dipole moments in isotropic (c) and nanoparticle (d) situations. Adapted from Donaldson, P.M., Hamm, P., 2013. Gold nanoparticle capping layers: Structure, dynamics, and surface enhancement measured using 2D-IR spectroscopy. Angewandte Chemie – International Edition 52 (2), 634–638.

Another nanoporous material that has been studied by 2D IR is metal-organic-frameworks (MOFs). [FeFe](dcbdt)$(CO)_6$ is attached to a UiO-66 MOF to probe the dynamics of inner pores (Nishida *et al.*, 2014). Nishida, Fayer, and coworkers observed a slow 670 ps spectral diffusion constant in empty MOF pores, which is attributed to a large-scale motion of the MOFs. When filled with DMF, the spectral diffusion becomes even slower. This result indicates that MOF structures become more rigid and the large scale motions are further slowed by hosting DMF. More interestingly, no large homogeneous broadening is observed. This is in sharp contrasts with the spectral lineshape of [FeFe](dcbdt)$(CO)_6$ in bulk DMF solutions, which is significantly homogeneously broadened. Since in solution the broad homogeneous lineshape can be attributed to solvent motions that occur faster than or on the same time scale as the time resolution of 2D IR, the lack of homogeneous broadening in the 2D IR spectra of DMF-infiltrated MOFs indicates that the motion of DMF molecules is largely constrained. These pioneering works demonstrated how 2D IR spectroscopy could provide insights into interactions between molecules and surface of materials in nanoscale, confined spaces.

Nanoscale materials are not only interesting scientific systems to be investigated by 2D IR spectroscopy, they can also advance 2D IR spectroscopic techniques (Donaldson and Hamm, 2013; Selig *et al.*, 2015). Donaldson and Hamm showed theoretically and experimentally that the 2D IR signal can be enhanced for molecules bound at the surface of metal nanoparticles (Donaldson and Hamm, 2013). This can be understood in a simplified scenario: when molecules bind to a metal nanoparticle, certain vibrational transition dipoles may orient perpendicular to the surface. These transition dipoles are also parallel to the near-field electric fields, which orient parallel to the surface normal. Therefore, the vibrational modes of surface-bound molecules are aligned uniformly with the electric fields (**Fig. 5(d)**). As a result, the dipole interaction with the electric field is strengthened relative to the solution phase, where the relative angles between transition dipoles and electric fields are isotropically distributed (**Fig. 5(c)**). In this case, the 2D IR signal is enhanced 81 times. The more general scenario is described in **Fig. 5(a)** and **(b)**, in which the enhancement factor was calculated as a function of distance to the metal surface, R, and tilt angle to surface normal, θ_d. The enhancement is strongest when R is small and θ_d is 0. As R becomes large and θ_d approaches to 90, the enhancement factor decrease to zero. Experimentally, 2D IR signal enhancement is measured using amide-PEG coated gold nanoparticles, where the distance to gold surface and particle sizes are varied. It is found that enhancement occurs up to 1 nm from the surface and the enhancement is larger when large (radius = 10 nm) nanoparticles are used.

Molecular electrolytes

Ion structure and dynamics of electrolytes

Molecular properties of electrolytes directly influence the performance of batteries, particularly in lithium ion batteries which are the most commonly used energy storage devices. A wide range of ion-molecule and ion-ion interactions exist in the electrolyte which can influence ion diffusions, pairing and solvation (Kwon and Park, 2015; Cui *et al.*, 2016; Fulfer and Kuroda, 2016). When complexes are formed between ions or between ions and molecules, the vibrational frequencies of the ions and molecules can shift due to differing electrostatic environments. As a result, the dissociated and associated configurations have two distinct vibrational peaks in FTIR. Similarly, 2D IR spectra contain diagonal peaks of different configurations and cross peaks that can reveal the exchange dynamics between these configurations. Thus, 2D IR spectroscopy is very useful to resolve dynamics of ion-ion or ion-molecule complex formations in electrolytes. The structure and dynamics of the electrolyte between $LiPF_6$ and organic carbonates have been studied using 2D IR spectroscopy (Fulfer and Kuroda, 2016). Together with DFT calculations, Fulfer and Kuroda find evidence to support that Li^+ and organic carbonates form two types of ion pairs, contact and solvent-separated ion pairs, which has implications in rational design of electrolyte composition for better Li batteries. In another work, a model Li^+ electrolyte system, Li^+ and CH_3SCN in acetonitrile, is investigated by Kwon and Park using 2D IR spectroscopy, where the SCN vibrational modes of CH_3SCN are probed (Kwon and Park, 2015). Two distinct diagonal peak pairs are resolved in the 2D IR spectrum at 1 ps (**Fig. 6**). The high frequency peak (C) corresponded to the Li^+-CH_3SCN complex, and the peak at low frequency (F) is assigned to the free CH_3SCN. Please note the second plot convention is used in these 2D IR spectra, as discussed in Section Brief Theory. As t_2 increases, cross peaks start to appear. Due to interference with the overtone signal of the diagonal peaks, part of the cross peaks (X2) does not appear clearly. Nevertheless, peaks that are well-separated from diagonal peaks (X1 and X3) can be identified.

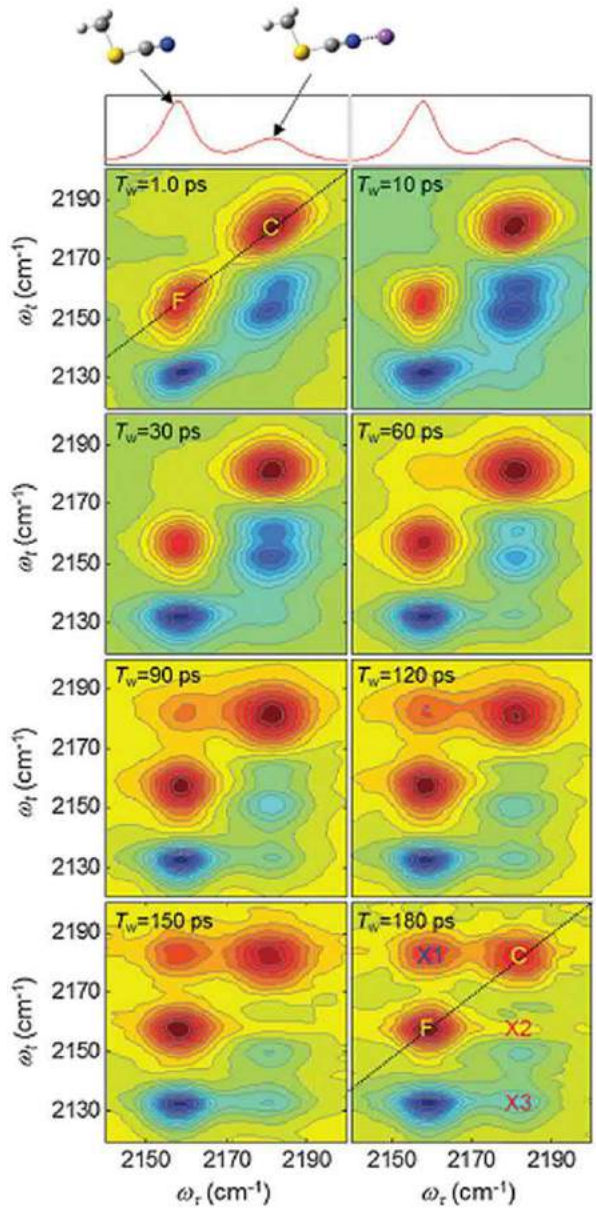

Fig. 6 Time-dependent 2D IR spectral series of Li^+ and CH_3SCN. The free CH_3SCN (F) and Li^+-CH_3SCN complex (C) are resolved along 2D IR diagonal peaks. As time increases, the free and complex forms exchange conformation and therefore cross peak (X) appears. Adapted from Kwon, Y., Park, S., 2015. Complexation dynamics of CH_3SCN and Li^+ in acetonitrile studied by two-dimensional infrared spectroscopy. Physical Chemistry Chemical Physics 17, 24193–24200.

X1 corresponds to transformation from F to C, which is the association process, and X3 corresponds to the dissociation reactions. By following the cross peak growth dynamics and fitting them using a kinetic model, the dissociation time constant is determined to be 160 ± 10 ps and the association time constant is 360 ± 20 ps. These quantitative results provide some fundamental parameters for understanding dissociation and association dynamics of electrolytes for Li^+ batteries.

Another special electrolyte material is an ionic liquid (Dutta *et al.*, 2015; Brinzer *et al.*, 2015; Ren *et al.*, 2015, 2014). Ionic liquids are molten salts at room temperature that are composed of organic cations and anions. Besides their use as electrolytes, ionic liquids also have applications in carbon capture and tribology. Multiple molecular forces exist in ionic liquids, which leads them to have unique but complicated molecular properties. For instance, it is observed that a decrease in hydrogen bonding can lead to a large increase in viscosity, which is counterintuitive. 2D IR spectroscopy has been used to understand the molecular origins of an ionic liquid's viscosity (Ren *et al.*, 2014). In this work, two ionic liquids, 1-butyl-3-methylimidazolium bis(trifluoromethylsulfonyl)imide ($[C_4C_1im][NTf_2]$) and 1-butyl-2,3-dimethylimidazolium bis(trifluoromethylsulfonyl)-imide ($[C_4C_1C_1{}^2im][NTf_2]$), are studied. These two ionic liquids only differ by one H atom that is replaced by CH_3 in $[C_4C_1im][NTf_2]$, which breaks the hydrogen bonds between the cation and anions. Ren, Garrett-Roe and co-workers measure spectral diffusion of

both compounds and find that the time constants of spectral diffusion dynamics, i.e. the ion complex breaking and reforming dynamics, are 47 ± 15 ps for the methylated ionic liquid ($[C_4C_1C_1{}^2im][NTf_2]$) and 26 ± 3 ps for ($[C_4C_1im][NTf_2]$). The authors further suggest that this difference in the ion-complex reorganization rate is responsible for the difference in viscosity between the two liquids. In a Stokes-Einstein picture, the fundamental limit to activate translational diffusion and viscosity is the time to break up the local ion-complex; the faster ion complex reorganization is, the lower viscosity will be. Besides viscosity, other properties of ionic liquids, such as solute-solvent interactions and interactions between CO_2 and ionic liquids, have also been investigated (Brinzer *et al.*, 2015; Ren *et al.*, 2015; Dutta *et al.*, 2015).

Organometallic chemical catalysts

Organometallic chemical catalysts have been developed and used for many reactions, such as CO_2 reduction (Kumar *et al.*, 2012), H_2O splitting (Weston *et al.*, 2011), and CH activation (Crabtree, 2004). During catalytic reaction cycles, organometallic catalysts undergo transitions from one intermediate state to another on a fast time scale that cannot be followed by techniques such as NMR spectroscopy. 2D IR spectroscopy has been used to explore the transitions between different intermediates with fs or ps temporal resolution (Jones *et al.*, 2011; Nee *et al.*, 2008; Anna and Kubarych, 2010; Cahoon *et al.*, 2008).

Dynamics of homogeneous catalysts

One beautiful example is a study of the famous Berry's pseudorotation of $Fe(CO)_5$ by Cahoon, Harris and co-workers (Cahoon *et al.*, 2008). Berry's pseudorotation predicts that $Fe(CO)_5$, which has a D_{3h} geometry, can undergo transformations through C_{4v} geometry and return back to the original D_{3h} geometry. During this process, however, the two axial CO groups exchange with two equatorial CO groups (**Fig. 7(B)**). $Fe(CO)_5$ has two IR active modes. The doubly degenerate e′ modes that involve three equatorial CO groups absorb at 1999 cm^{-1} and the a″ mode, which is composed of vibrations of axial CO groups, absorbs at 2022 cm^{-1}. It is observed that as t_2 increases, cross peaks between the e′ and a″ modes appear in the 2D IR spectra (**Fig. 7(A–C)**). The cross peak dynamics reflect Berry's pseudorotation: as axial and equatorial CO groups exchange their positions, the vibrational energy they carry would also exchange from a″ to e′ and vice versa. The authors model the vibrational energy transfer process and simulate corresponding 2D IR spectra (**Fig. 7(D–F)**). They found the pseudorotation occurs within 8.0 ± 0.6 ps, with an activation barrier of 2.13 kJ/mol (**Fig. 7(G)**). Besides this example, there are many other interesting studies on the isomerization and vibrational dynamics of organometallic catalysts in solution phase using 2D IR spectroscopy (Anna *et al.*, 2009; Nee *et al.*, 2008; Anna and Kubarych, 2010; Jones *et al.*, 2011). These works provide valuable insights on the dynamics of transition states in organometallic catalysts.

Surface attached catalysts

Another exciting application is to use reflection mode 2D IR spectroscopy to study surface-attached catalysts (Rosenfeld *et al.*, 2011, 2013; Yan *et al.*, 2015, 2016; Nishida *et al.*, 2016; Wang *et al.*, 2015; Li *et al.*, 2016a,b). Organometallic catalysts have been attached to metals and semiconductors to perform electrochemistry. Yet, there is a lack of fundamental knowledge on their structure and dynamics, which could be affected by the surface attachment. The Fayer group studied a set of fac-Re organometallic catalyst attached by click chemistry onto gold and silica surfaces through various lengths of linkers (Rosenfeld *et al.*, 2011, 2013, 2012; Yan *et al.*, 2015, 2016; Nishida *et al.*, 2016). By carefully controlling the surface density of these Re compounds, it is shown that the spectral diffusion

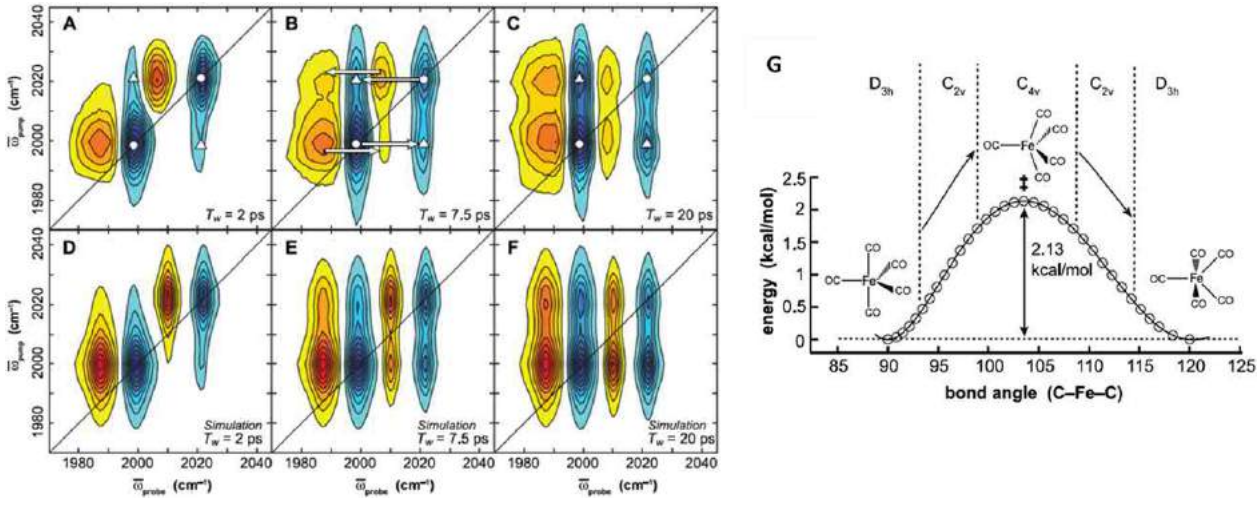

Fig. 7 2D IR measurement of Berry's pseudo rotation on $Fe(CO)_5$. (A–C) Experimental 2D IR spectra at different time delays show cross peak growth as waiting time increases. (D–F) Simulated 2D IR spectra based on a vibrational energy transfer model. (G) Depiction of the mechanism and energetics of a pseudorotation plotted as a function of the bond angle (C–Fe–C) between one axial and one equatorial CO ligand that exchange during the pseudorotation. Adapted from Cahoon, J.F., Sawyer, K.R., Schlegel, J.P., Harris, C.B., 2008. Determining transition-state geometries in liquids using 2D-IR. Science 319, 1820–1823.

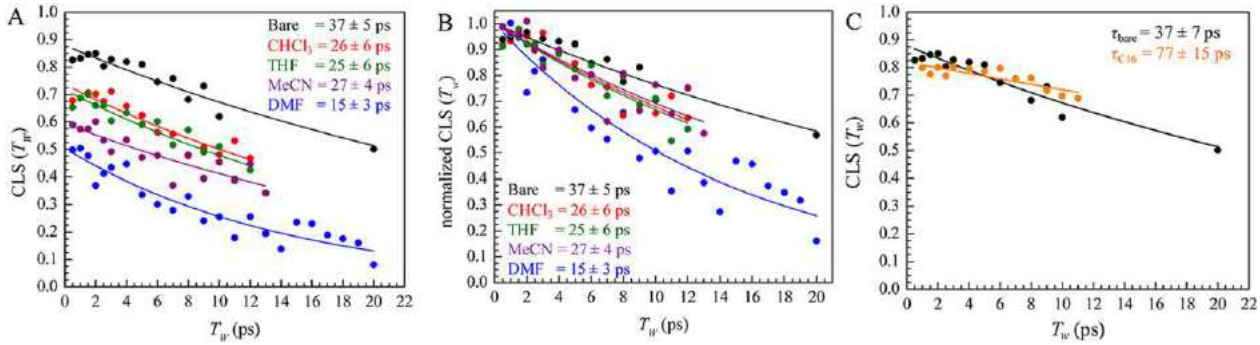

Fig. 8 Spectral diffusion measurements of surface-immobilized catalysts with different solvent environments. (A) Spectral diffusion and (B) normalized spectral diffusion dynamics in organic polar solvents. (C) Spectral diffusion in non-polar solvents. When surface-immobilized catalysts are immersed in solvents that can dissolve the catalysts, they have faster dynamics than under solvent-free environments. Adapted from Rosenfeld, D.E., Nishida, J., Yan, C., *et al.*, 2013. Structural dynamics at monolayer – liquid interfaces probed by 2D IR spectroscopy. Journal of Physical Chemistry C 117, 1409–1420.

measurement is primarily affected by structural dynamics such as constrained-rotation motions and surrounding solvent dynamics (Rosenfeld *et al.*, 2012, 2013). The spectral diffusion time constants are especially sensitive to the surrounding solvents. When polar solvents that can solvate the Re compounds are used, the corresponding spectral diffusion is faster than on a "dry" catalyst surface (Rosenfeld *et al.*, 2013). However, when solvents that do not solvate the Re compounds are used, the spectral diffusion dynamics are significantly slower – even slower than the "dry" surface (**Fig. 8**). The difference in spectral diffusion demonstrates that the local solvent structures differ greatly depending on the type of solvent used. The polar organic solvents could swell the surface and penetrate into the catalytic layer, whereas the other solvents might cause a compaction of the surface. Besides being highly sensitive to the solvent, it is also shown that the spectral dynamics are sensitive to the length of the linker, the surface coverage, and the substrate (Yan *et al.*, 2014, 2015). In general, the longer the linker, the faster the dynamics are. This agrees with the intuition that longer linkers are more flexible. Furthermore, it is found that the dynamics on gold surfaces is much slower than on SiO_2, because the thiol anchoring groups used on the gold surface are mobile and allow 2D "recrystallization" into a more compact monolayer, which slows the structural dynamics of the catalysts. A recent combination of 2D IR spectroscopy and MD simulations explores the molecular origin of the spectral diffusion of these catalysts. It is found many types of linker dynamics contribute to spectral diffusion, including dihedral flips in the alkyl chain and triazole ring flips, which drive subsequent chain motions (Yan *et al.*, 2016). These novel developments and applications to surface attached organometallic catalysts allow in-depth knowledge to be gained about the structure and dynamics of the immobilized catalysts and the surrounding local solvents.

Protein Structure and Folding Dynamics

The function of a protein is determined by its structure and all biological processes involve changes in protein conformation. A variety of techniques exist to study protein structure, including X-ray crystallography, NMR spectroscopy, and electron microscopy. These techniques are best applied to static structures, however, and rarely allow structural dynamics to be observed directly. In contrast, 2D IR spectroscopy can be used to observe protein conformational dynamics that occur over a wide range of timescales, from picoseconds to hours (Ganim *et al.*, 2008; Kim and Hochstrasser, 2009; Hamm *et al.*, 2008; Woutersen and Hamm, 2002; Zhuang *et al.*, 2009).

Most infrared studies of proteins focus on the backbone amide I mode, which is generated primarily by the C=O stretching motion with a smaller contribution from N-H stretching. Within a protein, the amide I modes of individual amino acids vibrate in unison to create delocalized normal mode vibrations that extend across multiple amino acids. This coupling, which depends strongly on the secondary structure of the protein, causes a change in the frequency of the amide I mode. For example, the random couplings between residues within disordered peptides cause the amide I mode to appear as a broad peak centered around 1650 cm^{-1}, while the regular coupling patterns within β-sheets structures causes the amide I mode to narrow and shift to lower frequencies, which can range from 1620 cm^{-1} for highly ordered, extended β-sheets to 1635 cm^{-1} for smaller β-sheets. Thus, changes in the frequency of the amide I mode indicate changes in protein secondary structure (**Fig. 9**). A few 2D IR studies of proteins have examined additional vibrational modes, including the backbone amide II mode (Maekawa *et al.*, 2009), which comprises N-H bending and C-N stretching, and sidechain modes (Buchanan *et al.*, 2014).

While the amide I mode can be used to determine the overall structural content of a protein, a 2D IR spectrum of a wild-type protein generally does not contain enough information to determine a more detailed protein structure. Yet, 2D IR spectroscopy is capable of residue-specific structural resolution when applied to labeled proteins. Isotope labels can be incorporated into the protein during synthesis or expression (Middleton *et al.*, 2010; Moran *et al.*, 2012) and do not perturb the structure or the dynamics. Commonly, either $^{13}C=^{16}O$ or $^{13}C=^{18}O$ labeling of the backbone carbonyl is used. The heavier nuclei cause the amide I frequency of the labeled residues to shift by 40 or 60 cm^{-1}, respectively, so that the labeled residues are spectrally resolved from the remaining unlabeled residues (**Fig. 9(b)**). As a result, the secondary structure of the labeled residue can be determined by analyzing the frequency, linewidth, and crosspeaks of the labeled mode.

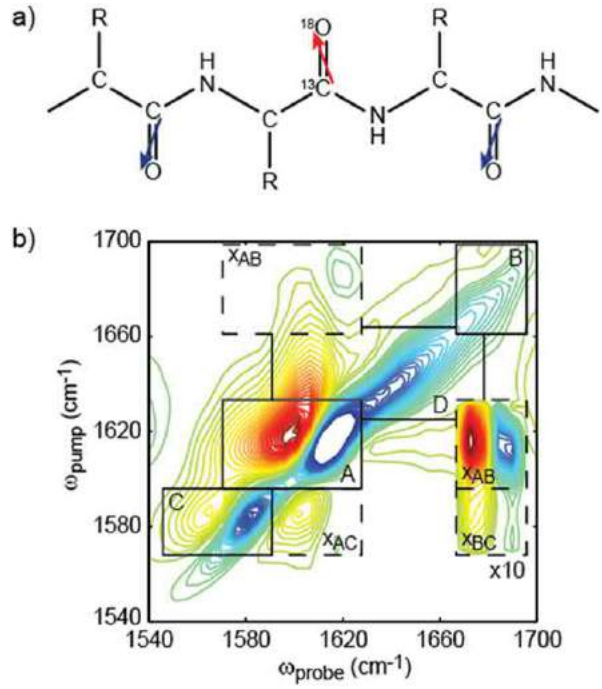

Fig. 9 Overview of protein 2D IR spectroscopy. (a) Peptide sequence with amide I modes marked with arrows. The dipole lies about 20° off the C=O bond. One carbonyl is isotope labeled with $^{13}C^{18}O$, which shifts the frequency of that mode by 60 cm^{-1}. (b) Example spectrum for an isotope-labeled peptide with diagonal peaks, enclosed in solid boxes, characteristic of β-sheet (A), β-turn (B), and disordered (D) structures. The isotope-labeled mode (C) appears well separated from the other amide I modes. Crosspeaks between diagonal modes, enclosed in dashed boxes, appear when the corresponding regions of secondary structure are coupled. Spectrum reproduced from Buchanan, L.E., Dunkelberger, E.B., Zanni, M.T., 2012. Examining amyloid structure and kinetics with 1D and 2D infrared spectroscopy and isotope labeling. In: Fabian, H., Naumann, D. (Eds.), Protein Folding and Misfolding: Shining Light by Infrared Spectroscopy, vol. 1. Springer, pp. 217–237.

In the follow sections, we will highlight a few examples of how 2D IR spectroscopy has been applied to amyloid fibrils and membrane proteins.

Amyloid proteins

The misfolding and aggregation of proteins into amyloid fibrils is associated with more than 20 human diseases including type II diabetes, Alzheimer's disease, prion diseases, and cataracts. Yet, there are also cases of functional amyloids that are required for necessary for biological functions such as the formation of melanin (Chiti and Dobson, 2006) and recent studies indicate that the ability to form amyloids may be common to all proteins (Goldschmidt *et al.*, 2010). Thus, an understanding of amyloid protein structure and formation is important not only for therapeutic applications but also for the broader field of protein folding.

Amyloid fibers are characterized by extended, intermolecular β-sheets in which the individual β-strands are oriented perpendicular to the fiber axis. Atomic-level structures of amyloid proteins are difficult to determine. While a few X-ray crystallography and NMR spectroscopy studies have been completed on amyloid fragments and polypeptides, the most common techniques for studying amyloids and their formation are circular dichroism spectroscopy, fluorescence spectroscopy, and transmission electron microscopy, which provide limited structural information beyond identifying the presence of fibers. Additionally, a mechanistic understanding of amyloid formation requires temporal resolution so that intermediate states may be identified. This is especially true in the case of amyloid diseases, where growing evidence suggests that prefibrillar intermediates rather than the fibrils themselves may be the pathogenic species.

2D IR spectroscopy has been used to study numerous amyloid proteins, including human amylin (type II diabetes) (Shim *et al.*, 2009; Buchanan *et al.*, 2013), amyloid beta (Alzheimer's disease) (Kim *et al.*, 2008, 2009), polyglutamine (Huntington's disease) (Buchanan *et al.*, 2014), and crystallins (cataracts) (Moran *et al.*, 2012). A variety of labeling schemes are used in these studies ranging from uniform labeling to residue-specific labeling.

Structure of polyglutamine fibrils

Several neurogenerative diseases, including Huntington's disease, are associated with the formation of amyloid fibers by proteins containing mutationally-expanded polyglutamine tracts. A wide variety of structures has been proposed for the polyglutamine aggregates, but the two most widely supported structures are the β-arc and β-turn models (**Fig. 10**, top; Kar *et al.*, 2013; Kajava *et al.*, 2016). Each model has the polyglutamine sequence forming two antiparallel β-strands separated by a short turn, but differ in the backbone hydrogen-bonding structure. In the β-arc model, the two β-strands within each monomer contribute to the formation of

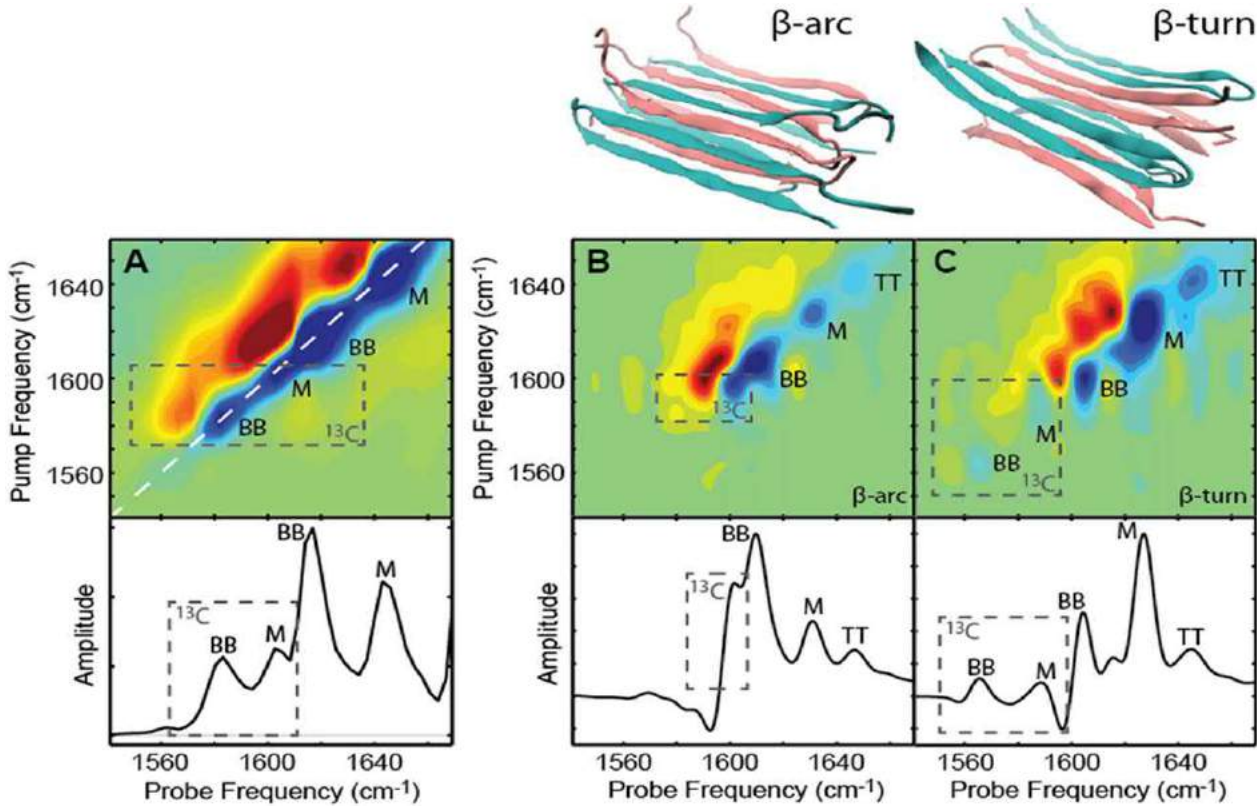

Fig. 10 Experimental and simulated 2D IR spectra of isotope-diluted Q24 fibrils. (A) Experimental 2D IR spectrum and diagonal intensity slice of fibrils formed from 10% ^{13}C-labeled Q24 mixed with natural abundance ^{12}C Q24. Simulated 2D IR spectra and slices for the (B) β-arc and (C) β-turn models of isotope-diluted Q24 fibrils of the same composition. The peaks are labeled BB for backbone modes, M for mixed backbone-sidechain modes, and TT for disordered modes at the turns and termini. Schematic visualizations of the fibril models are shown above their respective simulated spectra, with monomeric units shown in alternating colors for clarity. Adapted from Moran, S.D., Woys, A.M., Buchanan, L.E., *et al.*, 2012. Two-dimensional IR spectroscopy and segmental 13C labeling reveals the domain structure of human γD-crystallin amyloid fibrils. Proceedings of the National Academy of Sciences 109 (9), 3329–3334.

two separate β-sheets. In the β-turn model, the two β-strands contribute to the same β-sheet. In both structures, the side chains are intercalated to stabilize the stacked β-sheets.

While 2D IR spectra of Q24, a polyglutamine peptide with 24 glutamine residues, agree with the prediction that polyglutamine forms antiparallel β-sheets, infrared spectroscopy alone cannot differentiate between the two models described above because the backbone structures of the β-sheets are nearly identical between models. However, if it were possible to obtain the spectrum of an individual monomer within the fibril, the structure could be distinguished by determining whether the two β-strands were coupling to each other within a β-sheet (β-turn) or if they were isolated in separate β-sheets with minimal coupling between strands (β-arc). Buchanan *et al.* used mixtures of uniformly labeled Q24 peptides to vibrationally isolate monomers within the fibrils (Buchanan *et al.*, 2014). Q24 peptides were expressed in E. coli so they could be either uniformly ^{13}C-labeled or left at natural abundance ^{12}C. Samples were prepared of 10% ^{13}C-labeled Q24 mixed with 90% natural abundance Q24. At that ratio, it is statistically unlikely that any two ^{13}C-labeled peptides will be stacked consecutively within a fibril. Thus, the frequency of the ^{13}C amide I mode will reflect only couplings between strands of the same monomer. The experimental spectra (**Fig. 10(A)**) clearly identify the β-turn model as the structure for this peptide and show strong agreement with spectra computed from molecular dynamics simulations of these models (**Fig. 10(B)** and **(C)**).

Kinetics of amyloid formation by human amylin

Human amylin is a 37 amino acid polypeptide implicated in type II diabetes. Based on solid-state NMR measurements and molecular dynamics simulations, Tycko and co-workers proposed a model for amylin in which the monomers form two β-strands connected by a short disordered region (Luca *et al.*, 2007). The strands stack into two separate parallel β-sheets. This structure was confirmed by 2D IR measurements in which a series of fibers were formed in which a single amino acid had been labeled with ^{13}C=^{18}O via solid-phase peptide synthesis (Wang *et al.*, 2011). When amylin aggregates into in-register parallel β-sheets, the isotope labeled amino acids align within the sheets; this creates a column of coupled oscillators that form their own set of normal modes within the β-sheets, red-shifting the frequency of the isotope-labeled mode just as coupling within a β-sheet shifts the unlabeled amide I. Thus, the frequency of an isotope-labeled mode is an excellent indicator of whether the labeled residue adopts a β-sheet structure.

With the development of mid-IR pulse shaping, 2D IR spectra can be collected rapidly and continuously over the course of amyloid formation. Shim *et al.* used 2D IR spectroscopy to follow the aggregation of amylin (Shim *et al.*, 2009). They studied six separate samples of amylin, each having a different amino acid labeled with $^{13}C^{18}O$. As the disordered monomers assemble into parallel β-sheets, the uncoupled isotope-labeled feature at 1595 cm^{-1} disappears and a coupled isotope-labeled feature grows in between 1570 and 1585 cm^{-1}. By comparing the rate at which the coupled peak developed for each labeled residue, they determined the order in which the residues are incorporated into the fiber β-sheets. More recently, Buchanan and coworkers performed an analogous study that focused on labeling residues within the disordered turn of amylin (Buchanan *et al.*, 2013). They discovered an early β-sheet intermediate on the fibril formation pathway. The intermediate contains a transient β-sheet that is broken to form the disordered turn in the final fiber structure. The discovery of the intermediate represents a significant advance in the understanding of amyloid formation, as it both provides a rational for the initial lag phase observed in kinetics measurements before fibril β-sheets are observed and presents a target for future inhibitor studies and drug design.

Membrane proteins

Membrane proteins are involved in many basic cellular activities, such as ion transport and energy transduction (Engel and Gaub, 2008). Despite their prevalence in biology, our understanding of membrane proteins is limited by their hydrophobicity. 2D IR spectroscopy is capable of determining the structure of membrane proteins either solubilized by detergent or embedded in protein bilayer. Woys and coworkers used site-specific $^{13}C^{18}O$ labeling to measure the 2D IR lineshapes for 15 of the 18 residues of ovispirin, an antibiotic polypeptide that binds to the surfaces of membranes as an α-helix (Woys *et al.*, 2010). A regular oscillation was observed in inhomogenous linewidth with a 15 cm^{-1} difference between the largest and smallest linewidths. The period of the oscillation matches the periodicity of an α-helix. As the inhomogenous linewidth measures the structural disorder of the backbone and surrounding electrostatic environment, broader lineshapes indicate residues that lie closer to the membrane surface while narrower lineshapes indicate residues that lie within the hydrophobic membrane interior. Comparison with simulations allowed the orientation and depth of ovispirin within the membrane to be determined.

Beyond structural measurements, the application of 2D IR spectroscopy to ion channels, a special class of membrane proteins, has provided significant insight into their function.

Ion permeation through the K^+ ion channel

Potassium ion channels are responsible for setting the resting membrane potential of many cells and shaping the action potential of excitable cells such as neurons. Ion permeation through the channel is highly selective and determined by the selectivity filter, a narrow pore whose structure is highly conserved across all K^+ channels. The pore is lined by backbone carbonyls that coordinate the K^+ ions through the carbonyl oxygens. The series of carbonyls create four K^+ binding sites through which the ions must move sequentially. Kratchovil and coworkers used 2D IR spectroscopy to determine the mechanism for K^+ permeation through the KcsA ion channel (Kratochvil *et al.*, 2009). Previous experiments suggest that "knock-on" permeation, in which the pore is simultaneously occupied by two K^+ ions separated by water molecules, is the most likely mechanism for the transport of K^+ ions through the pore. However, these studies can not rule out alternative models such as "hard-knock" permeation, in which a pair of K^+ ions occupy adjacent sites in the pore with no intervening water molecules. The inherent picosecond time resolution of 2D IR spectroscopy allows it to provide a nearly instantaneous snapshot of the ion binding configurations in the pore. Using protein semisynthesis, two adjacent residues at the center of the poor were labeled with $^{13}C^{18}O$; these residues participate in the three of the four binding sites and their frequencies depend strongly on the occupancy of these sites. In combination with molecular dynamics simulations, the 2D IR data provides new experimental evidence that water and ions alternate through the pore during conduction and reveals a new conformation of the protein backbone that had only been observed in MD simulations.

Role of water in the influenza M2 channel

The M2 channel of the influenza virus transports protons across the viral envelope to acidify the interior, a process that is vital to the replication of the virus. Water is an integral part of the M2 proton channel as it is required for assisting proton conduction. The transmembrane domain of the M2 channel (M2TM) consists of a bundle of four identical α-helices that rearrange at acidic pH to open the channel and allow proton transport (Manor *et al.*, 2009). X-ray crystallography revealed a water cluster near the center of M2TM which is stabilized by the carbonyls of residues Gly34 and His37. Hochstrasser and coworkers used 2D IR spectroscopy to further explore the nature of the water within the channel (Ghosh *et al.*, 2011). Spectral diffusion measurements, in which the evolution of the 2D IR line shape is tracked as a function of the delay between pump and probe pulses, can be used to identify mobile water molecules. The rapid exchange of hydrogen bonds causes the vibrational frequencies of nearby amide groups to vary rapidly in time, leading to increasingly homogenous peaks as the waiting time between pulses is increased. They specifically probed the water dynamics at Gly24 by incorporating $^{13}C^{18}O$ isotope label at the backbone carbonyl. At pH 8, the water cluster near Gly24 is found to be immobilized on the 10 ps timescale as no appreciable spectral diffusion was observed. In contrast, at pH 5.2 when the channel is open, spectral diffusion is observed on the timescale of 1.3 ps, indicating the presence of highly mobile water molecules (**Fig. 11**). The flow of water in the open state of the channel thus facilitates the diffusion of the proton through the channel. A subsequent study found that channel-blocking drugs increase the mobility of water molecules within the channel, even at non-acidic pH (Ghosh *et al.*, 2014). Thus, drug binding leads to an increase in the entropy of the channel water, increasing the thermodynamic favorability of the drug-binding process as a whole.

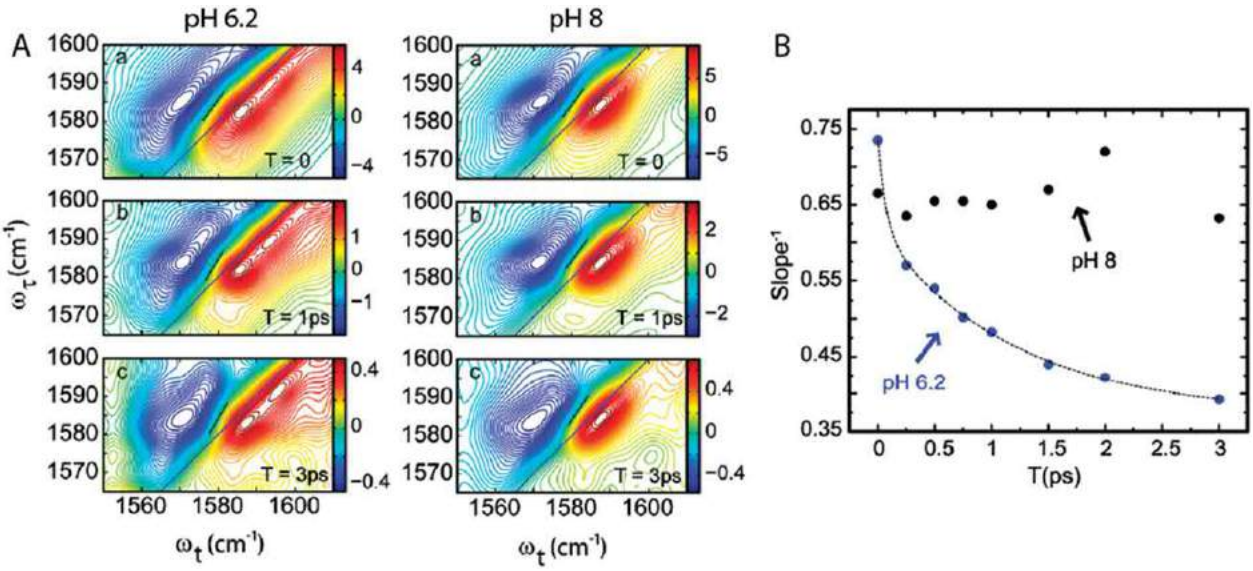

Fig. 11 Spectral diffusion measurements of the M2 proton channel. (A) 2D IR spectra of the isotope-labeled region of M2TM when Gly34 is labeled with $^{13}C^{18}O$. Spectra are collected at pH 6.2 (left) and pH 8.0 (right) at waiting times between 0 and 3 ps. Nodal slopes are indicated with black lines. (B) Plot of inverse nodal slope vs waiting time for M2TM. Adapted from Ghosh, A., Qiu, J., Degrado, W.F., Hochstrasser, R.M., 2011. Tidal surge in the M2 proton channel, sensed by 2D IR spectroscopy. Proceedings of the National Academy of Sciences 108 (15), 6115–6120.

Dynamics of Water

Although water is critical for many chemical and biological processes and therefore has been studied extensively using both experimental and theoretical approaches, much about water remains unclear. While repulsive forces typically dictate the structure of liquids, extensive hydrogen bonding causes water to exhibit many anomalous properties. Each water molecule can accept and donate two hydrogen bonds which rapidly interchange on the timescale of tens of femtoseconds to picoseconds (Roberts *et al.*, 2009). This rapidly evolving hydrogen bond network governs many aqueous processes including ion solvation and proton transport.

2D IR spectroscopy has provided a wealth of insight into the hydrogen bond switching dynamics of water (Loparo *et al.*, 2006). 2D IR anisotropy measurements of the OH stretching vibration of HOD were used to track the reorientation of water molecules as they undergo hydrogen bond exchange (Roberts *et al.*, 2009; Ramasesha *et al.*, 2011). The results suggest that the water hydrogen bond network rearranges through concerted motions that large angle molecular reorientations. Other 2D IR studies have explored the slowing of hydrogen bond dynamics as liquid water is cooled from ambient conditions to the metastable supercooled regime (Perakis and Hamm, 2011; Kraemer *et al.*, 2008). Beyond studies of pure water, 2D IR spectroscopy has been used to observe the exchange of hydrogen bonds between water molecules and ions (Moilanen *et al.*, 2009). These studies show that while water dynamics are slowed in salt solutions compared to pure water, they are only slowed by a factor of approximately 2. The influx of new insights into water structure and dynamics from 2D IR spectroscopy, as well as other ultrafast nonlinear optical techniques, has raised many questions about the suitability of existing water models and led theorists to develop new models for water to better explain experimental observations (Schmidt *et al.*, 2007; Ni and Skinner, 2014).

A recent 2D IR study by the Tokmakoff group focused on the proton-transport mechanism in water (Thamer *et al.*, 2015). Proton transport governs most biological redox processes as well as acid-base chemistry but little was known about how the water hydrogen-bonding network accommodates excess protons. The primary point of debate involved the dominant structure of the proton-water complex and how various species interconvert during proton transport. By exciting O-H stretching vibrations and detecting the response over a broad spectral window (1500–4000 cm^{-1}), they observed cross peaks at 3200 and 1760 cm^{-1} which are characteristic of the water stretching and bending vibrations of the Zundel complex $H(H_2O)_2^+$, a structure in which the proton is shared equally between two water molecules (**Fig. 12(A)**). The lifetime of the Zundel complex was determined to be at least 480 fs (**Fig. 12(B)**); this is long for the complex to be merely a transition state, which suggests it must play a key role in aqueous proton transfer.

Perspective

2D IR spectroscopy continues to rapidly progress with new implementations and applications. Several new directions have already emerged and could lead to major breakthroughs in the next 10 years.

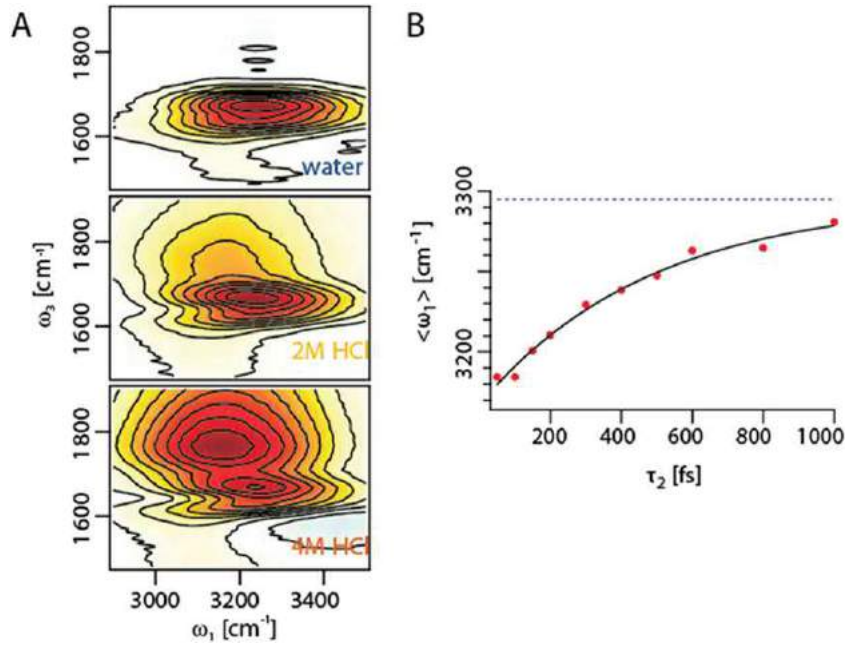

Fig. 12 2D IR crosspeaks reveal existence of Zundel complex in water. (A) Background subtracted 2D IR spectra for water, 2M HCl, and 4M HCl. A previously unobserved cross peak corresponding to the Zundel bend transition appears at $\omega_3 = 1760$ cm^{-1} in the acid spectra and increases linearly in intensity with acid concentration. (B) As waiting time increases, the Zundel cross peak blue-shifts along the excitation axis, indicating that either proton transport or vibrational energy transfer is occurring between bulk water and the Zundel complex on the time scale of 480 fs. Adapted from Thamer, M., De Marco, L., Ramasesha, K., *et al.*, 2015. Ultrafast 2D IR spectroscopy of the excess proton in liquid water. Science 350 (6256), 78–82.

Surface- and Interface-Sensitive 2D Vibrational Spectroscopy

By combining 2D IR spectroscopy with surface- and interface-sensitive techniques, intrinsic surface- and interface-sensitive 2D vibrational spectroscopic techniques have been developed. These techniques have found applications in the study of catalytic surfaces (Xiong *et al.*, 2011; Wang *et al.*, 2015; Li *et al.*, 2016a,b), protein-functionalized surfaces (Laaser and Zanni, 2013; Ghosh *et al.*, 2015; Laaser *et al.*, 2014; Ho *et al.*, 2015), and air/water interfaces (Hsieh *et al.*, 2014; Piatkowski *et al.*, 2014; Livingstone *et al.*, 2016; Bredenbeck *et al.*, 2009, 2008; Zhang *et al.*, 2011a,b; Singh *et al.*, 2012, 2016; Inoue *et al.*, 2015; Nihonyanagi *et al.*, 2013) which are difficult to be study using bulk techniques. There are two main versions of surface- and interface-sensitive 2D vibrational spectroscopy. The Zanni, Bonn, Tahara, and Xiong groups implemented 2D vibrational sum frequency generation (2D SFG) spectroscopy, which combines the pulse sequences of 2D IR and SFG. The detected 2D SFG signal is a fourth-order nonlinear optical response; thus, it is intrinsically sensitive to surfaces and interfaces. An alternative approach is 2D attenuated total reflection (ATR) spectroscopy, developed by the Hamm group (Kraack *et al.*, 2014, 2015, 2016; Lotti *et al.*, 2016; Kraack and Hamm, 2016b). This technique applies the 2D IR pulse sequence at the surface of an ATR crystal, taking advantage of evanescent waves at the crystal surface to achieve interface- and surface-sensitivity. With improved signal-to-noise, these surface- and interface-sensitive 2D vibrational techniques are expected to unravel molecular conformations and dynamics at complex interfaces. An excellent review article has become available on this specific topic just recently (Kraack and Hamm, 2016a).

Broadband Light Sources

The Tokmakoff, Khalil and Peterson groups have worked to develop broadband IR laser pulses for 2D IR applications (Petersen and Tokmakoff, 2010; Calabrese *et al.*, 2012; Cheng *et al.*, 2012; Balasubramanian *et al.*, 2016). These broadband light sources are based on filamentation in gaseous media. When a 800 nm pulse and its harmonics are focused into the gaseous media, pulses that span from <400 to <5000 cm^{-1} have been realized (Calabrese *et al.*, 2012). It is feasible to compress the pulses close to the transform limit using a deformable mirror pulse shaper (Balasubramanian *et al.*, 2016). The broadband IR pulses have sufficiently high pulse energies to be used as IR_3 in the 2D IR pulse sequence to obtain partial broadband 2D IR spectra. This approach has been used to study couplings between stretching and bending motions of water in Zundel complex form (Thamer *et al.*, 2015) and electronic vibrational coupling (Gaynor *et al.*, 2016). In the future, there is great interest in developing broadband IR pulses with μJ pulse energies to enable full broadband 2D IR spectroscopy.

2D IR Microscopy

As with many spectroscopic techniques, developing imaging capabilities to spatially visualize samples with 2D IR spectral signals is an emerging field. This development will allow spatial heterogeneity to be resolved in 2D IR spectra and also enrich microscope

imaging with molecular information that is not available from traditional absorption- or fluorescence-based optical microscopy. To the best of our knowledge, there are only two pioneering 2D IR microscopy reports (Baiz *et al.*, 2014; Ostrander *et al.*, 2016). Both used a pulse shaper based approach to collect 2D IR spectra. FPA detectors reduced data acquisition times and allowed 3.48 μm spatial resolution to be achieved (Ostrander *et al.*, 2016). Although still in its infancy, these proof-of-principle works demonstrate the feasibility of collecting microscope images with 2D IR spectral information and this field could blossom in the next 10 years.

2D Electronic-Vibrational Spectroscopy and 2D Vibrational-Electronic Spectroscopy

Another notable development related to 2D IR spectroscopy is 2D Electronic-Vibrational (EV) spectroscopy and 2D Vibrational-Electronic (VE) spectroscopy, which probe the coupling between electronic and vibrational excitations of molecules. The Fleming group has pioneered 2D EV spectroscopy, in which electronic states of molecules are excited and the subsequent vibrational motions are probed. They showed that this technique is useful in understanding the interplay between electronic states and vibrational manifolds responsible for nonradioactive electronic relaxation pathways (Oliver *et al.*, 2014). 2D EV spectroscopy has also revealed electronic energy transfer between segments in photosynthetic complexes (Lewis and Fleming, 2016; Lewis *et al.*, 2016). Theoretical frameworks were developed to quantify correlated electronic and vibrational dynamics and to correlate electronic state populations with structural information (Lewis *et al.*, 2015a,b). 2D VE spectroscopy, developed by the Khalil group, is complementary to 2D EV spectroscopy. 2D VE spectroscopy excites vibrational modes of molecules and probes electronic states. They used this technique to study organometallic charge transfer compounds and were able to distinguish coherent and incoherent vibrational energy transfer among coupled vibrational modes and charge transfer transitions (Courtney *et al.*, 2015a,b). These pioneering studies demonstrate that 2D VE spectroscopy can reveal how molecular vibrations modulate energy and charge transfer in molecular systems. Together, 2D EV and 2D VE spectroscopy extend the ability of 2D spectroscopic techniques from probing dynamics and couplings between the same type of molecular motions to probing interactions between molecular motions with different energy and time scales.

Acknowledgements

Lauren E. Buchanan and Wei Xiong thank all the researchers who contributed to the development of 2D IR spectroscopy in the past 18 years. Lauren E. Buchanan and Wei Xiong thank Prof. Martin T. Zanni for being a fantastic mentor and for introducing them to 2D IR spectroscopy. Wei Xiong acknowledges his group members and the financial support from UC San Diego. Lauren E. Buchanan acknowledges her group members and financial support from Vanderbilt University.

See also: Tutorial on Multidimensional Coherent Spectroscopy

References

Anderson, N.A., Lian, T., 2005. Ultrafast electron transfer at the molecule–semiconductor nanoparticle interface. Annual Review of Physical Chemistry 56, 491–519.
Anna, J.M., Kubarych, K.J., 2010. Watching solvent friction impede ultrafast barrier crossings: A direct test of Kramers theory. Journal of Chemical Physics 133, 174506.
Anna, J.M., Ross, M.R., Kubarych, K.J., 2009. Dissecting enthalpic and entropic barriers to ultrafast equilibrium isomerization of a flexible molecule using 2DIR chemical exchange spectroscopy. Journal of Physical Chemistry A 113, 6544–6547.
Asbury, J.B., Steinel, T., Stromberg, C., *et al.*, 2004. Water dynamics: Vibrational echo correlation spectroscopy and comparison to molecular dynamics simulations. Journal of Physical Chemistry A 108, 1107–1119.
Asplund, M.C., Zanni, M.T., Hochstrasser, R.M., 2000. Two-dimensional infrared spectroscopy of peptides by phase-controlled femtosecond vibrational photon echoes. Proceedings of the National Academy of Sciences 97 (15), 8219–8224.
Auer, B., Kumar, R., Schmidt, J.R., Skinner, J.L., 2007. Hydrogen bonding and Raman, IR, and 2D-IR spectroscopy of dilute HOD in liquid D_2O. Proceedings of the National Academy of Sciences 104 (36), 14215–14220.
Baiz, C.R., Schach, D., Tokmakoff, A., 2014. Ultrafast 2D IR microscopy. Optics Express 22 (15), 18724.
Bakulin, A.A., Liang, C., Jansen, T.L.C., *et al.*, 2009. Hydrophobic solvation: A 2D IR spectroscopic inquest. Accounts of Chemical Research 42 (9), 1229–1238.
Balasubramanian, M., Courtney, T.L., Gaynor, J.D., Khalil, M., 2016. Compression of tunable broadband mid-IR pulses with a deformable mirror pulse shaper. Journal of the Optical Society of America B 33 (10), 2033–2037.
Barbour, L.W., Hegadorn, M., Asbury, J.B., 2006. Microscopic inhomogeneity and ultrafast orientational motion in an organic photovoltaic bulk heterojunction thin film studied with 2D IR vibrational spectroscopy. Journal of Physical Chemistry B 110, 24281–24286.
Bredenbeck, J., Ghosh, A., Nienhuys, H., Bonn, M., 2009. Interface-specific ultrafast two-dimensional vibrational spectroscopy. Accounts of Chemical Research 42 (9), 1332–1342.
Bredenbeck, J., Ghosh, A., Smits, M., Bonn, M., 2008. Ultrafast two dimensional-infrared spectroscopy of a molecular monolayer. Journal of the American Chemical Society 130, 2152–2153.
Bredenbeck, J., Helbing, J., Hamm, P., 2004. Labeling vibrations by light: Ultrafast transient 2D-IR spectroscopy tracks vibrational modes during photoinduced charge transfer. Journal of the American Chemical Society 126, 990–991. Available at: http://dx.doi.org/10.1063/1.4932983.
Brinzer, T., Berquist, E.J., Ren, Z., *et al.*, 2015. Ultrafast vibrational spectroscopy (2D-IR) of CO_2 in ionic liquids: Carbon capture from carbon dioxide's point of view. Journal of Chemical Physics 142, 212425.

Buchanan, L.E., Carr, J.K., Fluitt, A.M., *et al*., 2014. Structural motif of polyglutamine amyloid fibrils discerned with mixed-isotope infrared spectroscopy. Proceedings of the National Academy of Sciences 111 (16), 5796–5801.
Buchanan, L.E., Dunkelberger, E.B., Tran, H.Q., *et al*., 2013. Mechanism of IAPP amyloid fibril formation involves an intermediate with a transient beta-sheet. Proceedings of the National Academy of Sciences 110 (48), 19285–19290.
Buchanan, L.E., Dunkelberger, E.B., Zanni, M.T., 2012. Examining amyloid structure and kinetics with 1D and 2D infrared spectroscopy and isotope Labeling. In: Fabian, H., Naumann, D. (Eds.), Protein Folding and Misfolding: Shining Light by Infrared Spectroscopy, vol. 1. Springer, pp. 217–237.
Cahoon, J.F., Sawyer, K.R., Schlegel, J.P., Harris, C.B., 2008. Determining transition-state geometries in liquids using 2D-IR. Science 319, 1820–1823.
Calabrese, C., Stingel, A.M., Shen, L., Petersen, P.B., 2012. Ultrafast continuum mid-infrared spectroscopy: Probing the entire vibrational spectrum in a single laser shot with femtosecond time resolution. Optics Letters 37 (12), 2265–2267.
Cerullo, G., De, S.S., 2003. Ultrafast optical parametric amplifiers. Review of Scientific Instruments 74 (1), 1–18.
Cervetto, V., Helbing, J., Bredenbeck, J., Hamm, P., 2004. Double-resonance versus pulsed Fourier transform two-dimensional infrared spectroscopy: An experimental and theoretical comparison. Journal of Chemical Physics 121 (12), 5935–5942.
Cheng, M., Reynolds, A., Widgren, H., Khalil, M., 2012. Generation of tunable octave-spanning mid-infrared pulses by filamentation in gas media. Optics Letters 37 (11), 1787–1789.
Chiti, F., Dobson, C.M., 2006. Protein misfolding, functional amyloid, and human disease. Annual Review of Biochemistry 75, 333–366.
Cho, M., 2009. Two-Dimensional Optical Spectroscopy. CRC Press.
Courtney, T.L., Fox, Z.W., Estergreen, L., Khalil, M., 2015. Measuring coherently coupled intramolecular vibrational and charge-transfer dynamics with two-dimensional vibrational-electronic spectroscopy. Journal of Physical Chemistry Letters 6, 1286–1292.
Courtney, T.L., Fox, Z.W., Slenkamp, K.M., Khalil, M., 2015b. Two-dimensional vibrational-electronic spectroscopy T. Journal of Chemical Physics 143, 154201. Available at: http://dx.doi.org/10.1063/1.4932983.
Crabtree, R.H., 2004. Organometallic alkane CH activation. Journal of Organometallic Chemistry 689, 4083–4091.
Cui, Y., Fulfer, K.D., Ma, J., *et al*., 2016. Solvation dynamics of an ionic probe in chloride-based deep eutectic solvents. Physical Chemistry Chemical Physics 18, 31471–31479.
Demirdöven, N., Khalil, M., Tokmakoff, A., 2002. Correlated vibrational dynamics revealed by two-dimensional infrared spectroscopy. Physical Review Letters 89 (23), 237401.
Donaldson, P.M., Hamm, P., 2013. Gold nanoparticle capping layers: Structure, dynamics, and surface enhancement measured using 2D-IR spectroscopy. Angewandte Chemie – International Edition 52 (2), 634–638.
Dutta, S., Ren, Z., Brinzer, T., Garrett-Roe, S., 2015. Two-dimensional ultrafast vibrational spectroscopy of azides in ionic liquids reveals solute-specific solvation. Physical Chemistry Chemical Physics 17, 26575–26579.
Eaves, J.D., Loparo, J.J., Fecko, C.J., *et al*., 2005. Hydrogen bonds in liquid water are broken only fleetingly. Proceedings of the National Academy of Sciences 102 (37), 13019–13022.
Engel, A., Gaub, H.E., 2008. Structure and mechanics of membrane proteins. Annual Review of Biochemistry 77, 127–148.
Fayer, M.D., 2009. Dynamics of liquids, molecules, and proteins measured with ultrafast 2D IR vibrational echo chemical exchange spectroscopy. Annual Review of Physical Chemistry 60, 21–38.
Fayer, M.D., 2013. Ultrafast Infrared Vibrational Spectroscopy. CRC Press.
Fulfer, K.D., Kuroda, D.G., 2016. Solvation structure and dynamics of the lithium ion in organic carbonate-based electrolytes: A time-dependent infrared spectroscopy study. Journal of Physical Chemistry C 120, 24011–24022.
Fulmer, E.C., Mukherjee, P., Krummel, A.T., Zanni, M.T., 2004. A pulse sequence for directly measuring the anharmonicities of coupled vibrations: Two-quantum two-dimensional infrared spectroscopy. Journal of Chemical Physics 120 (7), 8067–8078.
Ganim, Z., Chung, H.O.I.S., Smith, A.W., *et al*., 2008. Amide I two-dimensional infrared spectroscopy of proteins. Accounts of Chemical Research 41 (3), 432–441.
Gaynor, J.D., Courtney, T.L., Balasubramanian, M., Khalil, M., 2016. Fourier transform two-dimensional electronic-vibrational spectroscopy using an octave-spanning mid-IR probe. Optics Letters 41 (12), 2895–2898.
Ghosh, A., Ho, J., Serrano, A.L., *et al*., 2015. Two-dimensional sum-frequency generation (2D SFG) spectroscopy: Summary of principles and its application to amyloid fiber monolayers. Faraday Discussions 177, 493–505.
Ghosh, A., Qiu, J., Degrado, W.F., Hochstrasser, R.M., 2011. Tidal surge in the M2 proton channel, sensed by 2D IR spectroscopy. Proceedings of the National Academy of Sciences 108 (15), 6115–6120.
Ghosh, A., Wang, J., Moroz, Y.S., *et al*., 2014. 2D IR spectroscopy reveals the role of water in the binding of channel-blocking drugs to the influenza M2 channel. Journal of Chemical Physics 140, 235105.
Goldschmidt, L., Teng, P.K., Riek, R., Eisenberg, D., 2010. Identifying the amylome, proteins capable of forming amyloid-like fibrils. Proceedings of the National Academy of Sciences 107 (8), 3487–3492.
Golonzka, O., Khalil, M., Demirdöven, N., Tokmakoff, A., 2015. Vibrational anharmonicities revealed by coherent two-dimensional infrared spectroscopy. Physical Review Letters 86 (10), 2154–2157.
Hamm, P., 2006. Three-dimensional-IR spectroscopy: Beyond the two-point frequency fluctuation correlation function. Journal of Chemical Physics 124, 124506.
Hamm, P., Helbing, J., Bredenbeck, J., 2008. Two-dimensional infrared spectroscopy of photoswitchable peptides. Annual Review of Physical Chemistry 59, 291–317.
Hamm, P., Lim, M., DeGrado, W.F., Hochstrasser, R.M., 2000. Pump/probe self heterodyned 2D spectroscopy of vibrational transitions of a small globular peptide. Journal of Chemical Physics 112 (4), 1907–1916.
Hamm, P., Lim, M., Hochstrasser, R.M., 1998. Structure of the amide I band of peptides measured by femtosecond nonlinear-infrared spectroscopy. Journal of Physical Chemistry B 102, 6123–6138.
Hamm, P., Zanni, M.T., 2011. Concepts and Methods of 2D Infrared Spectroscopy. New York, NY: Cambridge University Press.
Hasegawa, T., Tanimura, Y., 2008. Nonequilibrium molecular dynamics simulations with a backward-forward trajectories sampling for multidimensional infrared spectroscopy of molecular vibrational modes. Journal of Chemical Physics 128, 64511.
Hochstrasser, R.M., 2001. Two-dimensional IR-spectroscopy: Polarization anisotropy effects. Chemical Physics 266, 273–284.
Ho, J., Sko, D.R., Ghosh, A., Zanni, M.T., 2015. Structural characterization of single-stranded DNA monolayers using two-dimensional sum frequency generation spectroscopy. Journal of Physical Chemistry B 119, 105868–110596.
Hsieh, C., Okuno, M., Hunger, J., *et al*., 2014. Aqueous heterogeneity at the air/water interface revealed by 2D-HD-SFG spectroscopy. Angewandte Chemie International Edition 53, 8146–8149.
Huber, C.J., Egger, S.M., Spector, I.C., *et al*., 2015. 2D-IR spectroscopy of porous silica nanoparticles: Measuring the distance sensitivity of spectral diffusion. Journal of Physical Chemistry C 119, 25135–25144.
Huber, C.J., Massari, A.M., 2014. Characterizing solvent dynamics in nanoscopic silica sol – gel glass pores by 2D-IR spectroscopy of an intrinsic vibrational probe. Journal of Physical Chemistry C 118, 25567–25578.
Hunt, N.T., 2009. 2D-IR spectroscopy: Ultrafast insights into biomolecule structure and function. Chemical Society Reviews 38 (7), 1837–1848.
Hybl, J.D., Yu, A., Farrow, D.A., Jonas, D.M., 2002. Polar solvation dynamics in the femtosecond evolution of two-dimensional fourier transform spectra. Journal of Physical Chemistry A 106, 7651–7654.

Inoue, K., Nihonyanagi, S., Singh, P.C., *et al.*, 2015. 2D heterodyne-detected sum frequency generation study on the ultrafast vibrational dynamics of H_2O and HOD water at charged interfaces. Journal of Chemical Physics 142, 212431.
Ishizaki, A., Tanimura, Y., 2006. Modeling vibrational dephasing and energy relaxation of intramolecular anharmonic modes for multidimensional infrared spectroscopies. Journal of Chemical Physics 125, 84501.
Ishizaki, A., Tanimura, Y., 2007. Dynamics of a multimode system coupled to multiple heat baths probed by two-dimensional infrared spectroscopy. Journal of Physical Chemistry A 111, 9269–9276.
Jones, B.H., Huber, C.J., Massari, A.M., 2011. Solvation dynamics of Vaska's complex by 2D-IR spectroscopy. Journal of Physical Chemistry C 115, 24813–24822.
Kajava, A.V., Baxa, U., Steven, A.C., 2016. **B**-arcades: Recurring motifs in naturally occurring and disease-related amyloid fibrils. The FASEB Journal 24 (5), 1311–1319.
Kar, K., Hoop, C.L., Drombosky, K.W., *et al.*, 2013. β-Hairpin-mediated nucleation of polyglutamine amyloid formation. Journal of Molecular Biology 425 (7), 1183–1197.
Khalil, M., Demirdo, N., Tokmakoff, A., 2003a. Obtaining absorptive line shapes in two-dimensional infrared vibrational correlation spectra. Physical Review Letters 90 (4), 47401.
Khalil, M., Demirdoven, N., Tokmakoff, A., 2003b. Coherent 2D IR spectroscopy: Molecular structure and dynamics in solution. Journal of Physical Chemistry A 107, 5258–5279.
Kim, Y.S., Hochstrasser, R.M., 2005. Chemical exchange 2D IR of hydrogen-bond making and breaking. Proceedings of the National Academy of Sciences 102 (32), 11185–11190.
Kim, Y.S., Hochstrasser, R.M., 2009. Applications of 2D IR spectroscopy to peptides, proteins, and hydrogen-bond dynamics. Journal of Physical Chemistry B 113 (24), 8231–8251.
Kim, Y.S., Liu, L., Axelsen, P.H., Hochstrasser, R.M., 2008. Two-dimensional infrared spectra of isotopically diluted amyloid fibrils from Abeta40. Proceedings of the National Academy of Sciences of the United States of America 105, 7720–7725.
Kim, Y.S., Liu, L., Axelsen, P.H., Hochstrasser, R.M., 2009. 2D IR provides evidence for mobile water molecules in beta-amyloid fibrils. Proceedings of the National Academy of Sciences of the United States of America 106 (42), 17751–17756.
Kraack, J.P., Hamm, P., 2016a. Surface-sensitive and surface-specific ultrafast two-dimensional vibrational spectroscopy. Chemical Reviews 117 (16), 10623–10664. doi:10.1021/acs.chemrev.6b00437.
Kraack, J.P., Hamm, P., 2016b. Vibrational ladder-climbing in surface-enhanced, ultrafast infrared spectroscopy. Physical Chemistry Chemical Physics 18, 16088–16093.
Kraack, J.P., Kaech, A., Hamm, P., 2016. Surface enhancement in ultrafast 2D ATR IR spectroscopy at the metal–liquid interface. Journal of Physical Chemistry C 120, 3350–3359.
Kraack, J.P., Lotti, D., Hamm, P., 2014. Ultrafast, multidimensional attenuated total reflectance spectroscopy of adsorbates at metal surfaces. Journal of Physical Chemistry Letters 5, 2325–2329.
Kraack, J.P., Lotti, D., Hamm, P., 2015. 2D attenuated total reflectance infrared spectroscopy reveals ultrafast vibrational dynamics of organic monolayers at metal–liquid interfaces. Journal of Chemical Physics 142, 212413.
Kraemer, D., Cowan, M.L., Paarmann, A., *et al.*, 2008. Temperature dependence of the two-dimensional infrared spectrum of liquid H_2O. Proceedings of the National Academy of Sciences 105 (2), 437–442.
Kratochvil, H.T., Carr, J.K., Matulef, K., *et al.*, 2009. Instantaneous ion configurations in the K^+ ion channel selectivity filter revealed by 2D IR spectroscopy. Science 353, 1040–1044.
Kumar, B., Llorente, M., Froehlich, J., *et al.*, 2012. Photochemical and photoelectrochemical reduction of CO_2. Annual Review of Physical Chemistry 63, 541–569.
Kurochkin, D.V., Naraharisetty, S.R.G., Rubtsov, I.V., 2007. A relaxation-assisted 2D IR spectroscopy method. Proceedings of the National Academy of Sciences 104 (36), 14209–14214.
Kwac, K., Cho, M., 2003a. Molecular dynamics simulation study of N-methylacetamide in water. II. Two-dimensional infrared pump–probe spectra. Journal of Chemical Physics 119 (4), 2256–2263.
Kwac, K., Cho, M., 2003b. Two-color pump–probe spectroscopies of two- and three-level systems: 2-Dimensional line shapes and solvation dynamics. Journal of Physical Chemistry A 107, 5903–5912.
Kwac, K., Lee, H., Cho, M., 2004. Non-gaussian statistics of amide I mode frequency fluctuation of N-methylacetamide in methanol solution: Linear and nonlinear vibrational spectra. Journal of Chemical Physics 120 (3), 1477–1490.
Kwon, Y., Park, S., 2015. Complexation dynamics of CH_3SCN and Li^+ in acetonitrile studied by two-dimensional infrared spectroscopy. Physical Chemistry Chemical Physics 17, 24193–24200.
Laaser, J.E., Sko, D.R., Ho, J., *et al.*, 2014. Two-dimensional sum-frequency generation reveals structure and dynamics of a surface-bound peptide. Journal of the American Chemical Society 136, 956–962.
Laaser, J.E., Zanni, M.T., 2013. Extracting structural information from the polarization dependence of one- and two-dimensional sum frequency generation spectra. Journal of Physical Chemistry A 117, 5875–5890.
Lewis, N.H.C., Dong, H., Oliver, T.A.A., Fleming, G.R., 2015a. A method for the direct measurement of electronic site populations in a molecular aggregate using two-dimensional electronic-vibrational spectroscopy. Journal of Chemical Physics 143, 124203.
Lewis, N.H.C., Dong, H., Oliver, T.A.A., Fleming, G.R., 2015b. Measuring correlated electronic and vibrational spectral dynamics using line shapes in two-dimensional electronic-vibrational spectroscopy. Journal of Chemical Physics 142, 174202.
Lewis, N.H.C., Fleming, G.R., 2016. Two-dimensional electronic-vibrational spectroscopy of chlorophyll a and b. Journal of Physical Chemistry Letters 7, 831–837.
Lewis, N.H.C., Gruenke, N.L., Oliver, T.A.A., *et al.*, 2016. Observation of electronic excitation transfer through light. Journal of Physical Chemistry Letters 7, 4197–4206.
Lin, Z., Rubtsov, I.V., 2012. Constant-speed vibrational signaling along polyethyleneglycol chain up to 60-Å distance. Proceedings of the National Academy of Sciences 109 (5), 1413–1418.
Livingstone, R.A., Zhang, Z., Piatkowski, L., *et al.*, 2016. Water in contact with a cationic lipid exhibits bulklike vibrational dynamics. Journal of Physical Chemistry B 120, 10069–10078.
Li, Y., Wang, J., Clark, M.L., *et al.*, 2016a. Characterizing interstate vibrational coherent dynamics of surface adsorbed catalysts by fourth-order 3D SFG spectroscopy. Chemical Physics Letters 650, 1–6.
Li, Z., Wang, J., Li, Y., Xiong, W., 2016b. Solving the "magic angle" challenge in determining molecular orientation heterogeneity at interfaces. Journal of Physical Chemistry C 120 (36), 20239–20246.
Loparo, J.J., Roberts, S.T., Tokmakoff, A., *et al.*, 2006. Multidimensional infrared spectroscopy of water. II. Hydrogen bond switching dynamics multidimensional infrared spectroscopy of water. II. Hydrogen bond switching dynamics. The Journal of Chemical Physics 125, 194522.
Lotti, D., Hamm, P., Kraack, J.P., 2016. Surface-sensitive spectro-electrochemistry using ultrafast 2D ATR IR spectroscopy. Journal of Physical Chemistry C 120, 2883–2892.
Luca, S., Yau, W., Leapman, R., Tycko, R., 2007. Peptide conformation and supramolecular organization in amylin fibrils: Constraints from solid-state NMR. Biochemistry 46, 13505–13522.
Luther, B.M., Tracy, K.M., Gerrity, M., *et al.*, 2016. 2D IR spectroscopy at 100 kHz utilizing a Mid-IR OPCPA laser source. Optics Express 24 (4), 10095–10100.
Maekawa, H., De Poli, M., Toniolo, C., Ge, N.-H., 2009. Couplings between peptide linkages across a 3(10)-helical hydrogen bond revealed by two-dimensional infrared spectroscopy. Journal of the American Chemical Society 131 (6), 2042–2043.
Manor, J., Mukherjee, P., Lin, Y., *et al.*, 2009. Article gating mechanism of the influenza A M2 channel revealed by 1D and 2D IR spectroscopies. Structure/Folding and Design 17 (2), 247–254. Available at: http://dx.doi.org/10.1016/j.str.2008.12.015.

Middleton, C.T., Woys, A.M., Mukherjee, S.S., Zanni, M.T., 2010. Residue-specific structural kinetics of proteins through the union of isotope labeling, mid-IR pulse shaping, and coherent 2D IR spectroscopy. Methods 52 (1), 12–22.
Moilanen, D.E., Wong, D., Rosenfeld, D.E., *et al.*, 2009. Ion–water hydrogen-bond switching observed with 2D IR vibrational echo chemical exchange spectroscopy. Proceedings of the National Academy of Sciences 106 (2), 375–380.
Moran, S.D., Woys, A.M., Buchanan, L.E., *et al.*, 2012. Two-dimensional IR spectroscopy and segmental 13C labeling reveals the domain structure of human γD-crystallin amyloid fibrils. Proceedings of the National Academy of Sciences 109 (9), 3329–3334.
Mukamel, S., 2000. Multidimensional femtosecond correlation spectroscopies of electronic and vibrational excitations. Annual Review of Physical Chemistry 51, 691–729.
Mukherjee, P., Kass, I., Arkin, I.T., Zanni, M.T., 2005. Picosecond dynamics of a membrane protein revealed by 2D IR. Proceedings of the National Academy of Sciences 103 (10), 3528–3533.
Nee, M.J., Baiz, C.R., Anna, J.M., *et al.*, 2008. Multilevel vibrational coherence transfer and wavepacket dynamics probed with multidimensional IR spectroscopy. Journal of Chemical Physics 129, 84503.
Nee, M.J., Mccanne, R., Kubarych, K.J., 2007. Two-dimensional infrared spectroscopy detected by chirped pulse upconversion. Optics Letters 32 (6), 713–715.
Nihonyanagi, S., Mondal, J.A., Yamaguchi, S., Tahara, T., 2013. Structure and dynamics of interfacial water studied by vibrational sum-frequency generation. Annual Review of Physical Chemistry 64, 579–603.
Nishida, J., Tamimi, A., Fei, H., *et al.*, 2014. Structural dynamics inside a functionalized metal–organic framework probed by ultrafast 2D IR spectroscopy. Proceedings of the National Academy of Sciences 111 (52), 18442–18447.
Nishida, J., Yan, C., Fayer, M.D., 2014. Dynamics of molecular monolayers with different chain lengths in air and solvents probed by ultrafast 2D IR spectroscopy. Journal of Physical Chemistry C 118, 523–532.
Nishida, J., Yan, C., Fayer, M.D., 2016. Orientational dynamics of a functionalized alkyl planar monolayer probed by polarization-selective angle-resolved infrared pump – probe spectroscopy. Journal of the American Chemical Society 138, 14057–14065.
Ni, Y., Skinner, J.L., 2014. Ultrafast pump–probe and 2DIR anisotropy and temperature-dependent dynamics of liquid water within the E3B model. Journal of Chemical Physics 141, 24509.
Oliver, T.A.A., Lewis, N.H.C., Graham, R., 2014. Correlating the motion of electrons and nuclei with two-dimensional electronic-vibrational spectroscopy. Proceedings of the National Academy of Sciences 111 (28), 10061–10066.
Ostrander, J.S., Serrano, A.L., Ghosh, A., Zanni, M.T., 2016. Spatially resolved two-dimensional infrared spectroscopy via wide- field microscopy. ACS Photonics 3, 1315–1323.
Oudenhoven, T.A., Joo, Y., Laaser, J.E., *et al.*, 2015. Dye aggregation identified by vibrational coupling using 2D IR spectroscopy. Journal of Chemical Physics 142, 212449.
Paarmann, A., Hayashi, T., Mukamel, S., Miller, R.J.D., 2009. Nonlinear response of vibrational excitons: Simulating the two-dimensional infrared spectrum of liquid water. Journal of Chemical Physics 130, 204110.
Park, K., Cho, M., 1998. Time- and frequency-resolved coherent two-dimensional IR spectroscopy: Its complementary relationship with the coherent two-dimensional Raman scattering spectroscopy. Journal of Chemical Physics 109 (24), 10559–10569.
Perakis, F., Hamm, P., 2011. Two-dimensional infrared spectroscopy of supercooled water. Journal of Physical Chemistry B 115, 5289–5293.
Petersen, P.B., Tokmakoff, A., 2010. Source for ultrafast continuum infrared and terahertz radiation. Optics Letters 35 (12), 1962–1964.
Piatkowski, L., Zhang, Z., Backus, E.H.G., *et al.*, 2014. Extreme surface propensity of halide ions in water. Nature Communications 5, 4083. Available at: http://dx.doi.org/10.1038/ncomms5083.
Ramasesha, K., Roberts, S.T., Nicodemus, R.A., *et al.*, 2011. Ultrafast 2D IR anisotropy of water reveals reorientation during hydrogen-bond switching. Journal of Chemical Physics 135, 54509.
Ren, Z., Brinzer, T., Dutta, S., Garrett-Roe, S., 2015. Thiocyanate as a local probe of ultrafast structure and dynamics in imidazolium-based ionic liquids: Water-induced heterogeneity and cation-induced ion pairing. Journal of Physical Chemistry B 119, 4699–4712.
Ren, Z., Ivanova, A.S., Couchot-Vore, D., Garrett-Roe, S., 2014. Ultrafast structure and dynamics in ionic liquids: 2D-IR spectroscopy probes the molecular origin of viscosity. Journal of Physical Chemistry Letters 5, 1541–1546.
Roberts, S.T., Ramasesha, K., Tokmakoff, A., 2009. Structural rearrangements in water viewed through two-dimensional infrared spectroscopy. Accounts of Chemical Research 42 (9), 1239–1249.
Rock, W., Li, Y., Pagano, P., Cheatum, C.M., 2013. 2D IR spectroscopy using four-wave mixing, pulse shaping, and IR upconversion: A quantitative comparison. Journal of Physical Chemistry A 117, 6073–6083.
Rosenfeld, D.E., Gengeliczki, Z., Smith, B.J., *et al.*, 2011. Structural dynamics of a catalytic monolayer probed by ultrafast 2D IR vibrational echoes. Science 334 (6056), 634–639.
Rosenfeld, D.E., Nishida, J., Yan, C., *et al.*, 2012. Dynamics of functionalized surface molecular monolayers studied with ultrafast infrared vibrational spectroscopy. Journal of Physical Chemistry C 116, 23428–23440.
Rosenfeld, D.E., Nishida, J., Yan, C., *et al.*, 2013. Structural dynamics at monolayer – liquid interfaces probed by 2D IR spectroscopy. Journal of Physical Chemistry C 117, 1409–1420.
Roy, S., Pshenichnikov, M.S., Jansen, T.L.C., 2011. Analysis of 2D CS spectra for systems with non-gaussian dynamics. Journal of Physical Chemistry B 115, 5431–5440.
Rubtsova, N.I., Qasim, L.N., Kurnosov, A.A., *et al.*, 2015. Ballistic energy transport in oligomers. Accounts of Chemical Research 48, 2547–2555.
Rubtsova, N.I., Rubtsov, I.V., 2013. Ballistic energy transport via perfluoroalkane linkers. Chemical Physics 422, 16–21.
Rubtsova, N.I., Rubtsov, I.V., 2015. Vibrational energy transport in molecules studied by two-dimensional infrared spectroscopy. Annual Review of Physical Chemistry 66, 717–738.
Sanda, F., Mukamel, S., 2006. Stochastic simulation of chemical exchange in two dimensional infrared spectroscopy. Journal of Chemical Physics 125, 14507.
Schmidt, J.R., Roberts, S.T., Loparo, J.J., *et al.*, 2007. Are water simulation models consistent with steady-state and ultrafast vibrational spectroscopy experiments? Chemical Physics 341, 143–157.
Selig, O., Siffels, R., Rezus, Y.L.A., 2015. Ultrasensitive ultrafast vibrational spectroscopy employing the near field of gold nanoantennas. Physical Review Letters 114, 233004.
Serrano, A.L., Ghosh, A., Ostrander, J.S., Zanni, M.T., 2015. Wide-field FTIR microscopy using mid-IR pulse shaping. Optics Express 23 (14), 17815.
Shim, S.-H., Gupta, R., Ling, Y.L., *et al.*, 2009. Two-dimensional IR spectroscopy and isotope labeling defines the pathway of amyloid formation with residue-specific resolution. Proceedings of the National Academy of Sciences 106 (16), 6614–6619. Available at:.http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2672516&tool=pmcentrez&rendertype=abstract.
Shim, S., Zanni, M.T., 2009. How to turn your pump–probe instrument into a multidimensional spectrometer: 2D IR and Vis spectroscopies via pulse shaping. Physical Chemistry Chemical Physics 11, 748–761.
Singh, P.C., Inoue, K., Nihonyanagi, S., *et al.*, 2016. Femtosecond hydrogen bond dynamics of bulk-like and bound water at positively and negatively charged lipid interfaces revealed by 2D HD-VSFG spectroscopy. Angewandte Chemie International Edition 55, 10621–10625.
Singh, P.C., Nihonyanagi, S., Yamaguchi, S., Tahara, T., 2012. Ultrafast vibrational dynamics of water at a charged interface revealed by two-dimensional heterodyne-detected vibrational sum frequency generation. Journal of Chemical Physics 137, 94706.
Sueur, A.L. Le, Horness, R.E., Thielges, M.C., 2015. Applications of two-dimensional infrared spectroscopy. Analyst 140, 4336–4349.
Thamer, M., De Marco, L., Ramasesha, K., *et al.*, 2015. Ultrafast 2D IR spectroscopy of the excess proton in liquid water. Science 350 (6256), 78–82.

Tracy, K.M., Barich, M.V., Carver, C.L., *et al*., 2016. High-throughput two-dimensional infrared (2D IR) spectroscopy achieved by interfacing microfluidic technology with a high repetition rate 2D IR spectrometer. Journal of Physical Chemistry Letters 7, 4865–4870.
Wang, J., Clark, M.L., Li, Y., *et al*., 2015. Short-range catalyst – surface interactions revealed by heterodyne two-dimensional sum frequency generation spectroscopy. Journal of Physical Chemistry Letters 6, 4204–4209.
Wang, L., Middleton, C.T., Singh, S., *et al*., 2011. 2DIR spectroscopy of human amylin fibrils reflects stable β-sheet structure. Journal of the American Chemical Society 133 (40), 16062–16071.
Weston, M., Britton, A.J., Shea, J.N.O., 2011. Charge transfer dynamics of model charge transfer centers of a multicenter water splitting dye complex on rutile TiO_2 (110). Journal of Chemical Physics 134, 54705.
Woutersen, S., Hamm, P., 2002. Nonlinear two-dimensional vibrational spectroscopy of peptides. Journal of Physics: Condensed Matter 14, R1035–R1062.
Woys, A.M., Lin, Y.S., Reddy, A.S., *et al*., 2010. 2D IR line shapes probe ovispirin peptide conformation and depth in lipid bilayers. Journal of the American Chemical Society 132 (8), 2832–2838.
Wright, J.C., 2011. Multiresonant coherent multidimensional spectroscopy. Annual Review of Physical Chemistry 62, 209–230.
Xiong, W., Laaser, J.E., Mehlenbacher, R.D., Zanni, M.T., 2011. Adding a dimension to the infrared spectra of interfaces using heterodyne detected 2D sum-frequency generation (HD 2D SFG) spectroscopy. Proceedings of the National Academy of Sciences 108 (52), 20902–20907.
Xiong, W., Laaser, J.E., Paoprasert, P., *et al*., 2009. Transient 2D IR spectroscopy of charge injection in dye-sensitized nanocrystalline thin films. Journal of the American Chemical Society 131, 18040–18041.
Yan, C., Yuan, R., Nishida, J., Fayer, M.D., 2015. Structural influences on the fast Dynamics of alkylsiloxane monolayers on SiO_2 surfaces measured with 2D IR spectroscopy. Journal of Physical Chemistry C 119, 16811–16823.
Yan, C., Yuan, R., Pfalzgraff, W.C., *et al*., 2016. Unraveling the dynamics and structure of functionalized self-assembled monolayers on gold using 2D IR spectroscopy and MD simulations. Proceedings of the National Academy of Sciences 113 (18), 4929–4934.
Zhang, Z., Piatkowski, L., Bakker, H.J., Bonn, M., 2011a. Communication: Interfacial water structure revealed by ultrafast two-dimensional surface vibrational spectroscopy. Journal of Chemical Physics 135, 21101.
Zhang, Z., Piatkowski, L., Bakker, H.J., Bonn, M., 2011b. Ultrafast vibrational energy transfer at the water/air interface revealed by two-dimensional surface vibrational spectroscopy. Nature Chemistry 3 (11), 888–893. Available at: http://dx.doi.org/10.1038/nchem.1158.
Zheng, J., Kwak, K., Asbury, J., *et al*., 2005. Ultrafast dynamics of solute–solvent complexation observed at thermal equilibrium in real time. Science 309, 1338–1343.
Zheng, J., Kwak, K., Fayer, M.D., 2007. Ultrafast 2D IR vibrational echo spectroscopy. Accounts of Chemical Research 40 (1), 75–83.
Zhuang, W., Hayashi, T., Mukamel, S., 2009. Coherent multidimensional vibrational spectroscopy of biomolecules: Concepts, simulations, and challenges. Angewandte Chemie International Edition 48, 3750–3781.

Two-Dimensional Electronic Spectroscopy

Yin Song, University of Michigan, Ann Arbor, MI, United States
Xiaoqin Li, The University of Texas at Austin, Austin, TX, United States
Jennifer P Ogilvie, University of Michigan, Ann Arbor, MI, United States

Introduction

At present, the state of the art of optical spectroscopy of condensed phase systems aims to perform multidimensional spectral characterization with ultrafast time resolution. Since the first experimental demonstrations, multidimensional visible and infrared (IR) spectroscopy have been applied to address many fundamental questions in condensed phase dynamics. Several excellent textbooks (Hamm and Zanni, 2011; Cho, 2009) and reviews (Jonas, 2003; Ogilvie and Kubarych, 2009; Cho, 2008) discuss the principles and applications of multidimensional spectroscopy. Here we focus on studies in the visible and near-IR. In this frequency regime, 2D spectroscopy is commonly referred to as "2D electronic spectroscopy (2DES)," to be distinguished from its infrared (2DIR) and ultraviolet (2DUV) counterparts. We discuss experimental considerations and common implementations of 2DES, and summarize recent applications to a broad range of condensed phase systems.

2DES is an extension of one-dimensional (1D) pump–probe spectroscopy that overcomes some of its limitations by resolving the third order spectroscopic signal as a function of excitation frequency. This additional frequency dimension can separate homogeneous and inhomogeneous broadening, reveal electronic coupling and expose elusive states that can be hidden in systems with rich excited-state structure and dynamics. In pump–probe spectroscopy, the electric field of the pump pulse interacts (via the dipole operator) with the molecules twice to promote the molecules from the ground state to the excited state. After a variable waiting time (T), the electric field of a probe pulse interacts with the system of interest to generate the third-order polarization, which radiates the signal field. The probe pulse also acts as a reference pulse, referred to as a 'local oscillator' to interfere with the signal field, enabling heterodyne-detection (here the term heterodyne-detection refers to the detection of the interference between the signal field with a "local oscillator" field with the same spectral content. This process enables signal amplification and detection of the complex signal field (spectrally-resolved phase and amplitude of the signal. The pump–probe signal ($S(T,\omega)$) is often presented as a function of waiting time and detection frequency/wavelength to reveal the time evolution of the excited-state populations. Since the pump pulse can only be either a spectrally narrow (temporally long), or a spectrally broad (temporally short) pulse, there is a trade-off between time and frequency resolution with respect to the excitation axis. Thus, pump–probe spectroscopy is limited in its ability to uncover dynamic correlations between direct photoexcited states and probing states in complex systems with spectrally-overlapping transitions and dynamics occurring on an ultrafast time scale.

Fourier transform 2D spectroscopy differs from pump–probe spectroscopy in that it utilizes two ultrashort pulses with a variable time delay (coherence time, τ) to excite the sample (see **Fig. 1(a)**). The first field-matter interaction creates a coherence (a superposition of the ground and excited states) while the second interaction promotes the molecules to a population or another coherence. After a waiting time (T), the third-order polarization is generated in the sample via the third field-matter interaction and the radiated signal is heterodyne-detected with a fourth 'local oscillator' pulse. For each waiting time, the coherence time is scanned from $-\tau$ to $+\tau$ to obtain a 2D map ($S(\tau, T, \omega_3)$), where ω_3 is the detection frequency. Fourier transformation of the signal along the coherence time axis gives the signal in the frequency domain $S(\omega_1, T, \omega_3)$, (where ω_1 is the excitation axis). This 2D dynamic map directly reveals the correlation of the photoexcited and probing states as a function of waiting time T. As the excitation information is recorded by using the time-domain technique, the excitation frequency resolution is determined by the length of the coherence time scan rather than the excitation bandwidth. In other words, 2DES circumvents the time-frequency resolution trade-off in pump–probe spectroscopy.

Challenges

A key challenge in Fourier-transform 2DES is creating and delivering the appropriate pulse sequence with variable, accurate and phase-stable relative time delays. Time delays must be recorded with high accuracy if they are to yield accurate Fourier-transformed frequencies. Phase stability, which to first order is equivalent to a timing jitter between the pulses, should be as high as possible to enable clean separation of the complex signal components with a high signal-to-noise ratio. Typically one requires an interferometric precision of $\sim\lambda/100$, corresponding to timing errors of $\sim$0.017 fs at 500 nm. As a consequence, the challenges of implementing 2DES increase for shorter wavelengths, where small mechanical vibrations or air current fluctuations lead to optical pathlength variations. Another challenge is that the 2D signal is relatively weak and must be distinguished from the incident pulse as well as from pump–probe signals, free-induction decays and transient-grating signals. Isolating the 2D signal from the background and noise (e.g. the scattering) can be achieved by using phase-matching and/or phase-cycling. In addition, to maximize the information available from a 2D measurement, both "rephasing" and "nonrephasing" complex signal fields must be recorded. To

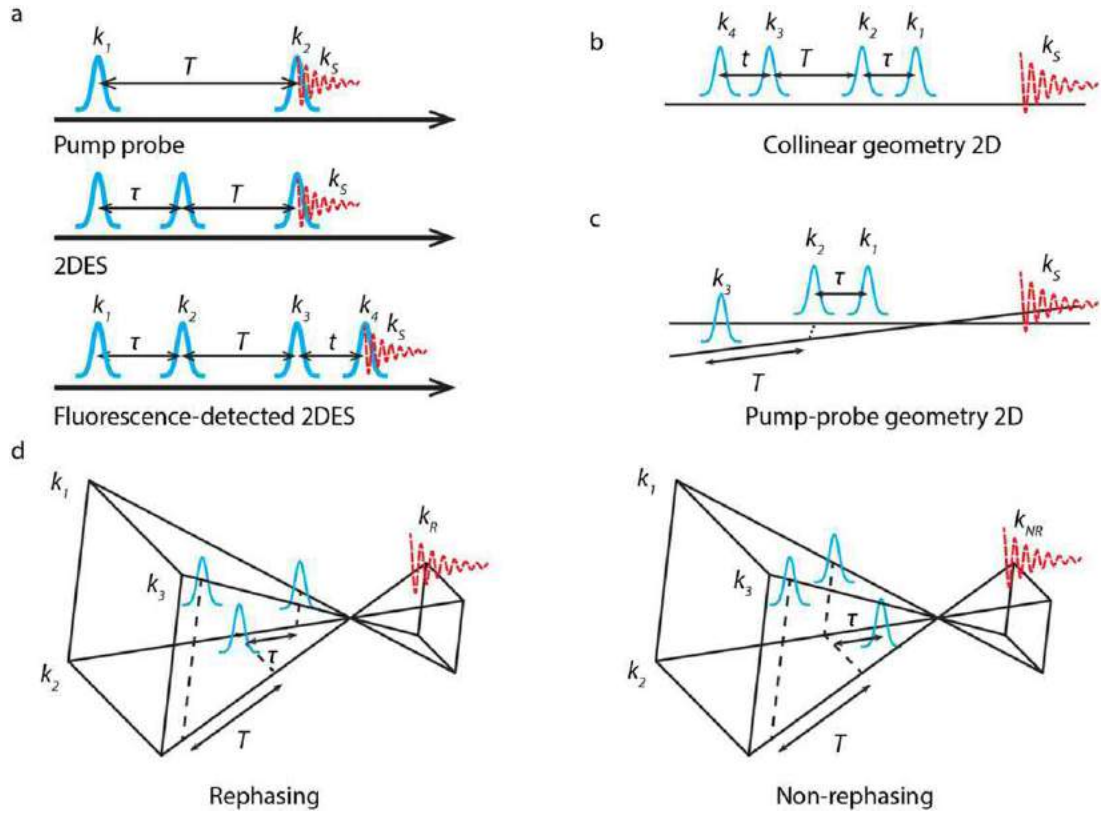

Fig. 1 (a) Pulse sequence in pump–probe, 2DES and fluorescence-detected 2DES; (b)–(d) cartoon illustrations of collinear geometry 2DES (fluorescence-detected), pump–probe geometry 2DES, BOXCARS rephasing and non-rephasing 2DES.

resolve the complex signal fields, the majority of 2DES setups employ spectral interferometry, in which interference between the signal and a reference pulse is measured in the frequency domain. A number of different experimental implementations of 2D spectroscopy have met the many challenges of 2D spectroscopy in different ways (recently reviewed by Fuller and Ogilvie (2015)) as discussed below.

Experimental Implementations

The first multidimensional optical spectroscopy measurements were pioneered independently by Joffre and Jonas. In 1996 the Joffre group, who developed the spectral interferometry method, used 2D spectroscopy to record the second-order response of a nonlinear crystal. The first third-order 2DES setup was reported by the Jonas group in 1998. In their implementation, two excitation pulses, the probe pulse and the signal were arranged in the BOXCARS geometry and the signal was emitted in a background-free direction. The coherence time was scanned using a translation stage that was calibrated by spectral interferometer, with the time zero determined by finding the symmetry-point of a pump–probe signal. The complex signal was recorded via spectral interferometry. A tracer pulse with pre-determined phase (by frequency-resolved optical gating (FROG)) was sent along the signal path and interferometry between the tracer and reference pulses was used to determine the signal phase. The phase stability of the interferometer-based setup was $\sim\lambda/27$ at 800 nm for 20 min (i.e. 0.1 fs rms drift of time).

Following these pioneering works, a number of approaches to 2DES were developed, aiming to improve phase stability, signal-to-noise ratio and ease of implementation. Three geometries, shown in **Fig. 1(b)–(d)**, are now commonly used to perform 2DES: the collinear geometry, the pump–probe geometry and the BOXCARS geometry. Below we describe commonly-used implementations of 2DES and how they overcome the various implementation challenges. **Fig. 2** shows the experimental setup for several typical 2DES implementations. The pros and cons of various approaches to 2DES are given in **Table 1**.

The BOXCARS Geometry

The BOXCARS geometry was the first to be implemented by the 2D community and remains the most common geometry. Its key advantage is that the signal emits in a background-free direction (spatially separated from the three pulses that generate the 2D signal). In 2DES, two different phase-matched signals are typically recorded and combined to obtain the "absorptive" 2DES

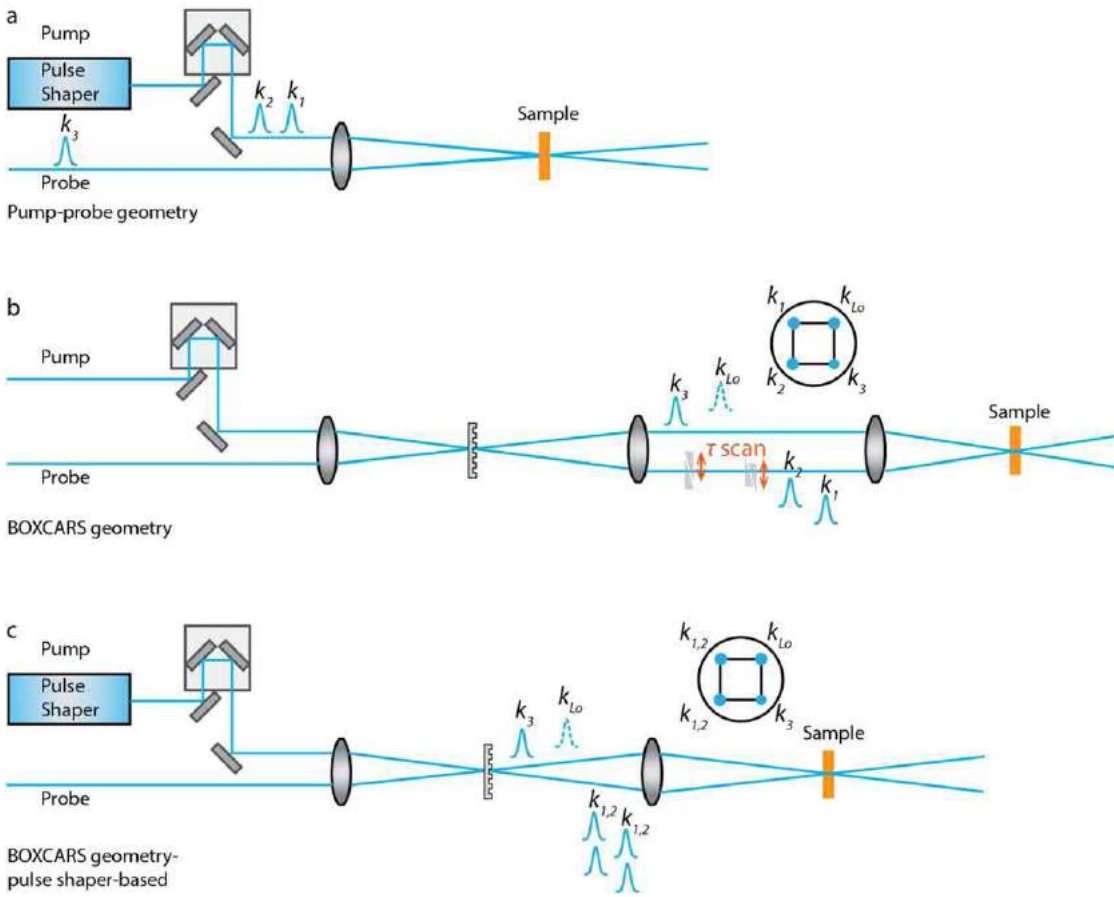

Fig. 2 Experimental implementations of several different 2DES setups. (a) Pulse-shaper-based pump–probe geometry 2DES; (b) BOXCARS geometry 2DES with refractive delays (τ), employing a diffractive optic (D.O.) beam splitter; and (c) pulse-shaper based BOXCARS geometry 2DES, employing a D.O. beam splitter.

spectrum, free from broadening refractive contributions. These are the "rephasing" and "nonrephasing" signals, emitted in the $\mathbf{k}_2 - \mathbf{k}_1 + \mathbf{k}_3$ and $\mathbf{k}_1 - \mathbf{k}_2 + \mathbf{k}_3$ directions, respectively. A related signal, with the phase-matching direction $\mathbf{k}_1 + \mathbf{k}_2 - \mathbf{k}_3$, yields information about double-quantum coherences. For mechanical delay lines, the rephasing and nonrephasing 2D signals are collected by scanning coherence time τ over positive and negative delays, respectively. As all four laser beams typically experience different optical paths in the BOXCARS geometry, the phase of the complex 2D signal is undetermined. The method of "phasing" the 2D spectrum will be discussed in more detail below.

Several methods have been used to generate the pulse pairs for 2DES in the BOXCARS geometry. The first 2DES setup used conventional beam splitters. Beam splitters pose three main problems for broadband pulses: (1) the beams may not be perfectly phase-matched at the focus (sample position); (2) the signal propagation direction may be slightly different from that of the local oscillator, reducing signal and resulting in spectral distortions; and (3) pulse-front tilt reduces the time resolution of the experiment. By constructing robust interferometers with carefully mounted optics and balancing the glass thickness in the pulse pairs, the phase differences and fluctuations can be minimized. To avoid the problem of the pulse-front tilt and imperfect phase matching, diffractive optic (D.O.) beam splitters were pioneered by the Miller and Nelson groups to generate pulse pairs, initially for transient-grating spectroscopy. Depending on the type of the light source and the applications, some groups have used a single 2D D.O. or 2 perpendicularly-mounted 1D D.O. to split a single beam into four beams at the corners of a square; other groups use two light sources as pump and probe beam, crossing them at a 1D D.O. to split each beam into two. The use of a single laser source gives better phase-matching while the use of two light sources with a 1D D.O. enables two-color 2D measurements. The use of D.O.s provides easy alignment of the signal and local oscillator for heterodyne detection. The D.O. based setup has the limitation of the spectral bandwidth since the diffracted beams are overlapped when the bandwidth reaches an octave of the wavelength.

Another method to generate the pulse pairs is to use a spatial filter with four holes at corners of a square. This method was employed first by the Nelson group and later by the Piefier group to generate four pulses from one original beam. Nelson and coworkers demonstrated that there is considerable power loss of the beam by using a spatial filter and that scattering from the edges of the holes can increase background signals.

Table 1 Summary of various implementations of 2DES, pros and cons and bandwidth limitations.

Method	*Pros*	*Cons*	*Bandwidth limitations*
Standard interferometer	Broadband and heterodyne detection	Motorized delays can cause time zero uncertainties and timing errors Requires either active stabilization to achieve linearly spaced time delays or high-precision time-delay measurements Rephasing and nonrephasing spectra are not acquired simultaneously Requires phasing to obtain absorptive spectra	Limited only by optical components (beam splitters, mirrors)
Diffractive optics	High passive phase stability, ease of alignment and heterodyne detection	Requires phasing to obtain absorptive spectra Rephasing and nonrephasing spectra are not acquired simultaneously	Limited by overlapping octaves and optical components (grating efficiency, mirrors, etc.)
Pulse shaping	Phase cycling to remove scatter and separate signals Automatic absorptive spectra in pump–probe geometry No uncertainty in interpulse pump delay Phase locking allows for spectra to be collected in fewer time steps.	Polarization control may be difficult, depending on pulse shaper and geometry May require phasing to obtain absorptive spectra, depending on geometry Pulse shapers can be lossy	Likely limited by pulse shaper Often limited by coatings for spatial-light-modulator shapers
Fluorescence-detected two-dimensional Fourier transform via phase modulation	Insensitive to mechanical/phase instabilities 1/f noise suppressed by lock-in detection No need for phasing data Single element detector, facilitating use with high repetition rate laser sources.	Requires the construction of reference signals Requires an additional motorized time delay to be scanned	Likely limited by acousto-optic modulators
Single shot	Reduced acquisition time, insensitive to laser noise	Requires separate recording of rephasing and nonrephasing signals Requires phasing to obtain absorptive spectra Laser spatial mode quality requirements are more stringent Requires a 2D camera to record the spatial dimension	Limited only by optical components (beam splitters, mirrors)
All reflective approaches	Broadband	Requires phasing to obtain absorptive spectra, depending on geometry May have reduced phase stability, depending on exact approach used Rephasing and nonrephasing spectra are not acquired simultaneously Motorized delays can cause time zero uncertainties and timing errors	Limited only by optical components (beam splitters, mirrors)
Birefringent interferometer (TWINS)	Collection in a partially rotating frame allows for spectra to be collected in fewer time steps.	Cannot implement phase cycling, or modulation of single pulses in a pulse pair	Limited only by optical components (beam splitters, mirrors, birefringent crystals)

(*Continued*)

Table 1 Continued

Method	*Pros*	*Cons*	*Bandwidth limitations*
	Compensates group delay dispersion introduced by changing wedge thickness over a time scan Permits pump–probe geometry measurements without need for a pulse-shaper	Motorized delays can cause time zero uncertainties and timing errors (reduced in the case of refractive delay lines)	

Source: Fuller, F.D., Ogilvie, J.P., 2015. Annual Reviews of Physical Chemistry 66, 667–690.

The use of D.O.s or spatial filters to generate the four beams for a BOXCARS 2DES experiment promises high passive phase stability since the phases of any beams originating from a single beam are correlated. Thus when common optics are used to direct them to the sample, relative phase errors between pulse pairs can be minimized, producing a passively phase stable signal. To precisely scan the time delay and have the cancellation of the phase fluctuations in the signal, two designs of optical paths have been developed. One design uses a geometry similar to D.O.-based transient grating experiments where all four beams hit common reflective mirrors between the D.O. and the sample. In order to precisely scan the coherence time while maintaining passive-phase stability, refractive delays may been used, either by adjusting the material insertion using a pair of wedges mounted on a translation stage, or by rotating cover slides mounted on a rotational stage. The use of the refractive delays introduces additional chirp during the coherence scan and may pose some problems for broadband laser pulses and longer time scans. For a 10 fs Fourier-transform limited visible pulse centered at 600 nm, the temporal width can be distorted by $\sim$6% when scanning up to 300 fs. The refractive delay works better for the coherence scan where optical dephasing $<$100 fs. As an alternative to refractive delays, other designs using all reflective optics may be used. This approach was pioneered by the Miller group, who showed that a carefully designed optical path in which appropriate pulse pairs hit common optics can minimize phase drifts in the detected signal. They also improved the phase stability with a design that decouples the coherence time scan from the waiting time delays.

As an alternative to the passive phase-stabilization approach, the Cundiff group was the first implement an actively phase-stabilized 2DES experiment. In their approach, the four beams for the BOXCARS geometry are generated by two interferometers, one between the two excitation pulses and another between the probe and reference pulse. A third interferometer is used to actively control the relative phase of the output from the other two interferometers. Changes in the relative time delays of the four pulses are measured by monitoring the interference of a continuous HeNe laser co-propagating through the interferometers. The HeNe measurements provide feedback to actively phase-stabilize the setup by adjusting mirrors mounted on piezoelectric transducers. In this approach, the time delays are scanned by conventional delay stages and actively stabilized at each coherence and population time step, for which the 2D signal may be averaged over the desired number of laser shots. This setup is particularly powerful for enabling the measurement of double quantum coherence signals that require phase-stabilization over the population time delay T, and for experiments that require the scanning of long coherence time delays.

Pulse-shapers

Pulse-shapers can be used to replace one or two delay lines in 2DES experiments. Removing the mechanical delay lines can improve the phase stability and provide the opportunity to control the relative pulse phases, enabling "phase-cycling" for scatter reduction and the isolation of particular signals of interest. The Nelson group has developed a spatial-temporal pulse-shaper in which the four laser pulses, generated either by a D.O. or by a spatial light modulator (SLM), are incident at the different locations of a 2D SLM pulse-shaper, enabling independent control of each pulse delay and phase. They have shown that space-time coupling effects in the setup can be largely corrected. The Ogilvie group has combined the use of an acousto-optic (Dazzler) pulse-shaper and D.O. for 2DES. In their implementation, the pulse-shaper is placed before the D.O. to generate a pair of collinear pump pulses with controlled relative phase. Thus instead of having four laser pulses (Fig. 2), there are six pulses incident on the sample. Two pairs of the pulses are time overlapped and generate two transient grating signals at different waiting times in the direction of the local oscillator. Rephasing and nonrephasing signals are emitted simultaneously in the phase-matching direction and can be separated via phase-cycling. Simultaneous collection of both rephasing and nonrephasing signals facilitates excellent "phasing" of the data. In general, the use of pulse-shapers offers the advantages of ease-of-use and phase-cycling that can isolate signals of interest and suppress background scatter that differs from the desired signal in its phase dependence on the incident pulses. Pulse-shapers that can operate at high update rates also enable detection of the signal in the rotating frame with coarse time steps on a shot-to-shot basis, speeding up the experiments. The disadvantages of pulse-shaper-based 2DES designs stem from the spectral bandwidth (typically $\sim$100–200 nm are achievable), maximum time delays (typically picoseconds) and overall throughput (ranging from $\sim$10–75%), each of which is pulse-shaper and implementation-dependent.

"Phasing" the 2DES signal

In principle, if all the pulses generated from a single light source experience the same optical elements and the reference beam follows the signal path, the absolute phase from the different pulses cancels in the signal. However, in the BOXCARS geometry, the

absorptive and dispersive part of the 2D signal are mixed even with the use of a D.O., requiring that a polynomial phase be added to separate the real and imaginary part of the complex signal. The phase distortion of the 2D signal may be due to the imperfect optical surfaces, the zero time offset between pump pulses and unbalanced material dispersion in the beam paths. A few methods have been proposed to phase the 2D spectrum.

The Jonas group first proposed to determine the absolute signal phase by utilizing a tracer pulse to follow the optical path of the signal. Spectral interferometry could then be used to determine the difference between the relative phases of the tracer and the reference, and the reference and signal. With measurement of the spectral phase of the tracer as determined by frequency resolved optical gating (FROG), the absorptive and refractive parts of the 2D spectrum could be separated. Alternatively, the Cundiff and Hamm groups have employed spatial interference for phasing the signal.

The projection-slice theorem is the most commonly used approach for phasing 2D spectra. It has been shown by the Jonas group that the integration of the 2D signal over the excitation axis is equivalent to the pump–probe signal. Thus, the signal phase can be determined by a fitting procedure that finds the optimum variable spectral phase for which the integrated 2D signal matches the spectrally-resolved pump–probe signal for the same waiting time delay. Two potential issues may occur when applying this theorem. First, the pump–probe signal obtained in the BOXCARS geometry is weaker than the 2D signal and can be noisy, in particular for scattering samples. Secondly, the pump–probe signal is often obtained by blocking one pump beam and one probe beam in the BOXCARS geometry. Since the geometrical configuration of the pump–probe and 2DES measurements are different this may result in a difference of the signal amplitude and potentially spectral differences between the integrated 2DES and pump–probe data.

To resolve the issue of the weak pump–probe signal, the Hauer group has employed two D.O.s in the 2DES setup in such a way that both the rephasing and nonrephasing signals can be collected simultaneously. Through balanced detection, an automatically-phased heterodyne-detected transient grating signal (TG) can also be obtained in this configuration. The phased heterodyne-detected TG is much stronger than the pump–probe signal, and can be used in its place to phase the 2D spectrum using the projection-slice theorem.

2DES in the pump–probe geometry

2DES in the pump–probe geometry was first proposed by the Jonas group in 1999 and implemented first in the infrared by Zanni and Tokmakoff and later in the visible by Ogilvie. The pulse sequence is showed in **Fig. 1(c)**, where a collinear pair of pump pulses excite the sample, which is then probed by a third noncollinear probe pulse. Since the wavevectors of the pump pulses are identical, the rephasing and non-rephasing signals are emitted in the same direction, along with two pump–probe signals and the probe beam. An advantage of the pump–probe geometry is that the relevant pulse pairs hit common optics and phase fluctuations cancel in the signal, providing excellent passive phase stability. In addition there is no unknown relative phase between the local oscillator (the transmitted probe) and the signal. This obviates the need for a phasing procedure as the absorptive part of the 2D spectrum is automatically obtained. Compared to the BOXCARS geometry, the phase-matching of broadband pulses is easier to achieve in the pump–probe geometry. As the probe beam follows the signal path and enters the detector, an attenuator may need to be placed before the detector in order to avoid detector saturation. As a result, the signal is also attenuated and the weak signal becomes more susceptible to noise, including scatter. The inability to control the relative amplitude of signal and local oscillator limits the ability to optimize the signal-to-noise ratio. Recently, the Jonas group showed that this problem can be solved with a modified pump–probe geometry in which the sample is placed in a Sagnac-interferometer. This setup enables the attenuation of the local oscillator without affecting the intensity of the probe beam. Thus a stronger 2D signal can be generated without saturating the detector and the signal-to-noise ratio can be optimized.

The coherence time scan in the pump–probe geometry is not as straightforward as the BOXCARS geometry since two excitation pulses are overlapped in space. Several methods have been used to implement the coherence scan. Tokmakoff has used conventional delay stages, while the Zanni and Ogilvie groups have utilized an AOM-based pulse-shapers, either in a 4f geometry or using a transmissive shaper (Dazzler) to generate the pump pulse pairs and scan the coherence time. As discussed previously, the use of pulse-shapers makes it possible to collect the signal in a rotating frame scheme on an (almost) shot-to-shot basis which greatly decreases the experimental time. Phase cycling can be used to reduce scatter and isolate signals of interest such as rephasing and nonrephasing components. The insertion of a pulse-shaper into the pump arm of a frequency-resolved pump–probe setup enables its straightforward conversion into a 2DES experiment.

The Cerullo group developed a method called Translating-Wedge-Based Identical Pulses eNcoding System (TWINS) to generate the time-delayed pulse pairs for 2DES in the pump–probe geometry. The method employs a 45° polarized beam incident upon several birefringent elements, including wedges mounted on a translation stage, to generate two orthogonally-polarized pulses with a variable time delay. After the birefringent materials, a 45° polarizer projects the time-delayed collinear pulse pair to the same polarization. Although refractive delays are used, this setup is capable of broadband operation, with $\sim$10 fs pulses in either UV, visible or mid-IR regimes by a careful selection of the birefringent materials. In this setup, the coherence delay is calibrated using spectral interferometry between two pump pulses in order to achieve an accurate excitation frequency axis. A phasing procedure owing to the uncertainty of the time zero offset is required to obtain the absorptive 2D spectrum.

The Fully Collinear Geometry

The collinear geometry 2D setup was pioneered by the Warren group in 2003, using an acousto-optic pulse-shaper to generate a sequence of three pulses for the experiment. In this geometry the 2D signal is spatially overlapped with the incident pulse sequence, as well as the linear and other higher order signals. In this case a combination of fluorescence detection to enable

spectral filtering of the incident pulses, and phase-cycling can be used to extract the desired signal. The collinear geometry avoids possible spectral distortions that can arise from imperfect phase-matching in other 2DES geometries. This is particularly useful for studies of samples such as molecular clusters or single molecules, the size of which is smaller than the optical wavelength.

The Marcus group has also pioneered fluorescence-detected 2DES in the collinear geometry. Instead of using a pulse-shaper, they employ a phase-modulation approach, using two Mach-Zender interferometers to generate two independently phase-modulated collinear pulse pairs. Each beam is tagged with a unique frequency via an acousto-optic modulator (AOM), such that each pulse pair is modulated by the difference frequency of the AOMs. A third delay line scans the waiting time between the excitation and probing pulse pairs. The nonlinear signal of interest can be selected via lock-in detection using an appropriately constructed reference frequency. This reference frequency can be generated by combining the pulse-pair interference signals from each Mach-Zender interferometer, detected in separate monochromators. The double phase modulation can significantly suppress noise and remove the effect of time jitter caused by timing errors in the mechanical delay lines.

We note that the use of fluorescence signal detection requires that the final pulse leave the system in an excited state population. Compared to the standard 2DES three pulse sequence, an additional field-matter interaction is required. This can be realized by adding a fourth pulse, as is done by the Marcus group, or by having one of the three pulses interact with the sample twice. In the Warren implementation, the second pulse plays this role, yielding a 2D spectrum at zero waiting time. The additional field-matter interaction means that there are differences between the quantum pathways accessed in fluorescence-detected and standard 2DES experiments, making them complementary approaches. As an alternative to fluorescence detection, the fully collinear phase-modulation approach can be combined with other measurements that detect excited state populations. For example, both the Marcus and Cundiff groups have used this approach in combination with photocurrent signal detection. Aeschlimann and coworkers have used photoemission electron microscopy to read-out the excited state population, achieving nanoscopic spatial resolution.

Noise Suppression

2DES signals are often contaminated by unwanted spurious signals as well as scattering. In the BOXCARS geometry, a linear response contribution from the local oscillator, as well as a third-order signal will be detected along with the desired 2D signal. These unwanted signals can be removed via Fourier transformation along the detection frequency axis. However, scattering from excitation and probe pulses can also emit along the signal detection direction, and can significantly degrade the signal quality. An early solution was to use a shutter/chopper in the probe beam path to remove the scattering from the excitation pulses. Recently, the Zigmantas group implemented a double modulation scheme to suppress the noise and signals from slowly accumulating long-lived species.

Pulse-shapers also offer easy noise suppression methods. If scattering or other unwanted signals have different phase dependence than the 2D signal, phase cycling can be used to effectively isolate the desired 2D response. This method is particularly useful in the pump–probe and fully collinear geometries where the signal is overlapped with large background optical responses. The particular experimental implementation and desired signal dictate the appropriate phase-cycling schemes.

The Cerullo group has employed a noise suppression approach wherein a loudspeaker is attached to a probe mirror mount. The vibration induced by the speaker modulates the 2D signals on a shot-to-shot basis. The scattering can be suppressed by carefully tuning the amplitude and phase of the vibration. A similar approach that uses a wobbling Brewster window or photoelastic modulator has been used by the Hamm group in 2D infrared experiments.

Single-Shot Measurements

The Engel group developed a single-shot 2DES method called gradient-rapid-assisted-pulse-echo spectroscopy (GRAPES). In analogy to approaches used in single-shot pulse characterization methods, they use spatial encoding of a time delay. This is achieved by focusing the two excitations pulses using a cylindrical lens. As the two excitation beams cross at an angle near the focus, the wavefront tilt produces a spatial encoding of the coherence time delay for a fixed waiting time. The third-order signal is then imaged onto a two-dimensional CCD camera in which the vertical axis resolves the signal at different coherence time delays. In this approach rephasing and nonrephasing signals are collected independently and the focusing geometry defines the range of coherence times that can be scanned, which is typically ~ 300 fs. The transient grating signal can be extracted from either the rephasing or nonrephasing spectrum at $\tau = 0$. Since the transient grating signal is equivalent to 2D spectrum integrated over the excitation axis, the rephasing and nonrephasing spectrum can be phased independently by using the projection slice theorem. The single-shot approach has the advantage of being robust against laser fluctuations that can degrade the signal-to-noise ratio in 2DES experiments. The rapid acquisition time also enables higher order multidimensional spectroscopies as recently demonstrated by the Harel group.

Applications of 2D Electronic Spectroscopy

Early 2D Studies

The development of 2D spectroscopy was inspired by Ernst, who suggested in 1976 that the technique of multidimensional NMR could be extended to other frequency regimes. With the development of ultrafast lasers, this idea was realized by Joffre in 1996. In the first implementation of 2D spectroscopy, Joffre and coworkers characterized the second-order response of a nonlinear crystal (potassium dihydrogen phosphate) (KDP). The signal $\chi^{(2)}$ was presented as a function of both the difference and sum frequencies

between the two excitation frequencies. It was suggested that such a 2D map could provide relevant information about the coupling between excited states.

The first third-order Fourier transform 2D spectroscopy was implemented by the Jonas group in 1998. A laser dye–IR144 was used to demonstrate the feasibility and capability of the experiment. The solvation dynamics of IR144 was investigated later by the same group. They found that the lineshape of 2D absorptive spectra evolves from a symmetrical diagonally elongated shape at 0 fs to a round square-like shape at 100 fs. The lineshape evolution reflects the fast lost of the correlation between excitation and detection frequency, induced by vibrational relaxation and solvation.

Since these early studies, 2DES is being used in an increasing number of applications. These applications exploit the strengths of the method that are particularly suited to the system of interest. Below we detail several examples of 2DES applications to a wide range of systems.

Photosynthesis

In photosynthetic systems, arrays of multi-pigment antenna complexes harvest sunlight and transfer the excitation energy to reaction centers where the energy is converted to a stable charge-separated state that fuels the slower biochemical steps of photosynthesis. The pigment-protein antennae and reaction center complexes make ideal targets for 2DES studies for a number of reasons. Designed to absorb light energy over a wide range of wavelengths, photosynthetic systems use a relatively small number of types of pigments, held in place by the surrounding protein. The protein environment tunes their electronic resonances and fixes their relative separations and orientations to define the energy landscape that guides excitation energy transfer and charge separation pathways. As a result, the linear absorption spectra of photosynthetic complexes are highly inhomogeneously broadened, and the electronic coupling between pigments varies from weak, where excitations are localized on individual pigments, to strong, where excitations are shared among pigments and an "excitonic" picture is appropriate. 2DES provides unique information about the electronic structure of photosynthetic systems that is buried in the linear absorption spectrum. 2DES as a function of waiting time details the photosynthetic energy transfer and charge separation pathways, enabling tests of transfer mechanisms. Below we give a brief overview of a subset of the work in this growing field and refer the reader to reviews (Ogilvie and Kubarych, 2009; Ginsberg *et al.*, 2009; Lewis and Ogilvie, 2012; Mukamel *et al.*, 2009) for a more comprehensive discussion.

In 2005 the first 2DES study of a photosynthetic complex was made by Fleming and coworkers, who examined the Fenna–Matthews–Olson (FMO) complex from green sulfur bacteria. Its simple and well-characterized structure has made the FMO complex an important model system to investigate photosynthetic energy transfer. The FMO complex is a trimer of identical subunits, consisting of 7 bacteriochlorophylls (BChl) (although an 8th BChl is present in some preparations) that serves as a bridge structure to transfer energy from the light-harvesting antennae to the reaction center. Within each subunit, the intermolecular electronic coupling in FMO leads to seven excitonic states delocalized over a subset of BChl molecules as depicted in **Fig. 3**. Fleming and coworkers mapped the energy transfer pathways in FMO by examining the waiting-time dependence of the cross peak amplitudes that indicate population transfer between the different excitonic states. Although consistent with overall downhill energy flow, their results suggested that there are two dominant pathways in FMO. Recently, Zigmantas and coworkers

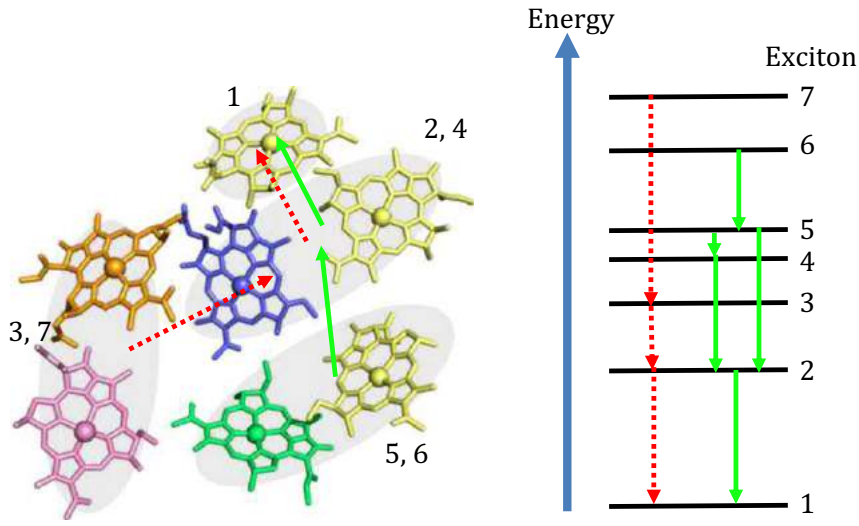

Fig. 3 The seven bacteriochlorophyll pigments of FMO (left), and the excitonic level structure (right). Gray shading indicates the delocalized nature of the excitons. Energy transport pathways identified by Fleming and coworkers (Brixner *et al.*, 2005) are depicted in red (dashed) and green (solid). Adapted from Ogilvie, J.P., Kubarych, K.J., 2009. Advances in Atomic, Molecular, and Optical Physics 57, 249–321.

revisited this system by using 2DES with full polarization-controlled beams. Controlling the polarizations of excitation and probing beams make it possible to suppress diagonal (or cross) peaks and enhance visibility of the cross (or diagonal) ones. This technique was first utilized by Hochstrasser in the infrared and has been employed by several groups for 2DES studies. Through global analysis of the polarization-controlled 2D data Zigmantas and coworkers were able to resolve the 8th exciton state and further unravel the complex energy-transfer pathways.

Another example in which 2DES was used to map energy-transfer pathways was a study by the Scholes group of the light-harvesting II (LH2) complex from purple bacteria. LH2 contains two types of chromophores: carotenoids (Car) and BChl. The electronic structure of Car is often described by a three-level model – a ground state, a dark low-lying state (S_1, $2A_g^-$) and a bright high-lying state (S_2, $1B_u^+$). Two energy-transfer pathways from Car to BChl have been suggested: a direct energy transfer from S_2 state to the BCh1 Q_x state and an indirect energy-transfer pathway via S_1 state to the BCh1 Q_y. The theoretically estimated overall energy-transfer efficiency from the two pathways was in poor agreement with experimental measurements. It had been proposed that a dark/weakly allowed state (X) might exist between S_2 and S_1 which was missing in the energy-transfer calculation and unresolved in 1D spectra. The Scholes group showed that this state could be clearly resolved by 2DES. To study the role of the dark state in energy transfer, they performed a global-target analysis, finding that the 2D evolution associated spectra (EAS) exhibited multiple cross peaks corresponding to the X, S_1, S_2 and Q_x states. Their analysis revealed a clear map of energy transfer via the dark state, resolving the discrepancy between theoretical calculations and experiments.

The Ogilvie group was the first to study photosynthetic reaction centers with 2DES. They examined the photosystem II reaction center (PSII RC), which is at the heart of oxygenic photosynthesis, taking excitation energy from neighboring antennae complexes and forming a charge separated state on ultrafast timescales. The PSII RC contains six chlorophylls (Chl) and two pheophytins (Pheo) arranged in two symmetric branches, with charge separation occurring on one side. The spectral overlap of all of the RC pigments, and the similar timescales of energy transfer and charge separation in this system have made it particularly challenging to understand both the electronic structure and charge separation mechanism of the PSII RC. To extract the heterogeneous kinetics of the system they fit the 2D dataset at each frequency-frequency point to a function composed of four exponential-decay terms. From this they constructed 2D decay-associated spectra to expose connections between the initially excited states and subsequent intermediates throughout the processes of rapid energy transfer to the charge separated state. In collaboration with Abramavicius and Mukamel, the Ogilvie group has used 2DES to test excitonic and charge separation models of the PSII RC (see **Fig. 4**). More recently they have performed 2DES experiments on the PSII "core" complex, revealing the spectral signatures of energy equilibration and transfer between neighboring antennae and the RC.

Coherence

Excitation with broadband pulses can readily excite superpositions of electronic, vibrational and mixed electronic-vibrational (vibronic) states, leading to coherent oscillations observable in time-resolved spectroscopies. Of particular interest to the photosynthesis and artificial light-harvesting communities have been observations of coherent processes on timescales concurrent with energy transfer and charge separation, leading to considerable speculation about the possible functional importance of the observations. The first report of coherent dynamics in a photosynthetic system was made by Vos, Martin and coworkers in the early 1990s in their pioneering pump–probe studies of the bacterial reaction center. Their work stimulated considerable experimental and theoretical effort. The interest in coherence in photosynthetic function was revived by the Fleming group in 2007 when they reported long-lived coherent amplitude oscillations in 2DES studies of the FMO complex at 77 K. These oscillations appeared at cross-peaks in the 2DES data as a function of waiting time. They originally attributed the observations to electronic coherence between exciton states, and proposed that the long-lived nature of the coherence could have important functional implications for efficient and coherent energy transfer. Similar coherent dynamics have now been observed by 2DES in a number of other light-harvesting antennae and reaction centers over a wide range of temperatures.

Considerable effort by the 2DES community has been devoted to understanding the physical origin of coherent dynamics in 2D spectra. Experiments by the Kauffman and Ogilvie groups on simple laser dyes showed that purely vibrational coherence can generate coherent oscillations in 2DES spectra. When the electronic structure of a system is well understood, the ways in which coherence between different states can be generated by a 2D pulse sequence can be readily enumerated and mapped to the appropriate (ω_1, ω_3) position on a 2D spectrum. Turner and coworkers performed this analysis for simple systems as illustrated in **Fig. 5**. They find that purely electronic coherence produces oscillations in the cross peaks of rephasing spectra and the diagonal peaks of nonrephasing spectra, while purely vibrational coherence produces a distinctly different pattern of oscillating signals. Egorova has also performed a similar study, as have Butkus and coworkers. The latter have Fourier transformed the 2D spectra (or rephasing or nonrephasing spectra) along the waiting time T, extracting Fourier-transformed maps (often called coherence amplitude maps or beating maps). Such maps distinguish electronic and vibrational coherence and separate vibrational coherence from ground- and excited- electronic states at certain peak positions. Inspired by this work, Seibt and Pullerits Fourier transformed the complex 2D signal and found that the coherence amplitude maps further separates into positive and negative frequencies, which can help distinguish signals from different pathways. In addition to these data analysis methods, the Ogilvie group developed an experimental approach termed 'two-color rapid acquisition coherence spectroscopy' (T-RACS), to separate vibrational coherence from ground- and excited-electronic states, demonstrating that this could be done for some vibrational modes of chlorophyll in solution. This experiment is a variant of 2D spectroscopy and has the advantage to greatly reduce the experimental time.

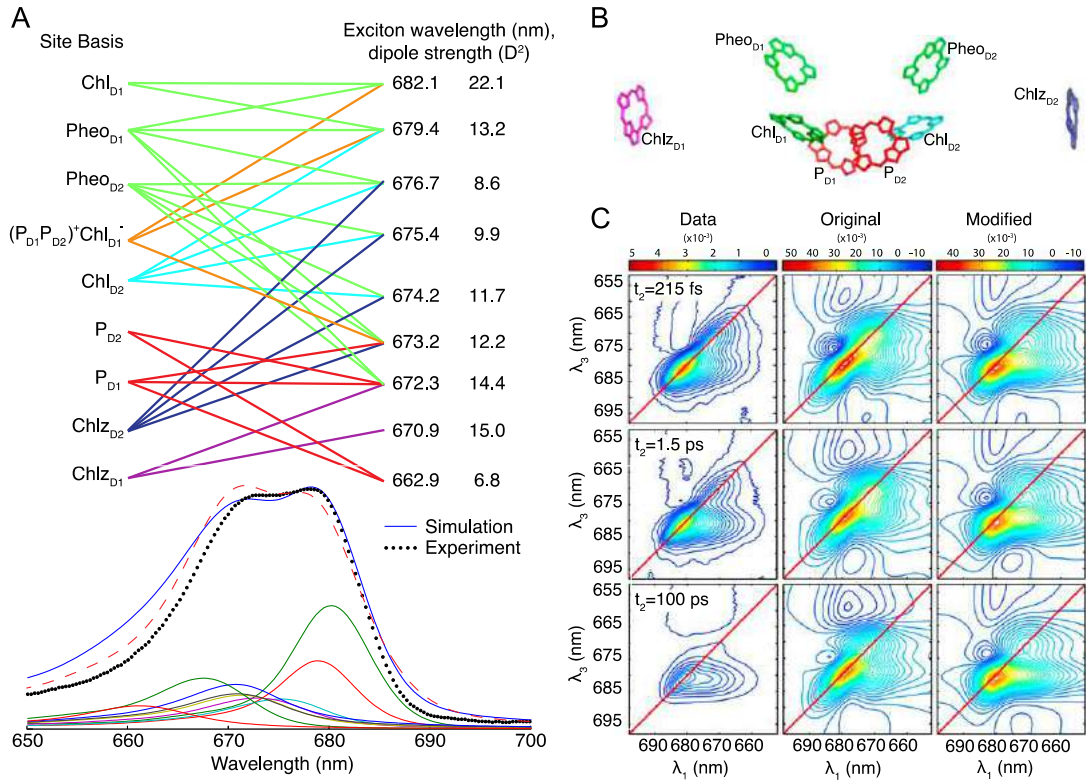

Fig. 4 (A) Excitonic model of the PSII RC. Lines denote any pigments with greater than 10% probability of participating in the connecting exciton state. Also given are the dipole strengths. Bottom, the experimental (black dotted) and simulated 77 K absorption spectrum of the Qy band for the modified (blue solid) and original Novoderezhkin model (red dashed). Also shown are the underlying excitonic contributions calculated for 5000 realizations of disorder for the modified model. (B) Structural arrangement of the PSII RC chlorophyll (Chl) and Pheophytin (Pheo) pigments (carotenoids present in the structure are not shown), where the D1 and D2 subscripts reflect the location of the pigments in the D1 and D2 proteins of the PSII RC. The most strongly coupled pair of Chls are denoted P_{D1} and P_{D2}. (C) Two-dimensional electronic spectra of the PSII RC at 77 K for waiting times T=215 fs, 1.5 ps and 100 ps. Experimental data (left) and simulations based on the Novoderezhkin and modified excitonic models are shown in the middle and right. Adapted from Lewis, K.L.M., Fuller, F.D., Myers, J.A., *et al.*, 2013. Journal of Physical Chemistry A 117 (1), 34–41.

Although the physical origin of coherence can be readily established in simple systems, photosynthetic complexes have poorly understood electronic structure with higher-lying excited states, electronic and vibrational coupling, protein-pigment interactions and dark states. Nevertheless there has been considerable progress made in understanding the 2DES observations of coherence in photosynthetic systems. Jonas and coworkers considered a vibronic dimer and realized that the nonadiabatic electronic and vibrational mixing leads to new pathways for vibrational coherence in the ground electronic states. Within this model they were able to describe the main 2DES signatures of coherence in FMO. In agreement with work by Moran on a different photosynthetic system, they noted that resonance between electronic energy gaps and pigment vibrational modes could have important implications for energy transfer. The Ogilvie and van Grondelle groups have reported coherent dynamics in the PSII RC and have noted the existence of electronic-vibrational resonances in this system, proposing their possible importance for charge separation. In general there is a growing consensus that the coherent dynamics that have been observed to date in photosynthetic system by 2DES are largely vibrational and/or vibronic in origin. It remains an active area of research to understand the functional relevance of electronic-vibrational resonances for energy transfer and charge separation.

Excitonic Electronic Structure and Dynamics in J-Aggregates

Molecular aggregates are clusters of molecules that self-assemble into superstructures with separations typically on the order of their component size. Similar structures are found in nature, for example in the chlorosome antenna of green sulfur bacteria. J-aggregates from cyanine dyes, have been developed as artificial light-harvesting systems aimed at mimicking the efficient energy transfer properties of photosynthetic antennae. These aggregates can be prepared in aqueous solution and form a double-layered tubular strand of $\sim$11 nm outer diameter, total tube wall thickness of $\sim$4 nm, with a lamellar spacing of the individual chromophore layers of $\sim$2.2 nm. Within either chromophore layer, the dye molecules can be considered to be arranged in a brickwork structure. The different molecular electronic coupling strength within layers and between layers leads to a complex electronic structure and rich excitonic dynamics. The linear absorption spectrum of this system exhibits four main peaks, but it is unclear whether or not some

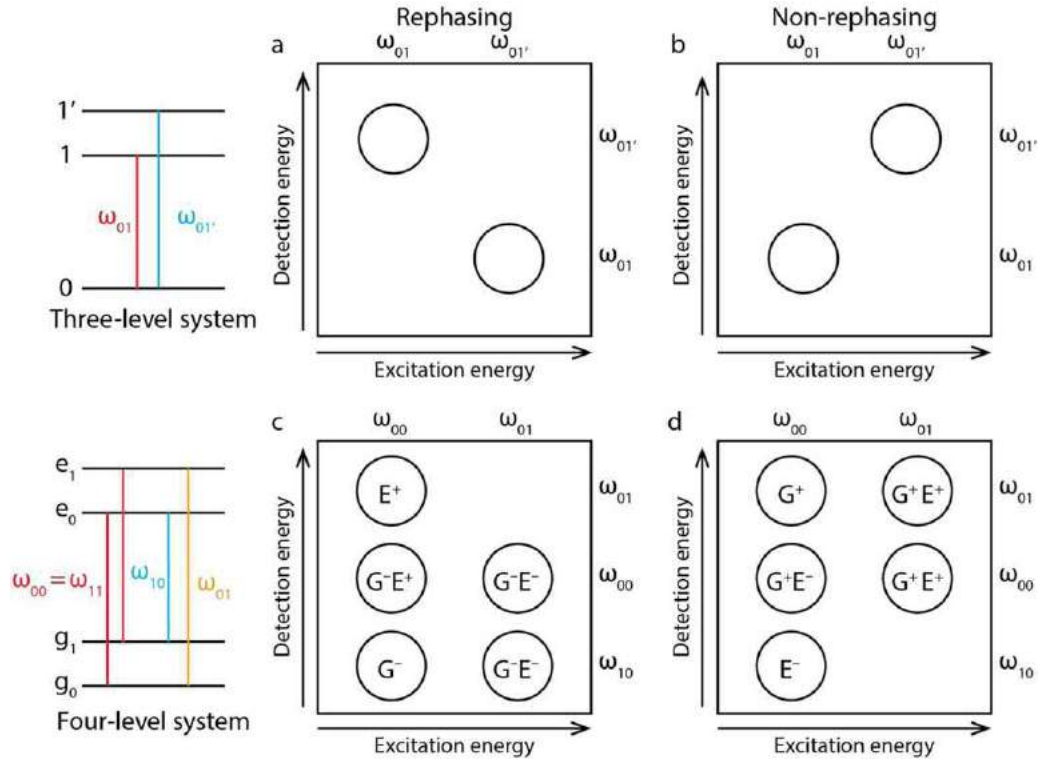

Fig. 5 Coherence amplitude maps for a three-level system (composed of a ground and two excited electronic states) (a, b) and a four-level system (composed of ground and excited electronic states each coupled to one vibrational mode with a single excited vibrational level shown on each electronic state) (c, d). Both rephasing (a, c) and non-rephasing maps (b, d) are displayed. The maps are generated by using double-sided Feynmann diagrams under the assumption that all transitions are allowed. In Fig. (c) and (d), G and E denote vibrational coherence from ground- and excited-electronic states, respectively. +/– denote the signal appears at the positive/negative frequency (f_2).

peaks are from non-aggregated monomers. Kauffman, Hauer and coworkers have performed a series of 2DES studies to investigate the electronic structure and energy transfer pathways within these aggregates. The cross peaks in the 2D map provide clear evidence of the delocalized nature of the absorption bands. By using this model system, the authors also demonstrated the capability to perform global analysis on the 2D dataset to reveal the detailed excitonic relaxation pathways in this system.

Charge Transfer in Organic Photovoltaic Materials

Organic photovoltaics (OPVs) use a blend of an electron donor such as a conjugated polymer, and an electron acceptor such as fullerene to convert sunlight into electricity. The electron donor often serves as the main light-harvester for the system. The primary processes in OPVs are the light absorption by electron donors, the transfer of energy to the donor/acceptor interface, and charge separation at the interface. Mapping the pathways of energy transfer and charge separation and elucidating their mechanism in OPVs is not only of fundamental importance, but also aids in the optimization of devices. Scholes and coworkers investigated charge transfer dynamics in a blend of poly-(3-hexylthiophene) (P3HT) and [6,6]-phenyl-C61 butyric acid methyl ester (PCBM) by using 2D spectroscopy. This blend has very rapid electron transfer. This process is seen as a rise of the P3HT hole signal, indicated by a photoinduced absorption (PIA) that is clearly resolved as a cross-peak in 2DES. One intriguing finding in this study is that vibrational coherence corresponding to C=C stretching mode has been transferred from the excited state to the P3HT hole. It is likely that the coherence here is a spectator mode and does not affect the charge-transfer dynamics. However, the existence of coherence on the P3HT hole implies that charge transfer precedes vibrational relaxation, exposing the significance of hot electron transfer.

Solid State Systems

2DES has been applied to study solids (almost exclusively semiconductors) in the past 10 years. Semiconductors lie at the foundation of the electronic and photonic industry. They are a class of materials with an energy band gap of 0.5 eV to a few eV. There are two different types of semiconductors, indirect gap and direct gap materials. Si and Ge are the most widely applied in indirect gap materials while GaAs, GaP, and InAs are the well-known direct gap materials. Direct gap materials are efficient light absorbers and emitters, thus widely used in light emitting devices such light emitting diodes (LEDs) and lasers. Most LEDs and lasers are built on quantum confined structures such as quantum wells and quantum dots in which the electrons are confined to a

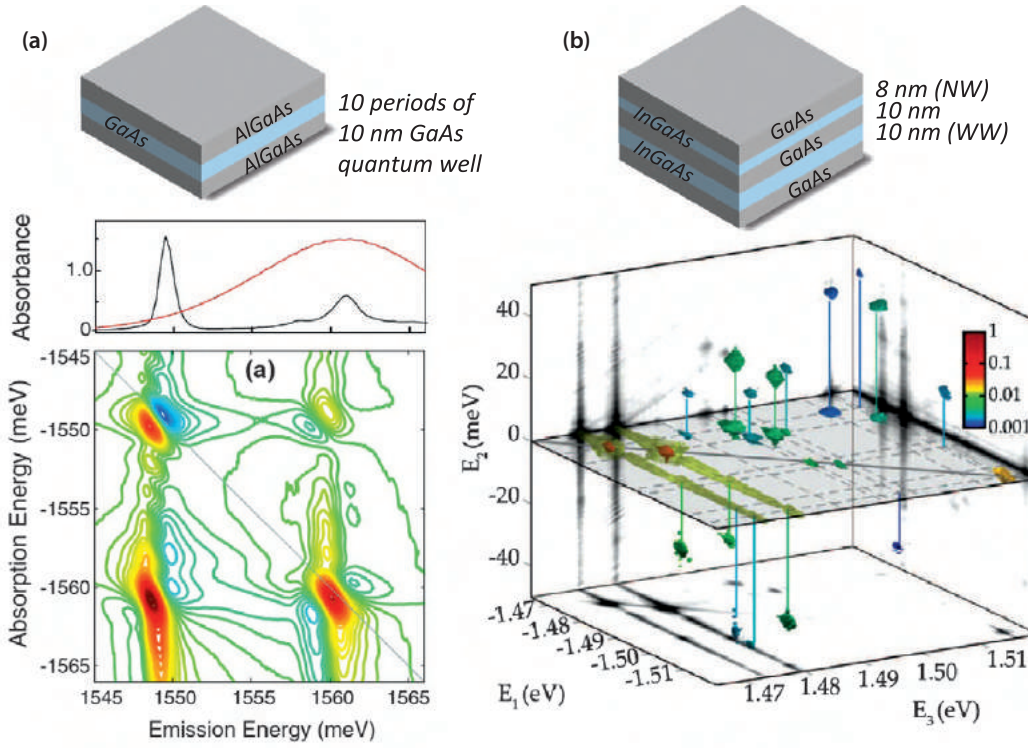

Fig. 6 Illustration of a semiconductor single quantum well (a), a double quantum well (b) and measured 2D spectra from them. (a) The dispersive line shape in the real part of the nonlinear signal originates from many-body interaction induced energy shifts of the excitons. (b) A total of 33 peaks are identified in a 3D spectrum from a double quantum well. Weak diagonal peaks are from dark excitons. Coupling between bright and dark excitons enhances the signal from dark excitons.

two-dimensional plane or localized in all three dimensions. Understanding electronic coupling and quantum coherence are critical for building both conventional optoelectronic devices and future quantum information processing devices.

2DES was first applied to study many body interactions between excitons in quantum wells by Cundiff and coworkers. Excitons in semiconductors are known as Wannier excitons with larger Bohr radius than the Frenkel excitons found in molecular and organic materials. A number of mechanisms such as the local field, biexciton formation, excitation induced dephasing, and excitation induced shift are known to contribute to many body interactions. Using conventional 1D spectroscopy, however, it is difficult to distinguish between these different mechanisms. As one of the earliest applications of 2DES to solids, Cundiff and coworkers explored the phase resolved nonlinear signal in 2DES, which provides valuable information for determining the dominant interaction mechanism. In the example shown in **Fig. 6(a)**, heavy-hole and light-hole exciton resonances in a GaAs/AlGaAs quantum well are simultaneously excited and identified as the diagonal peaks while the coupling between them is manifested as off-diagonal peaks. Most interestingly, the dispersive lineshape of these peaks suggests that the leading many-body effect is excitation induced energy shift.

As a natural extension, 3DES has been used to probe semiconductor asymmetric double quantum wells. Double-well potentials are encountered in many important problems and systems such as diffusion of defects in solids, structural relaxation in glasses, and proton tunneling between DNA bases. A semiconductor double quantum well hosts a rich variety of exciton resonances with tunable energies, thus serving as a useful model system for investigating coupling and dynamics encountered in double well potentials in general. In addition to bright excitons, Davis and coworkers have observed two types of dark excitons: parity-forbidden excitons and spatial indirect excitons in asymmetric InGaAs/GaAs double quantum wells. The observation of these dark excitons with small oscillator strengths was made by possible via off-diagonal coupling peaks between the bright and dark excitons. Such coupling effectively amplifies the signal from dark excitons. In the spectrum shown in **Fig. 6(b)**, a total of 33 peaks were identified, providing unprecedented and comprehensive information on excitonic resonances, coupling among them, and associated quantum dynamics in double quantum wells.

In semiconductor quantum dots, excitons are confined in all three dimensions, leading to bound many-body states. For example, trions (excitons bound with an additional charge) and biexcitons (a bound state of two excitons) have all been clearly identified in 2DES performed on an ensemble of quantum dots. The biexcitons are clearly identified as an off-diagonal peak corresponding to absorption at the exciton state and emission at an energy shifted toward lower energy by an amount corresponding to the biexciton binding energy. The high sensitivity of 2DES allows nonlinear signals from just a few quantum dots to be measured. The excitons confined in each quantum dot are slightly shifted in energy due to different confinement potentials, leading to multiple isolated diagonal peaks. Quantum dots that are coupled to each other can be identified by off-diagonal peaks.

Spatial imaging of individual quantum dots further determines the separation between the coupled quantum dots. As expected, the coupling strength reduces as the separation between quantum dots increases.

References

Brixner, T., Stenger, J., Vaswani, H.M., *et al.*, 2005. Nature 434 (7033), 625–628.
Cho, M., 2009. Two-Dimensional Optical Spectroscopy. New York: CRC Press, p. 385.
Cho, M.H., 2008. Chemical Reviews 108 (4), 1331–1418.
Fuller, F.D., Ogilvie, J.P., 2015. Annual Reviews of Physical Chemistry 66, 667–690.
Ginsberg, N.S., Cheng, Y.C., Fleming, G.R., 2009. Accounts of Chemical Research 42 (9), 1352–1363.
Hamm, P., Zanni, M.T., 2011. Concepts and Methods of 2D Infrared Spectroscopy. Cambridge: Cambridge University Press, p. 286.
Jonas, D.M., 2003. Annual Review Of Physical Chemistry 54, 425–463.
Lewis, K.L.M., Ogilvie, J.P., 2012. Journal of Physical Chemistry Letters 3, 503–510.
Mukamel, S., Abramavicius, D., Yang, L.J., *et al.*, 2009. Accounts of Chemical Research 42 (4), 553–562.
Ogilvie, J.P., Kubarych, K.J., 2009. Advances in Atomic, Molecular, and Optical Physics 57, 249–321.

Multidimensional Terahertz Spectroscopy

Michael Woerner, Klaus Reimann, and Thomas Elsaesser, Max Born Institute for Nonlinear Optics and Short-Pulse Spectroscopy, Berlin, Germany

Introduction

Over the last two decades, nonlinear terahertz (THz) spectroscopy (Luo *et al.*, 2004; Gaal *et al.*, 2006) has developed as a new field of condensed matter research, due to both substantial progress in the generation of THz pulses with high electric field amplitudes and new experimental concepts such as, e.g., THz pump–midinfrared (MIR) probe experiments (Gaal *et al.*, 2007; Günter *et al.*, 2009; Huber *et al.*, 2001, 2005; Kampfrath *et al.*, 2011; Pashkin *et al.*, 2011; Hwang *et al.*, 2015). THz pulses have a subpico- to picosecond duration and cover a frequency range from 0.3 to 30 THz or 10–1000 cm^{-1}, corresponding to photon energies from 1.25 to 125 meV and far-infrared wavelengths between 1 mm and 10 μm. Elementary excitations of condensed matter in this energy range include nuclear motions, i.e., translations, intra- and intermolecular vibrations as well as phonons in solids, and low-frequency electronic excitations such as excited-state absorption in large molecules or intraband, interband, defect-related and intra-excitonic transitions in solids (Tonouchi, 2007). Nonequilibrium dynamics and correlations of low-frequency excitations are not fully understood and call for time-resolved nonlinear spectroscopies in that spectral range. A second class of phenomena are processes driven by strong nonresonant electric fields such as carrier transport in the quantum kinetic regime or field-induced shifts of electronic states and transitions between them.

A preeminent feature of THz spectroscopy consists in the phase-resolved detection of electric field transients, allowing for direct time-domain measurements of THz fields in amplitude and absolute phase (Wu and Zhang, 1995). Both the spectral (Somma *et al.*, 2015) and the dynamic range of the field measurement has been improved significantly, now allowing for sophisticated experimental techniques involving phase-locked sequences of THz pulses (Leitenstorfer *et al.*, 1999; Huber *et al.*, 2000; Brodschelm *et al.*, 2001; Kübler *et al.*, 2004, 2005; Sell *et al.*, 2008; Junginger *et al.*, 2010; Schubert *et al.*, 2011). Nonlinear two-dimensional (2D) THz spectroscopy is a prominent method of this type, allowing for a measurement of couplings between electronic or vibrational excitations, for mapping structural fluctuations via time-dependent lineshapes, and for following chemical exchange processes in time (Mukamel, 2000; Jonas, 2003; Hamm *et al.*, 1998; Asplund *et al.*, 2000).

Multi-dimensional experiments in nuclear magnetic resonance (NMR) are typically performed in the nonperturbative regime using more than three radio-frequency pulses (Ernst *et al.*, 1997), while 2D optical spectroscopy has mainly been performed in the third-order or $\chi^{(3)}$ limit by applying pump–probe or three-pulse photon-echo techniques in noncollinear four-wave-mixing geometries (Hamm and Zanni, 2011). In a three-pulse photon-echo experiment, the pulses interact with the sample in a noncollinear geometry with wavevectors $\vec{k}_A$, $\vec{k}_B$, and $\vec{k}_C$. The first two interactions with the sample are separated by the coherence time τ. The first pulse propagating along $\vec{k}_A$ generates a coherent polarization of the sample, which is transformed into a transient population by the interaction with the second pulse propagating along $\vec{k}_B$. After the waiting period T, during which the population evolves freely, the nonlinear signal is generated by interaction with the electric field of the last pulse propagating along $\vec{k}_C$. Third order, i.e., $\chi^{(3)}$ nonlinear signals are emitted into the directions $-\vec{k}_A + \vec{k}_B + \vec{k}_C$ and $\vec{k}_A - \vec{k}_B + \vec{k}_C$, which are different from those of the three pulses interacting with the sample. In noncollinear 2D spectroscopy, a phase-resolved detection of the emitted third-order signal can be accomplished, e.g., by heterodyning it with a fourth synchronized pulse, the so called local oscillator (LO) (Hamm and Zanni, 2011).

In the THz frequency range, noncollinear n-wave-mixing geometries are inherently difficult if not impossible to realize due to the pronounced diffraction of beams at the large wavelengths involved. As an alternative approach, the fully phase-resolved collinear 2D spectroscopy has been developed for the THz range and been applied to a variety of condensed-phase systems (Kuehn *et al.*, 2009, 2011a,b; Junginger *et al.*, 2012; Woerner *et al.*, 2013; Bowlan *et al.*, 2014a,b; Folpini *et al.*, 2015; Somma *et al.*, 2014, 2016a,b). In this article, we describe the basic concepts and the most recent development in multidimensional THz spectroscopy. The article is organized as follows. In Section "Concepts of Collinear Multidimensional THz Spectroscopy", we introduce the concept of frequency vectors which is fully equivalent to that of wavevectors in noncollinear 2D spectroscopy. Special diagrams exploiting chains of frequency vectors in 2D frequency space allow for disentangling quantum pathways in Liouville space, even in experiments where different orders of nonlinearities such as $\chi^{(3)}$, $\chi^{(5)}$, etc. occur simultaneously. In Section "Two-dimensional THz Spectroscopy with Three Phase-locked THz Pulses" we introduce an experimental setup for 2D THz spectroscopy with three independent THz pulses, followed by a discussion of two prototype experiments on intersubband coherences in semiconductor quantum wells (Section "THz-driven AC Stark Effect of Intersubband Transitions in Semiconductor Quantum Wells") and two-photon interband coherences in the semiconductor InSb (Section "Three-pulse 2D THz Spectroscopy on InSb"). A brief summary and outlook are given in Section "Conclusions".

Concepts of Collinear Multidimensional THz Spectroscopy

Multidimensional THz spectroscopy is typically implemented in a geometry where a sequence of THz pulses propagating in collinear direction interacts with the sample. The resulting nonlinear polarization (or current) emits a THz electric field which is

detected in the same direction. A pulse sequence applied in three-pulse experiments is schematically shown in **Fig. 1(a)** with pulses A (red) and B (green) separated by the coherence time τ and pulses B and C (blue) by the waiting or population time T. After the last interaction, the nonlinear polarization or current in the sample emits the electric field $E_{NL}(\tau,T,t)$ (black curve).

In the following, the coherence time τ is measured relative to the maximum of pulse B, whereas the population time T and the real time t are measured relative to the maximum of pulse C. As a result, the field $E_A(\tau,T,t)$ depends on τ, T and the real time t, the electric field $E_B(T,t)$ depends on T and t, and $E_C(t)$ on t only. A sequence with pulse A preceding pulse B preceding pulse C implies $\tau<0$ and $T<0$. Thus, the center of pulse A occurs at the real time $t_A=T+\tau$, the center of pulse B at $t_B=T$ and the center of pulse C at $t_C=0$. **Fig. 1(b)** shows a 2D plot of the timing scheme as a function of both real time t and coherence time τ with the colored lines representing the centers of the THz pulses. A 3D Fourier transform along τ, T, and t provides the frequency dependent fields $\tilde{E}_A(\nu_\tau,\nu_T,\nu_t)$, $\tilde{E}_B(\nu_T,\nu_t)$, and $\tilde{E}_C(\nu_t)$. As a result, the orientations of frequency vectors which are perpendicular to the phase fronts of the pulses in a space with three time dimensions, are linearly independent of each other (Kuehn *et al.*, 2011a). For the pulse sequence shown in **Fig. 1(b)** the corresponding frequency vectors are:

$$\vec{\nu}_A=\begin{pmatrix}\nu_\tau\\ \nu_T\\ \nu_t\end{pmatrix}=\begin{pmatrix}-\nu_0\\ \nu_0\\ \nu_0\end{pmatrix}\qquad \vec{\nu}_B=\begin{pmatrix}0\\ \nu_0\\ \nu_0\end{pmatrix}\qquad \vec{\nu}_C=\begin{pmatrix}0\\ 0\\ \nu_0\end{pmatrix} \tag{1}$$

The orientation of the frequency vectors of pulses A, B, and C are shown as red, green, and blue arrows in **Fig. 1(b)**, respectively. A nonlinear signal due to a certain interaction sequence in Liouville space can be identified by the orientation of its phase fronts, in reference to the phase fronts of the three pulses. For third-order, i.e., $\chi^{(3)}$ signals, two different types of interaction sequences are distinguished, the so-called rephasing response, in which the phase differences acquired during the coherence period τ are reversed by interaction with the third pulse, and the nonrephasing response in which phase differences from the different interactions add up (Khalil *et al.*, 2003). For the timing scheme shown in **Fig. 1(b)**, the phase fronts of the rephasing and nonrephasing nonlinear signals are shown as magenta and cyan lines, respectively. The phase fronts of the nonrephasing signal and of pulse 1 are parallel while the rephasing signal exhibits phase fronts perpendicular to those of pulse 1.

A standard approach for describing Liouville space pathways connected with a third-order nonlinear response are double-sided Feynman diagrams (Mukamel, 2000; Hamm and Zanni, 2011; Yee and Gustafson, 1978; Boyd and Mukamel, 1984; Mukamel, 1995; Yang *et al.*, 2008). The time evolution of the density matrix describing a rephasing contribution (left) and its nonrephasing counterpart (right) are shown in **Fig. 1(c)**. Double-sided Feynman diagrams clearly show the time evolution of the |ket⟩ and ⟨bra|

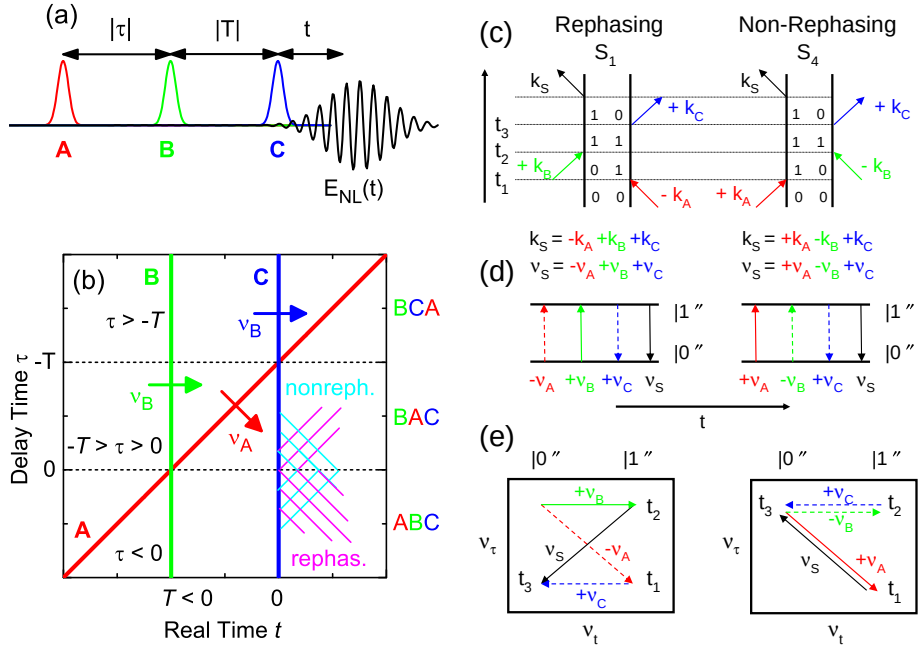

Fig. 1 (a) Timing scheme of a nonlinear, two-dimensional (2D) experiment using three THz pulses, pulse A (red), pulse B (green), and pulse C (blue). After the last pulse in the timing sequence the nonlinear current in the sample emits the electric field $E_{NL}(t)$ (black curve). (b) 2D plot of the timing scheme as a function of both real time t and delay time τ. Shown are the centers of pulses. For $\tau\leq 0$ (pulse sequence ABC) pulses A and B are separated by the coherence time τ, pulses B and C by the waiting time T, and the detection time t starts with pulse C. Arrows: orientation of the corresponding frequency vectors. The phase fronts of the rephasing and nonrephasing nonlinear signals are shown as magenta and cyan lines, respectively. (c) Standard double-sided Feynman diagrams describing the Liouville pathway of a third-order nonlinear response for a rephasing (left) and a nonrephasing contribution (right). (d) Liouville pathway diagrams in the style of Pollard *et al.* (1992). (e) Alternative diagrams exploiting frequency vectors used in the THz community (Kuehn *et al.*, 2009a,b).

Table 1 Coherence, waiting, and detection times and frequency vectors for different pulse sequences.

pulse sequence	*Coherence time*	*Waiting time*	*Detection time*	*Rephasing freq. vector*	*Nonrephasing freq. vector*
ABC	$t_B - t_A = -\tau$	$t_C - t_B = -T$	t	$\begin{pmatrix} \nu_0 \\ 0 \\ \nu_0 \end{pmatrix}$	$\begin{pmatrix} -\nu_0 \\ 0 \\ \nu_0 \end{pmatrix}$
BAC	$t_A - t_B = \tau$	$t_C - t_A = -T-\tau$	t	$\begin{pmatrix} -\nu_0 \\ 0 \\ \nu_0 \end{pmatrix}$	$\begin{pmatrix} \nu_0 \\ 0 \\ \nu_0 \end{pmatrix}$
BCA	$t_C - t_B = -T$	$t_A - t_C = T+\tau$	$t-T-\tau$	$\begin{pmatrix} -\nu_0 \\ 0 \\ \nu_0 \end{pmatrix}$	$\begin{pmatrix} -\nu_0 \\ 2\nu_0 \\ \nu_0 \end{pmatrix}$

side of the density matrix in Liouville space. Energetic aspects and phase matching conditions in wavevector or frequency vector space are not fully obvious in this picture. In panel (d), we show the same Liouville pathways in the style of Pollard *et al.* (1992). Here, the |ket⟩ evolution is represented by a chain of solid arrows whereas the ⟨bra| evolution by dashed arrows. Transition energies are clearly visible in this type of diagram, frequently used in recent 2D THz experiments (Somma *et al.*, 2016a,b). In the 2D THz community (Kuehn *et al.*, 2009, 2011a) a third, alternative type of diagrams has been developed [cf. **Fig. 1(e)**], which exploits the frequency vector components ν_τ and ν_t of Eq. (1) to visualize the Liouville pathway in 2D frequency space. The advantage of the latter diagrams is that the frequency vector chain on the |ket⟩ side of the density matrix (solid vectors) and that on the ⟨bra| side (dashed vectors) determine both the orientation of the phase fronts of the driving and nonlinearly emitted fields. For instance, the diagrams shown in **Fig. 1(e)** predict correctly the orientation of the phase fronts of the rephasing $-\vec{\nu}_A + \vec{\nu}_B + \vec{\nu}_C$ (magenta) and nonrephasing $\vec{\nu}_A - \vec{\nu}_B + \vec{\nu}_C$ (cyan) contributions for pulse sequence ABC as shown in panel (b). For the different pulse sequences ABC, BAC, and BCA both the frequency vectors and the coherence, waiting, and detection times change their roles. Details are summarized in **Table 1**.

In the collinear THz 2D experiments performed so far (Somma *et al.*, 2016a,b), the pulse separation $t_C - t_B = -T$ has been fixed with the implication that the waiting time period T and correspondingly the ν_T components of the frequency vectors are not accessible. Consequently, one cannot separate rephasing from nonrephasing signal contributions for pulse sequence BCA (last row in **Table 1** and $\tau \geq -T$ in **Fig. 1(b)**). For the pulse sequence BAC ($-T \geq \tau \geq 0$) the interpretation of the experiment is complicated by the fact that the waiting time $-T-\tau$ is simultaneously varied with the coherence time τ. Thus, the pulse sequence ABC ($\tau \leq 0$) is typically used to analyze nonlinear signals in the $\chi^{(3)}$ limit. As we shall see below, the simple $\chi^{(3)}$ classification scheme according to **Table 1** breaks down for higher-order nonlinearities $\chi^{(n)}$ with $n > 3$. In this regime, all pulse sequences ABC, BAC, and BCA are useful for the data analysis.

Two-Dimensional THz Spectroscopy With Three Phase-Locked THz Pulses

The experimental implementation of fully phase-resolved collinear 2D THz spectroscopy is discussed in this section (Kuehn *et al.*, 2009, 2011a,b; Junginger *et al.*, 2012; Woerner *et al.*, 2013; Bowlan *et al.*, 2014a,b; Folpini *et al.*, 2015; Somma *et al.*, 2014). The experiments of Somma *et al.* (2016a,b) were performed with three phase-locked THz pulses in the setup schematically shown in **Fig. 2**. A mode-locked Ti:sapphire oscillator generates linearly polarized 12 fs pulses centered at 800 nm at an 85 MHz repetition rate. A small part of the oscillator output is separated by a beam splitter to serve as a probe in free-space electrooptic (EO) sampling (Wu and Zhang, 1997). The major fraction of the oscillator pulses feeds two multi-pass amplifiers. Each amplifier delivers 25 fs pulses with an energy of up to 1 mJ at a repetition rate of 1 kHz. Pulses of the first amplifier are split into two parts of identical energy and mutual delay τ while the output pulse of the second amplifier is delayed by an additional time interval T. The three 800-nm beams are sent onto three separate GaSe crystals to generate the sub-picosecond phase-locked THz pulses A, B, and C. Phase-matched optical rectification within the pulse spectrum (Reimann *et al.*, 2003) provides almost octave-spanning pulses with a center frequency at 20 THz. Behind the GaSe crystals, the 800 nm beams are blocked by silicon wafers. Using a 90° off-axis parabolic mirror the three THz pulses are tightly focused on the sample. A couple of off-axis parabolic mirrors is used to image the THz electric fields transmitted through the sample onto a 10-μm-thick (110)-oriented ZnTe crystal to be detected by free-space electrooptic sampling (Wu and Zhang, 1997). By varying the real time t between the three THz pulses and the probe pulse from the Ti:sapphire oscillator, the electric field of the THz pulses is measured in amplitude and phase. In **Fig. 3(b)** we show the driving fields used in the 2D THz experiments of Somma *et al.* (2016a,b).

Mechanical choppers are placed in each beam path in order to retrieve the nonlinear response of the sample. The three choppers are synchronized to the 1 kHz repetition rate of the laser system and allow for measuring seven transients, one when all three pulses ABC interact with the sample, three for the pairs AB, BC, or CA, and three for the single pulses A, B, or C. From the

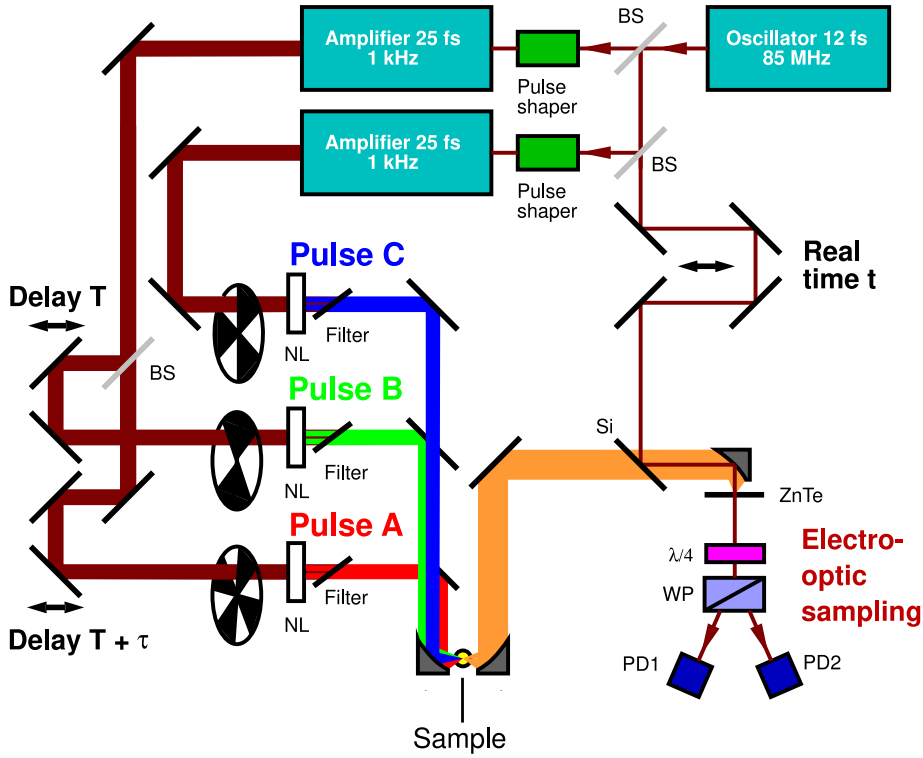

Fig. 2 Schematic view of the experimental setup for 2D THz spectroscopy used in Somma *et al.* (2016a,b). Intense THz electric fields are generated via optical rectification of near-infrared pulses in nonlinear crystals (NL) and detected by electrooptic (EO) sampling in a thin ZnTe crystal. Two acousto-optic pulse shapers, implemented in both amplifiers, serve to optimize the THz generation by tailoring the spectral components of the pulses from the oscillator.

measured transients, the nonlinear field emitted by the sample is given by

$$E_{\mathrm{NL}}(\tau,T,t) = E_{\mathrm{ABC}}(\tau,T,t)$$
$$-E_{\mathrm{AB}}(\tau,T,t) - E_{\mathrm{BC}}(T,t) - E_{\mathrm{CA}}(\tau,T,t)$$
$$+E_{\mathrm{A}}(\tau,T,t) + E_{\mathrm{B}}(T,t) + E_{\mathrm{C}}(t). \tag{2}$$

Most of the fully phase-resolved 2D collinear THz experiments reported so far were performed with just two THz pulses (Kuehn *et al.*, 2009, 2011a,b; Junginger *et al.*, 2012; Woerner *et al.*, 2013; Bowlan *et al.*, 2014a,b; Folpini *et al.*, 2015; Somma *et al.*, 2014). A special version with two pulses are experiments with an intense THz transient $E_{\mathrm{THz}}(\tau,t)$ and a weaker electric field transient in the midinfrared spectral range $E_{\mathrm{MIR}}(t)$ (Gaal *et al.*, 2006; Folpini *et al.*, 2015). In the two-pulse case the nonlinear signal has a simpler expression:

$$E_{\mathrm{NL}}(\tau,t) = E_{\mathrm{THz+MIR}}(\tau,t) - E_{\mathrm{THz}}(\tau,t) - E_{\mathrm{MIR}}(t). \tag{3}$$

THz-Driven AC Stark Effect of Intersubband Transitions in Semiconductor Quantum Wells

So far, nearly all studies based on 2D spectroscopy have concentrated on resonant interactions between light and the system under study, i.e., a sequence of pulses interacts with single- and/or multi-photon resonances. The application of distinctly off-resonant electric field transients for controlling resonant optical excitations has remained scarce (Wagner *et al.*, 2010). Off-resonant coherent control requires a high amplitude of the off-resonant field with frequency (ν) to make the interaction energy with the system comparable to the transition energy of the optical resonance.

A basic concept for describing off-resonant coherent control is the so called dressed-state picture developed by Autler and Townes (1955) and Bonch-Bruevich and Khodovoi (1967). In this picture, one considers not just the electronic states but combined states of electron and light, e.g., the state consisting of the electron in level 1 and one photon with energy $h\nu$, $|n=1, 1\ h\nu\rangle$ (see **Fig. 1(a)** of Folpini *et al.* (2015)). Because of the interaction with the electric field, the states $|n=1, 1\ h\nu\rangle$ and $|n=2, 0\ h\nu\rangle$ repel each other, leading to a blueshift of the transition frequency for $\nu < \nu_0$ and to a redshift for $\nu > \nu_0$ (ν_0 is the unperturbed transition frequency between states $n=1$ and $n=2$). We show in the following how an off-resonant perturbation of a midinfrared dipole transition by a strong THz field is mapped by 2D spectroscopy.

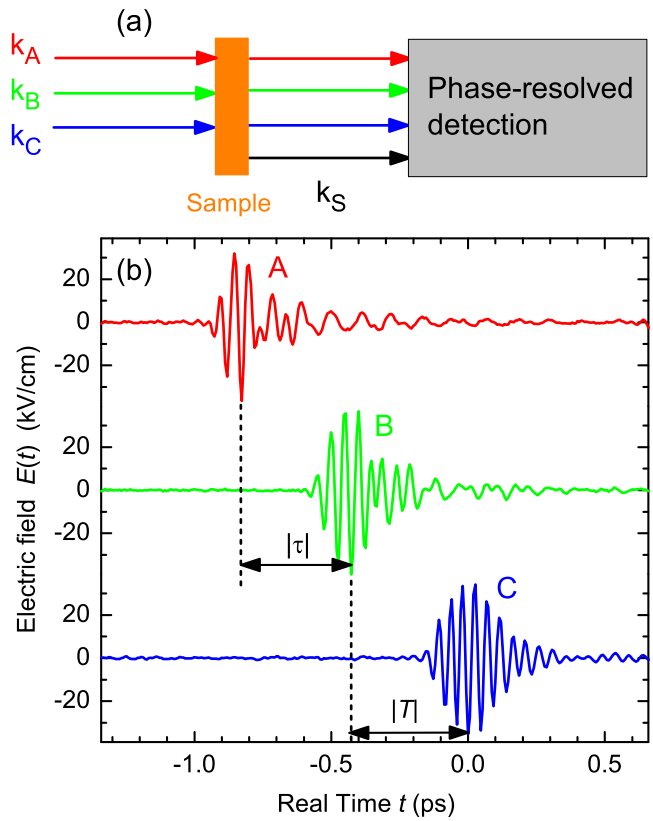

Fig. 3 (a) Schematic of the collinear interaction geometry in 2D terahertz (THz) spectroscopy. The electric field transmitted through the sample is measured in amplitude and phase by electrooptic sampling. (b) Sequence of three phase locked THz pulses: A and B are separated by the coherence time $\tau = -0.4$ ps and B and C are separated by the waiting time $T = -0.43$ ps. The electric field is plotted as a function of the real time t.

The model system under study is the optical transition between different confined subbands in a semiconductor quantum well (QW), a fundamental elementary excitation of free electrons confined in two dimensions (Helm, 2000). Such intersubband (IS) excitations represent sensitive probes of the complex many-body and scattering behavior of electrons and have been studied by both experiment and theory (Luo *et al.*, 2004; Delteil *et al.*, 2012; Wagner *et al.*, 2011; Dietze *et al.*, 2012; Shih *et al.*, 2005). In the present experiments, a sample grown by molecular beam epitaxy on an insulating GaAs substrate was investigated. It contains 20 GaAs QWs of $L = 11$ nm width, separated by 20 nm wide $Al_{0.35}Ga_{0.65}As$ barriers. The barriers are δ-doped to achieve an electron concentration of 5×10^{11} cm^{-2} per QW. As the thickness of the QW stack is small compared to both THz and MIR wavelengths, all QWs experience a similar electric field (Stroucken *et al.*, 1996). The $1 \rightarrow 2$ IS absorption is centered at a frequency of $\nu_0 = 21$ THz.

In the experiment, the nonresonant THz field manipulates the linear IS polarization, an effect manifested in a well-defined phase shift of the coherent IS emission. **Fig. 4** shows the time-domain two-color 2D signal from the semiconductor quantum well sample interacting with a THz pulse centered at 1.2 THz and a mid-infrared pulse centered at 21 THz. In panel (a) the sum of the electric fields $E_{\mathrm{THz}}(\tau,t) + E_{\mathrm{MIR}}(t)$ transmitted through the sample is plotted as a function of the real time t and the delay or coherence time τ. Panel (b) displays the nonlinear signal $E_{\mathrm{NL}}(t,\tau)$ in the mid-infrared according to Eq. (3). The dashed lines mark the centers of the two pulses in time and $\tau = 0$. A prototype electric field transient $E_{\mathrm{THz}}(\tau = 0,t) + E_{\mathrm{MIR}}(t)$ in the pulse overlap is shown in panel (c). The corresponding nonlinearly emitted field $E_{\mathrm{NL}}(\tau = 0,t)$ is shown in panel (d).

The nonlinear signal represents, in principle, a response up to arbitrary order in the incident fields. For the present, comparably small THz field amplitude, however, the nonlinear signal is well accounted for by considering the third-order response. The signal occurs predominantly at positive values of τ and mainly after the MIR pulse at long t. This time structure gives direct evidence of a perturbed free induction decay (PFID) (Chachisvilis *et al.*, 1995). To separate the different nonlinear contributions, we derive the 2D spectrum $E_{NL}(\nu_t,\nu_\tau)$ as a function of ν_t and ν_τ by a two-dimensional Fourier transform (Kuehn *et al.*, 2009, 2011a). As shown in **Fig. 5(b)**, only a THz-pump–MIR-probe signal E_{NL} at the frequency position $(\nu_t = \nu_0, \nu_\tau = 0)$ is present while MIR-pump–THz-probe and four-wave-mixing signals are absent. For a higher signal-to-noise ratio, $E_{NL}(\nu_t,\nu_\tau)$ is filtered to keep only the THz-pump–MIR-probe component and then back transformed to the time domain as shown in **Fig. 4(d)**. Using the latter nonlinear transient one can calculate the measured spectrally resolved THz-pump–MIR-probe signal as a function of ν_t and pump–probe delay τ, which is plotted in **Fig. 5(c)**. The dispersive lineshape shows clearly a THz-field-induced blue shift of the intersubband transition. Panel (d) displays the spectrally resolved pump–probe signal calculated with help of perturbation theory which will be discussed next.

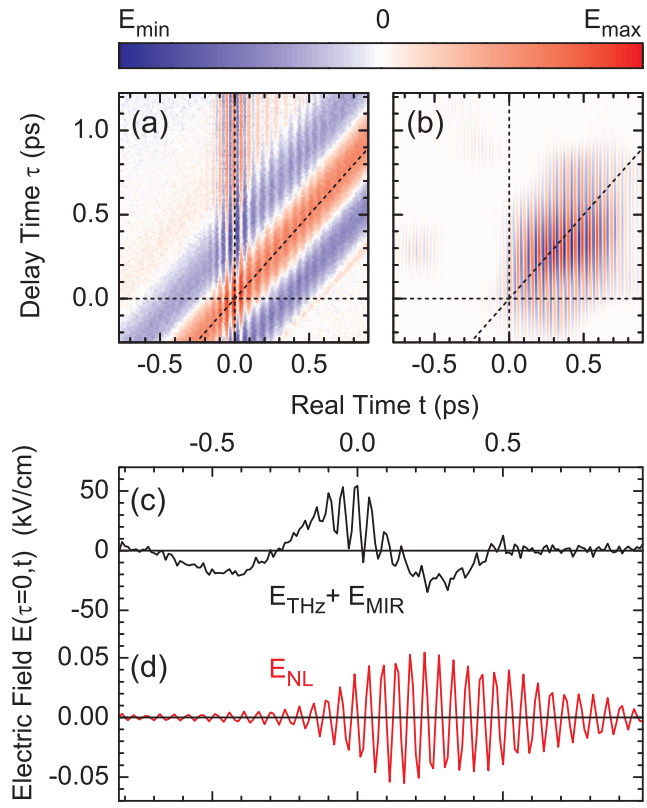

Fig. 4 Two-color 2D signal from a semiconductor quantum well sample interacting with a THz pulse centered at 1.2 THz and a mid-infrared pulse centered at 21 THz. (a) The sum of the electric fields $E_{THz}(\tau,t)+E_{MIR}(t)$ transmitted through the sample is plotted as a function of the real time t and the delay τ. (b) Nonlinear signal $E_{NL}(t,\tau)$ in the mid-infrared. The dashed lines show the centers of the two pulses and $\tau=0$. (c) Electric field transient $E_{THz}(\tau=0,t)+E_{MIR}(t)$. (d) The corresponding nonlinearly emitted field $E_{NL}(\tau=0,t)$ shows a perturbed free induction decay of the intersubband polarization as a function of real time t.

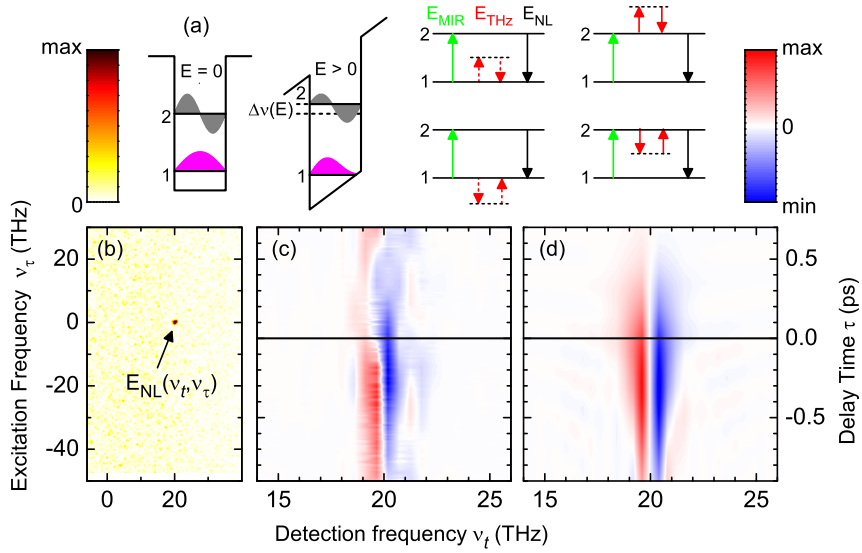

Fig. 5 (a) Schematic of the potential and wave functions of the quantum well without (left) and with (middle) applied THz field. For clarity, the electric field and the level shifts shown are much larger than in the actual experiment. (b) Right: lowest order (i.e. $\chi^{(3)}$) Liouville pathway diagrams in the style of Pollard *et al.* (1992) describing the AC Stark effect on intersubband transitions. (b) A 2D Fourier transform of the nonlinear signal shown in **Fig. 4(b)** gives the two-dimensional spectral density $E(\nu_\tau,\nu_t)$ as a function of the detection frequency ν_t and the excitation frequency ν_τ. (c) Measured spectrally resolved THz-pump–MIR-probe signal as a function of ν_t and pump–probe delay τ. (d) Calculated spectrally resolved pump–probe signal corresponding to the experiment shown in (c).

The electronic levels are determined by the confinement potential and wave functions of the quantum well without the applied THz field (lhs of **Fig. 5(a)**). In contrast to the dressed state picture, i.e., a modified QW potential (middle of **Fig. 5(a)**), perturbation theory treats the interaction of the intersubband transition with both the THz and the MIR field on equal footing. The rhs of **Fig. 5 (a)** shows the relevant lowest order, i.e., $\chi^{(3)}$ Liouville pathway diagrams in the style of Pollard *et al.* (1992). The short arrows describe interactions with the nonresonant THz field while the levels and the long arrows stand for IS excitations. The off-resonant interaction with the THz pulse allows for sequences in which first a THz photon is emitted and later reabsorbed by the system. Such 'acausal' interaction sequences are important when stepping through 'virtual' states. All the diagrams shown are nonrephasing as the THz field cannot revert the phase of the IS coherence. It is important to note that such nonrephasing diagrams do not survive the rotating wave approximation which is the basis of standard analysis in 2D spectroscopy (Hamm and Zanni, 2011). Moreover, the nonrephasing diagrams shown on the rhs of **Fig. 5(a)** do not have any rephasing counterparts. As a consequence, it is impossible to construct a purely absorptive 2D spectrum as the sum of rephasing and nonrephasing diagrams (Hamm and Zanni, 2011), whenever off-resonant interaction sequences in Liouville space dominate.

Three-Pulse 2D THz Spectroscopy on InSb

Off-resonant interaction sequences also dominate in our second example, in which 'dark' two-photon coherences in the semiconductor InSb are investigated. InSb is the III-V semiconductor with the smallest band gap, which has a value at room temperature of $E_g = 0.17$ eV corresponding to a frequency of $\nu_{THz} = E_g/h = 41$ THz. The well known electronic band structure of InSb (Kane, 1957) is shown in **Fig. 5(a)** of Somma et al. (2016a). The THz pulses of **Fig. 3(b)** with a center frequency of 21 THz are neither resonant to the bandgap nor to any other resonance in the unexcited crystal. The THz peak field strength $E \approx 50$ kV/cm and the extraordinarily large interband transition dipole $d_{cv} = e_0 \cdot (4\ \text{nm})$ at the Γ point (e_0: elementary charge) (Kane, 1957, 1959) result in a Rabi frequency $\Omega_{\text{Rabi}} = E\ d_{cv}/\hbar = 3 \times 10^{13}\ \text{s}^{-1}$, i.e., comparable to the THz carrier frequency. As a result, the 2D experiments presented in the following are in the strongly nonperturbative regime of light–matter interaction, which is characterized by the simultaneous occurrence of nonlinear polarizations of different orders $\chi^{(n)}$, similar to Rabi oscillations on intersubband transitions in quantum wells (Kuehn *et al.*, 2009; Folpini *et al.*, 2015). In particular, the nonperturbative regime allows for multiple interactions of a single THz transient with the InSb crystal. In Somma *et al.* (2016a,b) we reported the ultrafast dynamics of two-phonon coherences and signals related to multiple two-photon excitations of electron-hole pairs in the III-V semiconductor. In the following, we concentrate on the dynamics of 'dark' two-photon coherences in InSb, studying a 70 μm thick (100)-oriented InSb single crystal with a low *n*-type doping ($n \lesssim 10^{16}\ \text{cm}^{-3}$). All measurements were performed at ambient temperature (300 K).

Results of the 2D THz experiment performed with the three pulses A, B, and C of **Fig. 3(b)** are summarized in **Fig. 6**. In panel (a), the sum of electric field transients $E_A(\tau,T,t) + E_B(T,t) + E_C(t)$ transmitted through the InSb sample is shown in a contour plot as

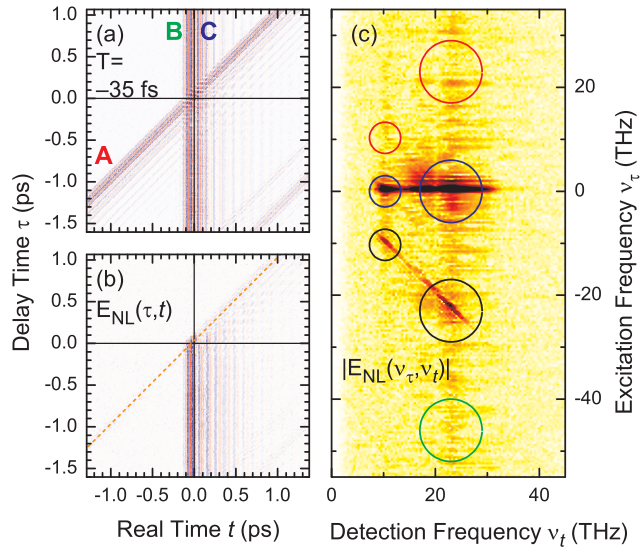

Fig. 6 Two-dimensional spectroscopy on InSb using the three THz pulses shown in **Fig. 3(b)**. (a) Contour plot of the sum of electric field transients $E_A(\tau,T,t) + E_B(T,t) + E_C(t)$ transmitted through the InSb sample as a function of the coherence time τ and real time t for the waiting time $T = -35$ fs. (b) Nonlinear signal $E_{NL}(\tau,T,t)$ according to Eq. (2). The orange dashed line indicates the center of pulse A. (c) Contour plot of the amplitude $|E_{NL}((\nu)_\tau,(\nu)_t)|$ which is the 2D Fourier transform of $E_{NL}(\tau,T,t)$. The colored circles indicate the position of relevant signals in the 2D frequency space. The linear amplitude scales of the 2D scans range over (a) ±76.5 kV/cm and (b) ±30 kV/cm.

a function of the coherence time τ and real time t for the waiting time $T=-35$ fs. In the following all contour plots are normalized to their respective maximum signals. The corresponding nonlinear signal $E_{NL}(\tau,T,t)$ derived with the help of Eq. (2) is shown in panel (b). In accordance with the law of causality a nonvanishing $E_{NL}(\tau,T,t)$ is exclusively observed starting with or after the last pulse in the timing sequence. For better orientation, the center of pulse A is indicated by the orange dashed line, which intersects the horizontal $\tau=0$ line (black) at $t=-T$. In **Fig. 6(c)** the contour plot of the amplitude $|E_{NL}(\nu_\tau,\nu_t)|$ is displayed, obtained by a 2D Fourier transform of $E_{NL}(\tau,T=-35\text{ fs},t)$.

The circles of different size and color indicate the positions of relevant signals in the 2D frequency space. Concerning the detection frequency ν_t, strong nonlinear signals occur in the spectral range of the driving pulses 15 THz $< \nu_t <$ 25 THz (large circles) and at the two-phonon resonance $\nu_t=10$ THz (small circles). The latter signals have been discussed in detail in Somma *et al.* (2016a). As a function of the excitation frequency ν_τ we observe significant nonlinear signals for $\nu_\tau=0$ (blue circles), $\nu_\tau=-\nu_t$ (black circles), $\nu_\tau=+\nu_t$ (red circles), and $\nu_\tau=-2\nu_t$ (green circle). The discussion in the following focuses of the signal in the green circle. It should be emphasized that the fully phase-resolved detection of electric field transients allows for a highly accurate measurement of the waiting time T. This knowledge becomes important in situations in which coherences evolve along the waiting period, the phase of which is determined by the exact value of T. For instance, recent 2D experiments on dipole-dipole interactions in a rubidium gas were detected via the time evolution of the two-atom coherence evolving along the waiting period T (Dai *et al.*, 2012; Gao *et al.*, 2016). The 'dark' two-photon coherence excited via two-photon excitation of InSb is measured in a similar scheme.

Fig. 7(a) and **(b)** shows the time domain nonlinear signal derived from the peak at ($\nu_\tau=-44$ THz,$\nu_t=+22$ THz) (green circle in **Fig. 6(c)**) by Gaussian frequency filtering and Fourier back-transform. The corresponding results for the red, blue, and black circles at frequency vectors ($\nu_\tau,\nu_t=22$ THz) have been presented and discussed in **Fig. 7** of Somma *et al.* (2016b), whereas data for ($\nu_\tau,\nu_t=10$ THz) have been discussed in detail inSomma *et al.* (2016a).

The orange dashed lines in **Fig. 7(a)** and **(b)** indicate the center of pulse A (waiting time $T=-35$ fs) with wavefronts parallel to the $t=\tau$ diagonal. In contrast, the nonlinear signal shows tilted phase fronts perpendicular to the frequency vector ($\nu_\tau=-44$ THz,$\nu_t=+22$ THz) (black dashed arrow in **Fig. 7(b)**). This signal is different from the $\chi^{(3)}$ signals typically observed in 2D spectroscopy (cf. **Fig. 1(b)** and **Table 1**). In the set of frequency vectors (Eq. (1)), $\vec{\nu}_A$ is the only one with a nonvanishing $\nu_\tau=-22$ THz component. Thus, generation of a nonlinear signal at $\vec{\nu}_S=(\nu_\tau=-44\text{ THz},\nu_t=+22\text{ THz})$ (dashed vector in **Fig. 7(b)** and **(d)**) requires at least two interactions of pulse A with the InSb sample. This fact excludes a $\chi^{(3)}$ scenario, because each pulse has interacted at least once with the sample, i.e., there is a minimum number of four interactions. On the other hand, the ν_t component of $\vec{\nu}_S$ is consistent with an odd total number of interactions only. Thus, the lowest order Liouville pathway corresponds at least to a $\chi^{(5)}$ susceptibility. The fact that the THz pulses are off-resonant to any dipole-allowed transition from the ground state of the sample and the observation that the electric field emitted by the nonlinear polarization is extremely short as a function of t (**Fig. 7(a)** and **(b)**), point to a scenario in which the respective last pulse in the interaction sequence interacts with a very broad band excited absorption of the crystal.

A $\chi^{(5)}$ scenario compatible with all experimental observations is represented by the Liouville space pathway shown in **Fig. 7(d)** and **(e)**. Two interactions with pulse A create a two-photon interband coherence in the crystal which evolves freely along the coherence time $t_A-t_B=-\tau$. Pulse B which overlaps in time with pulse A ($\tau=0$), completes the two-photon absorption event by two more interactions (Pidgeon *et al.*, 1979; Miller *et al.*, 1979). Finally, a single interaction with pulse C probes the broadband free carrier absorption in the conduction band of InSb. The time evolution of the two-photon interband coherence is plotted as a

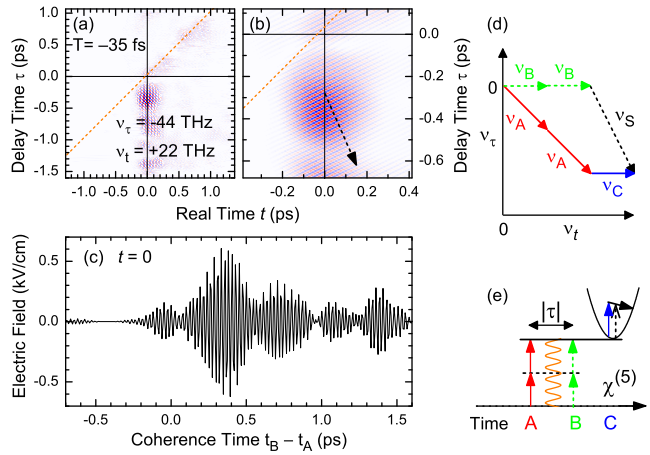

Fig. 7 (a), (b) Nonlinear signal for $T=-35$ fs at the frequency vector ($\nu_\tau=-44$ THz,$\nu_t=22$ THz) (green circle in **Fig. 6(c)**) as a function of the coherence time $t_A-t_B=-\tau$ and real time t. Panel (b) shows an enlarged view for small τ and t values. The tilted phase fronts are perpendicular to the direction of the frequency vector (black dashed arrow). (c) The cut along the ($t=0$) line of panel (a) shows directly the transient of the 'dark' two-photon interband coherence as a function of the coherence time $-\tau$. (d) Corresponding Liouville pathway diagram exploiting frequency vectors in 2D frequency space (cf. **Fig. 1(e)**). (d) Corresponding Liouville pathway diagram in the style of Pollard *et al.* (1992).

function of the coherence time $-\tau$ in Fig. 7(c). The fifth-order nonlinear signal shows quite a large amplitude (500 V/cm) and decays on a surprisingly long time scale. So far, there is no theoretical analysis of the microscopic dephasing mechanisms of two-photon coherences in semiconductors. One may speculate that reduced carrier-carrier scattering rates result in a surprisingly long decoherence time of the two-photon interband polarization. Four-wave-mixing experiments on semiconductor nanostructures (Kim *et al.*, 1992) have shown that electron-electron scattering is inhibited at the Fermi edge of the electron distribution, connected with the so-called Fermi edge singularity.

Conclusions

Multidimensional THz spectroscopy, a novel field of ultrafast science, has made rapid progress in recent years. The overview presented here combines an account of technological developments with selected prototype results for solids and nanostructures. Three-pulse 2D spectroscopy has been implemented in a collinear propagation geometry of three THz pulses and with fully phase-resolved detection of electric fields by electrooptic sampling. A key concept for analyzing 2D THz spectra is the introduction of frequency vectors in the multi-dimensional frequency space, which allow for identifying the dominating quantum pathways in Liouville space. Typical 2D experiments in the THz frequency range involve both off-resonant and multiple (i.e., beyond $\chi^{(3)}$) interaction sequences entering easily the nonperturbative regime of light–matter interaction. Two prototype experiments on intersubband coherences in quantum wells and the 'dark' two-photon coherence in the narrow gap semiconductor InSb illustrate the potential of multidimensional THz methods.

The generation of complex multicolor pulse sequences represents an important topic of current and future work. Another key issue is the generation of THz pulses of higher energy to study comparably weak dipole transitions such as intermolecular vibrations, which play a key role in the structural dynamics of liquids and other disordered systems. Such developments will broaden the range of nonlinear interactions to be studied and establish 2D THz techniques as a versatile tool, approaching the variability of multidimensional NMR methods.

See also: Coherent Terahertz Sources. Terahertz Detectors

References

Asplund, M.C., Zanni, M.T., Hochstrasser, R.M., 2000. Two-dimensional infrared spectroscopy of peptides by phase-controlled femtosecond vibrational photon echoes. Proc. Natl. Acad. Sci. USA 97, 8219–8224. doi:10.1073/pnas.140227997.

Autler, S.H., Townes, C.H., 1955. Stark effect in rapidly varying fields. Phys. Rev. 100, 703–722. doi:10.1103/PhysRev.100.703.

Bonch-Bruevich, A.M., Khodovoi, V.A., 1967. Current methods for the study of the Stark effect in atoms. Usp. Fiz. Nauk 93, 71–110. doi:10.1070/PU1968v010n05ABEH005850.

Bowlan, P., Martinez-Moreno, E., Reimann, K., Elsaesser, T., Woerner, M., 2014a. Ultrafast terahertz response of multi-layer graphene in the nonperturbative regime. Phys. Rev. B 89, 041408(R). doi:10.1103/PhysRevB.89.041408.

Bowlan, P., Martinez-Moreno, E., Reimann, K., Elsaesser, T., Woerner, M., 2014b. Terahertz radiative coupling and damping in multilayer graphene. New J. Phys. 16, 013027. doi:10.1088/1367-2630/16/1/013027.

Boyd, R.W., Mukamel, S., 1984. Origin of spectral holes in pump–probe studies of homogeneously broadened lines. Phys. Rev. A 29, 1973–1983. doi:10.1103/PhysRevA.29.1973.

Brodschelm, A., Tauser, F., Huber, R., Sohn, J.Y., Leitenstorfer, A., 2001. Amplitude and phase resolved detection of tunable femtosecond pulses with frequency components beyond 100 THz. In: Elsaesser, T., Mukamel, S., Murnane, M.M., Scherer, N.F. (Eds.), Ultrafast Phenomena XII. Berlin: Springer, pp. 215–217.

Chachisvilis, M., Fidder, H., Sundström, V., 1995. Electronic coherence in pseudo two-colour pump–probe spectroscopy. Chem. Phys. Lett. 234, 141–150. doi:10.1016/0009-2614(95)00041-2.

Dai, X., Richter, M., Li, H., *et al.*, 2012. Two-dimensional double-quantum spectra reveal collective resonances in an atomic vapor. Phys. Rev. Lett. 108, 193201. doi:10.1103/PhysRevLett.108.193201.

Delteil, A., Vasanelli, A., Todorov, Y., *et al.*, 2012. Charge-induced coherence between intersubband plasmons in a quantum structure. Phys. Rev. Lett. 109, 246808. doi:10.1103/PhysRevLett.109.246808.

Dietze, D., Darmo, J., Unterrainer, K., 2012. THz-driven nonlinear intersubband dynamics in quantum wells. Opt. Express 20, 23053. doi:10.1364/OE.20.023053.

Ernst, R.R., Bodenhausen, G., Wokaun, A., 1997. Principles of Nuclear Magnetic Resonance in One and Two Dimensions. Oxford: Clarendon Press.

Folpini, G., Morrill, D., Somma, C., *et al.*, 2015. Nonresonant coherent control-intersubband excitations manipulated by a nonresonant terahertz pulse. Phys. Rev. B 92, 085306. doi:10.1103/PhysRevB.92.085306.

Gaal, P., Kuehn, W., Reimann, K., *et al.*, 2007. Internal motions of a quasiparticle governing its ultrafast nonlinear response. Nature 450, 1210–1213. doi:10.1038/nature06399.

Gaal, P., Reimann, K., Woerner, M., *et al.*, 2006. Nonlinear terahertz response of *n*-type GaAs. Phys. Rev. Lett. 96, 187402. doi:10.1103/PhysRevLett.96.187402.

Gao, F., Cundiff, S.T., Li, H., 2016. Probing dipole–dipole interaction in a rubidium gas via double-quantum 2D spectroscopy. Opt. Lett. 41, 2954–2957. doi:10.1364/OL.41.002954.

Günter, G., Anappara, A.A., Hees, J., *et al.*, 2009. Sub-cycle switch-on of ultrastrong light–matter interaction. Nature 458, 178–181. doi:10.1038/nature07838.

Hamm, P., Lim, M., Hochstrasser, R.M., 1998. Structure of the amide I band of peptides measured by femtosecond nonlinear-infrared spectroscopy. J. Phys. Chem. B 102, 6123–6138. doi:10.1021/jp9813286.

Hamm, P., Zanni, M., 2011. Concepts and Methods of 2D Infrared Spectroscopy. Cambridge: Cambridge University Press.

Helm, M., 2000. The basic physics of intersubband transitions. In: Liu, H.C., Capasso, F. (Eds.), Intersubband Transitions in Quantum Wells: Physics and Device Applications I. Academic Press, pp. 1–100.

Huber, R., Brodschelm, A., Tauser, F., Leitenstorfer, A., 2000. Generation and field-resolved detection of femtosecond electromagnetic pulses tunable up to 41 THz. Appl. Phys. Lett. 76, 3191–3193. doi:10.1063/1.126625.

Huber, R., Kübler, C., Tübel, S., *et al.*, 2005. Femtosecond formation of coupled phonon-plasmon modes in InP: ultrabroadband THz experiment and quantum kinetic theory. Phys. Rev. Lett. 94, 027401. doi:10.1103/PhysRevLett.94.027401.

Huber, R., Tauser, F., Brodschelm, A., *et al.*, 2001. How many-particle interactions develop after ultrafast excitation of an electron-hole plasma. Nature 414, 286–289. doi:10.1038/35104522.

Hwang, H.Y., Fleischer, S., Brandt, N.C., *et al.*, 2015. A review of non-linear terahertz spectroscopy with ultrashort tabletop-laser pulses. J. Mod. Opt. 62, 1447–1479. doi:10.1080/09500340.2014.918200.

Jonas, D.M., 2003. Two-dimensional femtosecond spectroscopy. Annu. Rev. Phys. Chem. 54, 425–463. doi:10.1146/annurev.physchem.54.011002.103907.

Junginger, F., Mayer, B., Schmidt, C., *et al.*, 2012. Nonperturbative interband response of a bulk InSb semiconductor driven off resonantly by terahertz electromagnetic few-cycle pulses. Phys. Rev. Lett. 109, 147403. doi:10.1103/PhysRevLett.109.147403.

Junginger, F., Sell, A., Schubert, O., *et al.*, 2010. Single-cycle multiterahertz transients with peak fields above 10 MV/cm. Opt. Lett. 35, 2645–2647. doi:10.1364/OL.35.002645.

Kampfrath, T., Sell, A., Klatt, G., *et al.*, 2011. Coherent terahertz control of antiferromagnetic spin waves. Nature Photon. 5, 31–34. doi:10.1038/nphoton.2010.259.

Kane, E.O., 1957. Band structure of indium antimonide. J. Phys. Chem. Solids 1, 249–261. doi:10.1016/0022-3697(57)90013-6.

Kane, E.O., 1959. Zener tunneling in semiconductors. J. Phys. Chem. Solids 12, 181–188. doi:10.1016/0022-3697(60)90035-4.

Khalil, M., Demirdöven, N., Tokmakoff, A., 2003. Coherent 2D IR spectroscopy: molecular structure and dynamics in solution. J. Phys. Chem. A 107, 5258–5279. doi:10.1021/jp0219247.

Kim, D.-S., Shah, J., Cunningham, J.E., *et al.*, 1992. Carrier-carrier scattering in a degenerate electron system: strong inhibition of scattering near the Fermi edge. Phys. Rev. Lett. 68, 2838–2841. doi:10.1103/PhysRevLett.68.2838.

Kübler, C., Huber, R., Leitenstorfer, A., 2005. Ultrabroadband terahertz pulses: generation and field-resolved detection. Semicond. Sci. Technol. 20, S128–S133. doi:10.1088/0268-1242/20/7/002.

Kübler, C., Huber, R., Tübel, S., Leitenstorfer, A., 2004. Ultrabroadband detection of multi-terahertz field transients with GaSe electro-optic sensors: approaching the near infrared. Appl. Phys. Lett. 85, 3360–3362. doi:10.1063/1.1808232.

Kuehn, W., Reimann, K., Woerner, M., Elsaesser, T., 2009. Phase-resolved two-dimensional spectroscopy based on collinear *n*-wave mixing in the ultrafast time domain. J. Chem. Phys. 130, 164503. doi:10.1063/1.3120766.

Kuehn, W., Reimann, K., Woerner, M., Elsaesser, T., Hey, R., 2011a. Two-dimensional terahertz correlation spectra of electronic excitations in semiconductor quantum wells. J. Phys. Chem. B 115, 5448–5455. doi:10.1021/jp1099046.

Kuehn, W., Reimann, K., Woerner, M., *et al.*, 2011b. Strong correlation of electronic and lattice excitations in GaAs/AlGaAs semiconductor quantum wells revealed by two-dimensional terahertz spectroscopy. Phys. Rev. Lett. 107, 067401. doi:10.1103/PhysRevLett.107.067401.

Leitenstorfer, A., Hunsche, S., Shah, J., Nuss, M.C., Knox, W.H., 1999. Detectors and sources for ultrabroadband electro-optic sampling: experiment and theory. Appl. Phys. Lett. 74, 1516–1518. doi:10.1063/1.123601.

Luo, C.W., Reimann, K., Woerner, M., *et al.*, 2004. Phase-resolved nonlinear response of a two-dimensional electron gas under femtosecond intersubband excitation. Phys. Rev. Lett. 92, 047402. doi:10.1103/PhysRevLett.92.047402.

Miller, A., Johnston, A., Dempsey, J., *et al.*, 1979. Two-photon absorption in InSb and $Hg_{1-x}Cd_xTe$. J. Phys. C 12, 4839–4849. doi:10.1088/0022-3719/12/22/025.

Mukamel, S., 1995. Principles of Nonlinear Optical Spectroscopy. New York: Oxford University Press.

Mukamel, S., 2000. Multidimensional femtosecond correlation spectroscopies of electronic and vibrational excitations. Annu. Rev. Phys. Chem. 51, 691–729. doi:10.1146/annurev.physchem.51.1.691.

Pashkin, A., Kübler, C., Ehrke, H., *et al.*, 2011. Ultrafast insulator-metal phase transition in VO_2 studied by multiterahertz spectroscopy. Phys. Rev. B 83, 195120. doi:10.1103/PhysRevB.83.195120.

Pidgeon, C.R., Wherrett, B.S., Johnston, A.M., Dempsey, J., Miller, A., 1979. Two-photon absorption in zinc-blende semiconductors. Phys. Rev. Lett. 42, 1785–1788. doi:10.1103/PhysRevLett.42.1785.

Pollard, W.T., Dexheimer, S.L., Wang, Q., *et al.*, 1992. Theory of dynamic absorption spectroscopy of nonstationary states. 4. Application to 12-fs resonant impulsive Raman spectroscopy of bacteriorhodopsin. J. Phys. Chem. 96, 6147–6158. doi:10.1021/j100194a013.

Reimann, K., Smith, R.P., Weiner, A.M., Elsaesser, T., Woerner, M., 2003. Direct field-resolved detection of terahertz transients with amplitudes of megavolts per centimeter. Opt. Lett. 28, 471–473. doi:10.1364/OL.28.000471.

Schubert, O., Riek, C., Junginger, F., *et al.*, 2011. Ultrashort pulse characterization with a terahertz streak camera. Opt. Lett. 36, 4458–4460. doi:10.1364/OL.36.004458.

Sell, A., Leitenstorfer, A., Huber, R., 2008. Phase-locked generation and field-resolved detection of widely tunable terahertz pulses with amplitudes exceeding 100 MV/cm. Opt. Lett. 33, 2767–2769. doi:10.1364/OL.33.002767.

Shih, T., Reimann, K., Woerner, M., *et al.*, 2005. Nonlinear response of radiatively coupled intersubband transitions of quasi–two-dimensional electrons. Phys. Rev. B 72, 195338. doi:10.1103/PhysRevB.72.195338.

Somma, C., Folpini, G., Gupta, J., *et al.*, 2015. Ultra-broadband terahertz pulses generated in the organic crystal DSTMS. Opt. Lett. 40, 3404–3407. doi:10.1364/OL.40.003404.

Somma, C., Folpini, G., Reimann, K., Woerner, M., Elsaesser, T., 2016a. Two-phonon quantum coherences in indium antimonide studied by nonlinear two-dimensional terahertz spectroscopy. Phys. Rev. Lett. 116, 177401. doi:10.1103/PhysRevLett.116.177401.

Somma, C., Folpini, G., Reimann, K., Woerner, M., Elsaesser, T., 2016b. Phase-resolved two-dimensional terahertz spectroscopy including off-resonant interactions beyond the $\chi^{(3)}$ limit. J. Chem. Phys. 144, 184202. doi:10.1063/1.4948639.

Somma, C., Reimann, K., Flytzanis, C., Elsaesser, T., Woerner, M., 2014. High-field terahertz bulk photovoltaic effect in lithium niobate. Phys. Rev. Lett. 112, 146602. doi:10.1103/PhysRevLett.112.146602.

Stroucken, T., Knorr, A., Thomas, P., Koch, S.W., 1996. Coherent dynamics of radiatively coupled quantum-well excitons. Phys. Rev. B 53, 2026–2033. doi:10.1103/PhysRevB.53.2026.

Tonouchi, M., 2007. Cutting-edge terahertz technology. Nature Photon. 1, 97–105. doi:10.1038/nphoton.2007.3.

Wagner, M., Helm, M., Sherwin, M.S., Stehr, D., 2011. Coherent control of a THz intersubband polarization in a voltage controlled single quantum well. Appl. Phys. Lett. 99, 131109. doi:10.1063/1.3644988.

Wagner, M., Schneider, H., Stehr, D., *et al.*, 2010. Observation of the intraexciton Autler-Townes effect in GaAs/AlGaAs semiconductor quantum wells. Phys. Rev. Lett. 105, 167401. doi:10.1103/PhysRevLett.105.167401.

Woerner, M., Kuehn, W., Bowlan, P., Reimann, K., Elsaesser, T., 2013. Ultrafast two-dimensional terahertz spectroscopy of elementary excitations in solids. New J. Phys. 15, 025039. doi:10.1088/1367-2630/15/2/025039.

Wu, Q., Zhang, X.-C., 1995. Free-space electro-optic sampling of terahertz beams. Appl. Phys. Lett. 67, 3523–3525. doi:10.1063/1.114909.

Wu, Q., Zhang, X.-C., 1997. Free-space electro-optics sampling of mid-infrared pulses. Appl. Phys. Lett. 71, 1285–1286. doi:10.1063/1.119873.

Yang, L., Zhang, T., Bristow, A.D., Cundiff, S.T., Mukamel, S., 2008. Isolating excitonic Raman coherence in semiconductors using two-dimensional correlation spectroscopy. J. Chem. Phys. 129, 234711. doi:10.1063/1.3037217.

Yee, T.K., Gustafson, T.K., 1978. Diagrammatic analysis of the density operator for nonlinear optical calculations: pulsed and cw responses. Phys. Rev. A 18, 1597–1617. doi:10.1103/PhysRevA.18.1597.

Second Harmonic Generation Spectroscopy of Hidden Phases

Liuyan Zhao, University of Michigan, Ann Arbor, MI, United States
Darius Torchinsky, Temple University, Philadelphia, PA, United States
John Harter, University of California, Santa Barbara, CA, United States
Alberto de la Torre and David Hsieh, California Institute of Technology, Pasadena, CA, United States

Introduction

A cornerstone in condensed matter research is the emergence of the concept of symmetry and symmetry breaking upon an order formation (Landau, 1936, 1937), an idea that an ordered state can only have a true subset of symmetries of its unordered counterpart. Being able to experimentally capture the evolution of symmetries across a phase transition of an order is essential for understanding its microscopic mechanism and its macroscopic mechanical, electronic and magnetic properties as well (Birss, 1964; Nye, 1985). Traditionally, X-ray (Warren, 1969), neutron (Squires, 2012) and electron (Zou *et al.*, 2011) diffraction techniques are widely used in determining the symmetries of lattice, magnetic and charge orders, while resonant X-ray diffraction has recently been exploit to probe symmetries of more exotic orders such as orbital (Murakami *et al.*, 1998) and multipolar orders (Kuramoto *et al.*, 2009; Santini *et al.*, 2009). However, the accuracy of symmetry space group assignment through the above diffraction techniques critically depends on the ability of performing a unique fit to the diffraction pattern. This is not always feasible because of technical challenges including the presence of elements with a large adsorption or weak cross-section, the limited crystal size smaller than the probe beam spot, the coexistence of multiple orders with different symmetries, or the small domain size of multiple degenerate states of an order.

SHG (Franken *et al.*, 1961), the lowest-order degenerate nonlinear optical generation, is an alternative to diffraction techniques for resolving the symmetries of a material through the structure of its second order nonlinear optical susceptibility tensor (Boyd, 2003; Shen, 1984). Of particular interest is its extreme sensitivity to inversion symmetry because the leading order electric dipole (ED) contribution to SHG is only allowed wherever inversion symmetry is broken such as at interfaces of heterostructures of centrosymmetric crystals, i.e., the optical susceptibility tensor for this ED SHG process is zero if inversion symmetry is present. To experimentally determine the structure of the SHG susceptibility tensor for a crystal, a SHG rotation anisotropy (RA) measurement is typical performed, in which the intensity of the SHG light generated from the crystal is recorded as the crystal rotates about its surface normal.

The application of SHG in determining symmetry point group dates back to 1980s for studying lattice reconstructions on semiconductor surfaces (Heinz *et al.*, 1985; Tom *et al.*, 1983; Yamada and Kimura, 1993) and metal-electrolyte interfaces (Shannon *et al.*, 1987a,b; Shen, 1989), extends later on for probing magnetic ordering on metal surfaces (Dähn *et al.*, 1996; Gridnev *et al.*, 2001; Kirilyuk and Rasing, 2005; Nývlt *et al.*, 2008; Pan *et al.*, 1989; Reif *et al.*, 1991), in bulk semiconductors (Lafrentz *et al.*, 2012) and multiferroics (Fiebig *et al.*, 1994, 2000, 2005; Kumar *et al.*, 2008), and recently to characterizing two-dimensional atomic crystals (Clark *et al.*, 2014; Janisch *et al.*, 2014; Kim and Lim, 2015; Kumar *et al.*, 2013; Li *et al.*, 2013; Malard *et al.*, 2013; Seyler *et al.*, 2015; Wu *et al.*, 2014; Yin *et al.*, 2014) and topological materials (Hamh *et al.*, 2016; Hsieh *et al.*, 2011; McIver *et al.*, 2012; Wu *et al.*, 2016). However, it is only after overcoming a number of technical challenges very recently that this technique has been introduced as an effective probe for hidden phases in correlated electron systems. The most outstanding one is the need for sensitivity to the entire SHG susceptibility tensor, which requires performing the SHG RA measurements in the oblique incidence reflection geometry. This, under the traditional rotating-sample-based scheme, is achievable at the room temperature despite a number of alignment difficulties, but was almost impossible at cryogenic temperatures where electronic phase transitions usually happen. A novel SHG RA design based on a rotation of the light scattering-plane instead of the sample was first reported in 2014 (Harter *et al.*, 2015; Torchinsky *et al.*, 2014), and has been successfully applied in a few systems to determine their lattice (Harter *et al.*, 2016; Hogan *et al.*, 2016; Torchinsky *et al.*, 2015) and electronic symmetries (Zhao *et al, V1.36.*, 2016a,b; Harter *et al.*, 2017). Moreover, by keeping the sample stationary during the RA measurements, it allows to perform experiments not only at cryogenic temperatures, but also under static magnetic or strain fields.

This review is organized as follows. In Section "Second Harmonic Generation in Crystals", we introduce the theoretical background of SHG and its relationship to the lattice and electronic symmetries of a crystal. In Section "Generations of Rotational Anisotropy Experimental Designs", we describe the advances in the SHG RA designs with an emphasis on their capabilities and limitations. In Section "Hidden Phases in Correlated Electron Systems", we discuss hidden phases in correlated electron systems and the challenges for their experimental detections. In Section "Hidden Phases Revealed by SHG-based Techniques", we take examples of SHG-based techniques revealing hidden phases. Finally, in Section "Conclusions and Prospects", we discuss the prospects of generalizing SHG to non-degenerate second order nonlinear optical generations in detecting novel electronic orders.

Second Harmonic Generation in Crystals

Electromagnetic waves of the incident light propagating through a medium can induce electric dipole $\left(\vec{P}\right)$, magnetic dipole $\left(\vec{M}\right)$, electric quadrupole $\left(\overleftrightarrow{Q}\right)$, or even higher rank multipolar densities, all of which together compose the source term ($\vec{S}$; note that

this is not the Poynting vector) for the radiated light (Boyd, 2003; Fiebig *et al.*, 2005).

$$\vec{S} = \mu_0 \frac{\partial^2 \vec{P}}{\partial t^2} + \mu_0 \left(\nabla \times \frac{\partial \vec{M}}{\partial t} \right) - \mu_0 \left(\nabla \frac{\partial^2 \overleftrightarrow{Q}}{\partial t^2} \right) + \cdots \tag{1}$$

The electric dipole term ($\propto \vec{P}$) is the leading order contribution to $\vec{S}$, and is typically $\sim \lambda / a$ times stronger than the second order magnetic dipole ($\propto \vec{M}$) and electric quadrupole ($\propto \overleftrightarrow{Q}$) terms, where λ and a are the wavelength of the light and the lattice constant of the crystal respectively. This is precisely the reason why linear optical processes can be well described by the electric dipole approximation.

Generally, the induced dipolar/multipolar densities can be expressed as linear, second and higher order nonlinear expansions of electric $\vec{E}(\omega)$ and magnetic $\vec{H}(\omega)$ fields of the incident light.

$$\vec{P}(\omega, 2\omega, \ldots) \propto \chi^{pe} \vec{E}(\omega) + \chi^{pm} \vec{H}(\omega) + \chi^{pee} \vec{E}(\omega) \vec{E}(\omega) + \chi^{pem} \vec{E}(\omega) \vec{H}(\omega) + \chi^{pmm} \vec{H}(\omega) \vec{H}(\omega) + \mathcal{O}\left[\left(\vec{E}, \vec{H} \right)^3 \right] \tag{2}$$

$$\vec{M}(\omega, 2\omega, \ldots) \propto \chi^{me} \vec{E}(\omega) + \chi^{mm} \vec{H}(\omega) + \chi^{mee} \vec{E}(\omega) \vec{E}(\omega) + \chi^{mem} \vec{E}(\omega) \vec{H}(\omega) + \chi^{mmm} \vec{H}(\omega) \vec{H}(\omega) + \mathcal{O}\left[\left(\vec{E}, \vec{H} \right)^3 \right] \tag{3}$$

$$\overleftrightarrow{Q}(\omega, 2\omega, \ldots) \propto \chi^{qe} \vec{E}(\omega) + \chi^{qm} \vec{H}(\omega) + \chi^{qee} \vec{E}(\omega) \vec{E}(\omega) + \chi^{qem} \vec{E}(\omega) \vec{H}(\omega) + \chi^{qmm} \vec{H}(\omega) \vec{H}(\omega) + \mathcal{O}\left[\left(\vec{E}, \vec{H} \right)^3 \right] \tag{4}$$

The coefficients (χ) in Eqs. (2–4), i.e. optical susceptibility tensors, describe the optical processes happening in the crystal upon the driving fields, and reflects the physical properties of the crystal. For example, $\chi^{pee}_{ijk}(2\omega)$ characterizes the electric dipole SHG response in the crystal via two consecutive interactions with the incident electric field, where the first superscript denotes the electric dipole origin (p) of the induced source, the second and third superscripts denote the electric (e) nature of the driving field, and the subscripts denote the polarization components.

Neumann's principle dictates that property tensors, such as χ above, should remain invariant under transformations that respect the symmetries of the crystal (Birss, 1964). That is

$$\chi_{ijk\ldots} = R_{ii'} R_{jj'} R_{kk'} \ldots \chi_{i'j'k'\ldots} \tag{5}$$

where $R_{ii'}$, $R_{jj'}$, $R_{kk'}$ are matrices for the symmetry transformations, and $\chi_{i'j'k'\ldots}$ and $\chi_{ijk\ldots}$ are the property tensors before and after the transformation respectively. This enforces a set of relationships among the tensor elements, and therefore significantly reduces the number of independent non-zero tensor elements. As a result, the symmetry properties of the crystal are encoded in the structure of the property tensors, and we can in principle determine the lattice and electronic symmetries of the crystal via measuring the property tensors. Linear responses with the lowest rank property tensors have limited symmetry sensitivity due to their degenerate structure for the symmetry point groups within the same crystal system, while nonlinear responses with higher rank tensors gain greater symmetry resolutions through different tensor forms for different point groups. The second order nonlinear optical responses, including SHG, sum frequency generation (SFG) and difference frequency generation (DFG), stand out in particular because of their extreme sensitivity to inversion symmetry, i.e., their leading order electric dipole contribution $\left(\chi^{pee}_{ijk}\right)$ is only allowed wherever inversion symmetry is broken such as surfaces and interfaces of centrosymmetric crystals. However, here we draw your attention to that the higher order multipolar contributions, such as magnetic dipole $\left(\chi^{mee}_{ijk}\right)$ and electric quadrupole $\left(\chi^{qee}_{ijk}\right)$, are present even under inversion symmetry, despite their much lower radiation efficiency. In this review, we will focus on SHG, the degenerate second order nonlinear optical response, and its recent applications in revealing hidden electronic phases in correlated condensed matter systems.

Quantum mechanical description of optical SHG susceptibility is expressed via equations, for example, in the electric dipole approximation (Boyd, 2003; Zhao *et al.*, 2016a),

$$\chi^{pee}_{ijk}(2\omega) \propto \sum_{\vec{k}} \sum_{g,m,f} \left[\frac{\langle g, \vec{k} | P_i | f, \vec{k} \rangle \langle f, \vec{k} | P_j | m, \vec{k} \rangle \langle m, \vec{k} | P_k | g, \vec{k} \rangle}{\left(\hbar\omega_{fg} - 2\hbar\omega - i\gamma_{fg} \right) \left(\hbar\omega_{mg} - \hbar\omega - i\gamma_{mg} \right)} + \cdots \right] f_g(\vec{k}) \tag{6}$$

which involves a two-photon adsorption process driven by two consecutive electric dipole transitions at an energy of $\hbar\omega$ from the initial state $|g, \vec{k}\rangle$ to the intermediate state $|m, \vec{k}\rangle$ and from the intermediate state $|m, \vec{k}\rangle$ to the final state $|f, \vec{k}\rangle$ respectively, followed by a single-photon emission process driven by one electric dipole transition at an energy of $2\hbar\omega$ from $|f, \vec{k}\rangle$ back to $|g, \vec{k}\rangle$. The energy difference between the initial state and the intermediate (final) state is given by $\hbar\omega_{mg}$ ($\hbar\omega_{fg}$), and the damping rate for the transition between the initial and the intermediate (final) state is represented by γ_{mg} (γ_{fg}) with $\gamma_{m(f)g} = \frac{1}{2}\left(\tau^{-1}_{m(f)} + \tau^{-1}_{g}\right)$ where τ_g and $\tau_{m(f)}$ are the lifetime for the initial and the intermediate (final) states respectively. $f_g(\vec{k})$ is the Fermi distribution function for the initial state $|g, \vec{k}\rangle$. Similar as linear optics, the dominant contributions to $\chi^{pee}_{ijk}(2\omega)$ are from the resonant optical transitions where $2\hbar\omega$ ($\hbar\omega$) matches the energy difference between the initial and final (intermediate) states, $\hbar\omega_{fg}(\hbar\omega_{mg})$, and therefore $\chi^{pee}_{ijk}(2\omega)$ responds most when the resonant transition happens at energies with most spectra-weight transfer across a phase transition. This statement is generally true not only for $\chi^{pee}_{ijk}(2\omega)$, but also for any single optical process in Eqs. (2–4). More than linear optics, SHG has a better ability of probing the symmetry properties of the relevant states. Furthermore, SHG can be sensitive to inversion symmetry breaking phases even if the photon energy is off resonance with the most responsive electronic

states because of the stronger radiation efficiency of the activated leading order electric dipole contribution ($\chi^{pee}_{ijk}(2\omega)$) from the inversion symmetry breaking order.

The current understanding of SHG as a symmetry probe of electronic states is at a qualitative level where we only discuss if the SHG susceptibility tensor elements are zero or not so as to infer what symmetry operations are present in the system. Compared to linear optics, we lack comprehensive understanding on the amplitudes of SHG susceptibility tensor elements which in fact contain key information of the nature and the strength of the order parameters. Thereby, theoretical inputs in SHG, or more generally nonlinear optical generations, from novel electronic states will be greatly appreciated. In the following parts of this review, we will mainly discuss the experimental aspects of SHG from hidden electronic phases.

Generations of Rotational Anisotropy Experimental Designs

In addition to its sensitivity to inversion symmetry, SHG is capable of detecting rotational symmetries, mirror symmetries and time reversal symmetry by performing a RA measurement to access as many SHG susceptibility tensor elements as possible.

A SHG RA measurement is to record the SHG intensity ($I^{2\omega}$) as a function of the azimuthal angle ((φ)) about the crystal surface normal shown in **Fig. 1**. The incident light at an optical frequency of ω and the reflected light at a frequency of 2ω form a scattering plane. The angle between the scattering plane and the crystal axis a is defined as the azimuthal angle (φ). The polarizations of the incident and reflected light can be independently selected to be either within ($P_{in/out}$) or normal ($S_{in/out}$) to the scattering plane. In total, it gives four independent polarization geometries ($P/S_{in} - P/S_{out}$) that together access all the SHG susceptibility tensor elements and each select different combinations of the tensor elements.

Traditionally, the full RA measurement is obtained via rotating the crystal about its surface normal (**Fig. 1**), and has been successfully applied in studies of molecule self-assembly on crystal surfaces, semiconductors and transition metal oxide interfaces. However, this scheme has a number of alignment challenges including beam walking on the sample and precession of the reflected beam during the RA acquirement, and therefore it not only requires large crystal sizes, but also causes aberrations in the RA patterns. More importantly, the rotation of the crystal prohibits the measurements taken under various experimental conditions such as at cryogenic temperature, in magnetic fields, under high pressure and so on. In strongly correlated electron systems, temperature, magnetic field and pressure are precisely the experimental parameters that we tune to realize various electronic phases of matter.

To overcome the limitations on experimental conditions posed by the rotation of the crystal, a different design was introduced to keep the sample stationary during the RA measurements. As shown in **Fig. 2(a)**, instead of oblique incidence in **Fig. 1**, it takes the normal incidence geometry. In this design, the polarizations of the incoming and the reflected/transmitted light can be selected to be either parallel or orthogonal to each other, and the RA data is taken via rotating the polarizations relative to the crystal axis a. As we see in **Fig. 2(a)**, both polarizations are parallel to the crystal surface (xy plane in the lab coordinate), missing the component in the direction of the surface normal (z direction). According to the relationships in Eqs. (2–4), such as $P_i = \chi^{pee}_{ijk} E_j E_k$, this scheme lacks the accessibility to the SHG susceptibility tensor elements whose subscripts includes z, e.g., χ^{pee}_{ijk} elements with $i=z$ or $j=z$ or $k=z$. For the high symmetry systems where the inaccessible SHG susceptibility tensor elements are zero, this scheme works successfully, and we will see one such example on $Cd_2Re_2O_7$ (Petersen *et al.*, 2006) in Section "A Parity-broken Nematic Order in $Cd_2Re_2O_7$". However, it faces limitations in the systems with lower symmetry where the inaccessible tensor elements are necessary for differentiating between symmetry point groups.

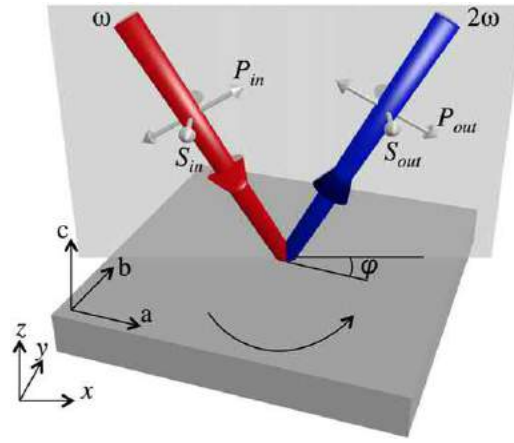

Fig. 1 Schematic of the rotating-sample-based SHG RA setup. The electric field polarizations of the obliquely incident fundamental beam (in) and outgoing SHG beam (out) can be independently selected to be either parallel (P) or perpendicular (S) to the light scattering plane (shaded vertical plane). SHG RA data are acquired by measuring the SHG intensity ($I^{2\omega}$) as a function of the angle (φ) between the fixed light scattering plane and the *ac* plane of the rotating sample. xyz is the lab coordinate and abc is the sample coordinate.

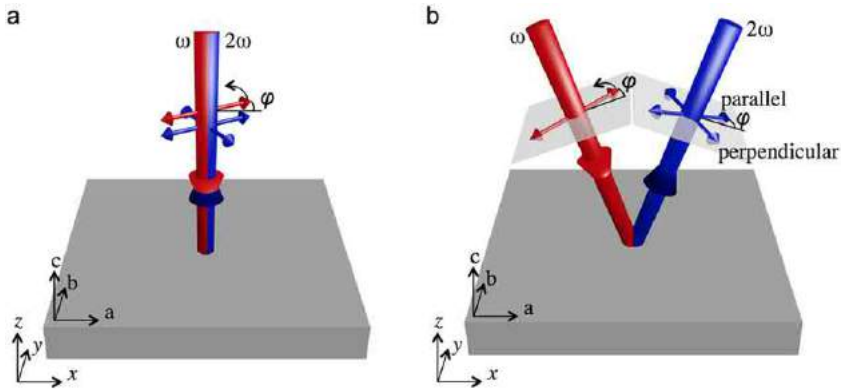

Fig. 2 Schematic of the rotating-polarization-based SHG RA setups. (a) In the normal incidence geometry, the polarizations of the incoming (ω) and outgoing (2ω) beams can be selected to be either parallel or perpendicular to each other. The SHG RA data is to record the SHG intensity $I^{2\omega}$ as a function of the angle (φ) between the a-axis and the polarization of the incoming light. (b) A derivative of (a) to the oblique incidence geometry. The polarizations of the incoming (ω) and outgoing (2ω) beams can be selected to be parallel, same angle (φ) between the polarization and the intersect of ac-plane and the wave-plane (shaded gray planes), or perpendicular, angle (φ) for incoming and (φ) + 90° for outgoing.

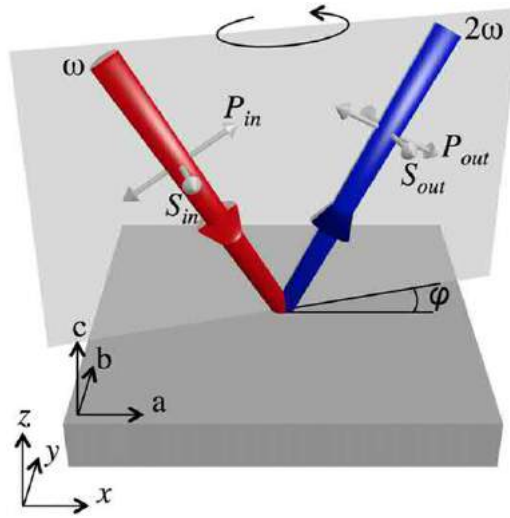

Fig. 3 Schematic of the rotating-scattering-plane-based SHG RA setups. This is equivalent to the one in **Fig. 1** but with a rotating scattering-plane instead of a rotating sample.

A derivative of this scheme was therefore developed, as shown in **Fig. 2(b)**. It keeps the idea of rotating the polarizations of both incident and reflected light, but deliberately chooses the oblique incidence to have polarization component along z direction. However, it still faces a couple of limitations. First, different from the last two setups in **Figs. 1** and 2(a), the measurement process from this design itself breaks rotational symmetries, and consequently the RA data would seemingly obscure the rotational symmetries of the crystal. Second, there are only two independent polarization channels, i.e., parallel- or cross-polarized between incoming and reflected/transmitted beams, in this scheme, as compared to four in the **Fig. 1** scheme. This limits the ability of accessing specific individual SHG susceptibility tensor elements , and therefore sometimes can overlook signatures of emergent phases. We will discuss this later in a concrete example inSection "An Odd-Parity Hidden Order in Sr_2IrO_4".

Recently, a rotating-scattering-plane-based RA setup design has been proposed and constructed (Torchinsky *et al.*, 2014). In the design shown in **Fig. 3**, instead of rotating crystals or polarizations, it rotates the scattering-plane while maintaining the polarizations locked with respect to the scattering plane. By doing so, it is in fact equivalent to the rotating-sample-based scheme in **Fig. 1**, but unifies the advantages of all three designs above. (1) By keeping the sample stationary in the current geometry, various sample environment conditions such as cooling the crystal to a cryogenic temperature, and directing a magnetic or strain field along a particular crystallographic direction are compatible with the SHG RA measurements, which empowers this technique to study the vast amount of emergent electronic phases in correlated materials. (2) The oblique incidence enables the accessibility to all susceptibility tensor elements, which is necessary to the sensitivity of differentiating symmetry point groups. In addition, the reflection geometry is especially suitable for the bulk single crystals, the common form for correlated electron materials, because

their thickness typically exceeds the penetration depth of light at the inter-band resonance frequencies. (3) The full rotation of 360° of the light scattering plane during the RA acquirement preserves all rotational symmetries, and therefore the RA patterns directly visualize the rotational symmetries in the crystal about its surface normal axis. Furthermore, the total four polarization geometries relative to the scattering plane maximize the number of independent SHG RA measurements of the susceptibility tensor for one crystallographic facet. (4) The design of 4*f* optical system, described in details in the following paragraph, minimizes misalignments including beam walking on the sample and precession of reflected beam. As a result, not only SHG RA experiments on small size single crystals (<1 *mm*) or spatially inhomogeneous samples can be easily performed, but also aberrations in RA SHG patterns are significantly reduced.

In practice, the rotation of the light scattering plane is realized by rotating a customized fused silica phase mask (PM) which diffracts the normal incident beam into $+1$ and -1 orders at an angle α relative to the optical axis. So far, two generations of RA designs have been developed. In the first generation shown in **Fig. 4**, the collimated incoming beam of a pulsed laser at a given photon energy first passes through a polarizer (P) to determine the incoming power and a half-wave plate (WP) to choose the incoming polarization with respect to the scattering plane (S_{in} for perpendicular and P_{in} for parallel), then gets focused by a lens (L1) onto the PM to produce $+/-1$ orders forming the light scattering plane. The two diffracted orders get collimated simultaneously and parallel to each other after a second lens (L2). One order is then blocked by a beam block (B) while the other one is filtered through a long pass filter (LPF) to eliminate any parasitic higher harmonics. The collimated incident beam is finally focused on the sample by a Cassegrain reflective objective (RO). The fundamental and higher harmonic beams reflected from the sample exit the RO from its diametrically opposite side, and get picked up by a D-cut silver coated pick-off mirror (DM). The SHG beam with a defined polarization of S_{out} or P_{out} is then selected through a polarization analyzer (A) and a set of spectra-filters including two short pass filters (SPF) and one interference filter (IF), and eventually directed into a photo-multiplier tube (PMT). In this design, the set of PM, L2, RO and sample form a 4*f* system that overcomes the alignment challenges in the rotating-sample-based RA design. During the RA data acquisitions, PM together with WP steps at discrete angles to illuminate the investigated area on the sample at a series of azimuthal angle (φ), and the detection arm consisting of DM, A, SPF/IF and PMT mounted on a big rotation stage rotates in sync with PM and WP to record the induced SHG intensity $I^{2\omega}_{P/S_{in}-P/S_{out}}((\varphi))$ at a given polarization geometry $P/S_{in}-P/S_{out}$ and an angle (φ). The stepped fashion of this design requires a long time (in order of 10^3s) to complete the full 360° sweep of (φ) for acquiring RA data, which is disadvantageous especially when the laser characteristics (power, point, pulse duration, etc.) have long-term fluctuations.

In order to overcome the drawback of long acquisition time, a second generation of the rotating-scattering-plane-based RA setup was developed, aiming at much shorter rotation cycles (Harter *et al.*, 2015). Shown in **Fig. 5**, a circularly polarized beam, produced from a linearly polarized beam using a quarter-wave plate (QWP), first passes through a linear polarizer (LP1) to select a polarization, either S_{in} or P_{in}, and then gets focused onto a fused silica binary PM to diffract the beam from the optical axis. The $+1$ order of the diffracted beam is then brought collimated and parallel to the optical axis via a lens (L1), further transmits through two tilted dichroic mirrors (DM1 and DM2 that transmit the fundamental beam but reflects the higher harmonic parasitic beams), and finally gets focused on the sample via an achromatic lens (L2). The reflected fundamental and higher harmonic beams return on the diametrically opposite side of L2, upon which it gets re-collimated. The SHG signal gets picked out by two tilted dichroic mirrors (DM2 and DM3) and one set of spectra filters (SF) with a chosen polarization of S_{out} or P_{out} via a linear polarizer (LP2), and draws a circle with varying intensity $I^{2\omega}_{S/P_{in}-S/P_{out}}((\varphi))$ on a cooled electron-multiplying CCD camera as the scattering plane rotates. In this design, LP1, PM and LP2 are coupled to a common axle to eliminate any possible mechanical phase slip, and therefore can rotate co-axially at a high speed (4 Hz or even more), which is greater than the first generation by a factor of 10^3.

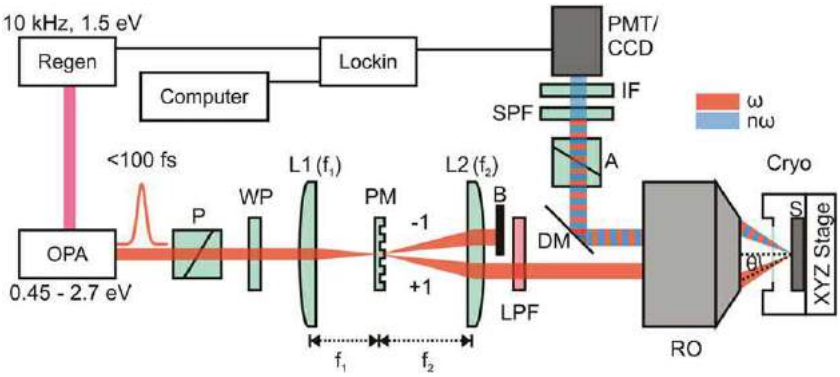

Fig. 4 Layout of the first generation rotating-scattering-plane-based SHG RA setup. A <100fs pulsed laser beam from an OPA, seeded by a Ti: sapphire regenerative amplifier, passes through a polarizer (P) and waveplate (WP), and is focused by the first lens (L1) on a phase mask (PM). PM, the second lens (L2), a reflective objective (RO) and a sample (S) form a 4f optical system. A first order ($+1$) diffracted beam gets focused on S through this 4f system while the other one (-1) is blocked. The reflected beam is collimated through RO, picked off by a D-cut mirror (DM), filtered through an analyzer (A), short-pass filters (SPF) and interference filters (IF), and finally detected by a photomultiplier tube (PMT) using lock-in detection or a CCD camera. Adapted from Torchinsky, D.H., *et al.*, 2014. A low temperature nonlinear optical rotational anisotropy spectrometer for the determination of crystallographic and electronic symmetries. Review of Scientific Instruments 85(8), 083102.

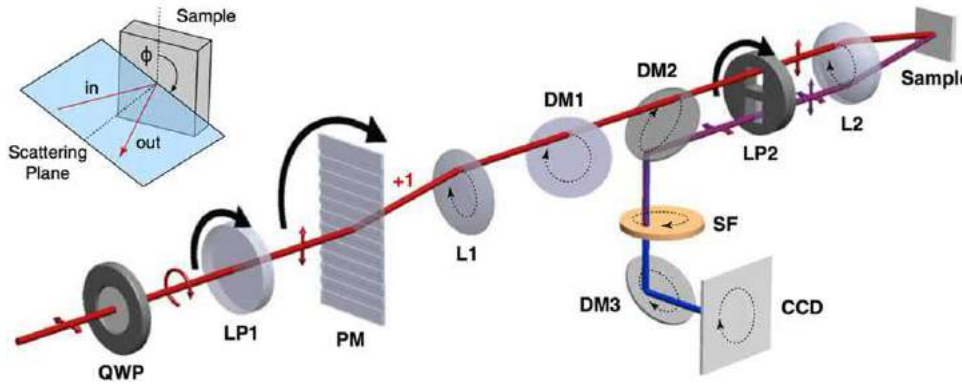

Fig. 5 Layout of the second generation rotating-scattering-plane-based SHG RA setup. Pulsed laser light passes through a series of optics including a quarter-wave plate (QWP), linear polarizers (LP1, 2), a phase mask (PM), collimating lens (L1), dichroic mirrors (DM1-3), achromatic objective lens (L2), spectral filter (SF), and CCD camera (CCD). LP2 is mounted on the detection side of a rotating wheel while the incoming side has a through-hole. The dashed arrows show the path traced by the beam on L1, L2, DM1-3, SF, and CCD during the RA acquisition. Top left inset shows the simplified geometry of a reflection-based SHG RA measurement. Adapted from Harter, J.W., *et al.*, 2015. High-speed measurement of rotational anisotropy nonlinear optical harmonic generation using position-sensitive detection. Optics Letters40(20), 4671–4674.

Using the rotating-scattering-plane-based RA technique, it has been shown successful in detecting hidden symmetry breaking orders in a pyrochlore $Cd_2Re_2O_7$ (Harter *et al.*, 2017), a perovskite iridate Sr_2IrO_4 (Zhao *et al.*, 2016b) and a cuprate $YBa_2Cu_3O_y$ (Zhao *et al.*, 2016a) as discussed in Sections "A Parity-broken Nematic Order in $Cd_2Re_2O_7$, An Odd-Parity Hidden Order in Sr_2IrO_4, and A Magnetic Quadrupole Order in the Pseudogap of $YBa_2Cu_3O_y$".

Hidden Phases in Correlated Electron Systems

In condensed matter systems, a hidden phase can either be a state of matter that cannot be reached in the thermodynamic equilibrated conditions (Huai and Nasu, 2002; Ichikawa *et al.*, 2011; Koshihara and Adachi, 2006; Nasu, 2004; Tokura, 2006; Zhang *et al.*, 2015), or be one in equilibrium that however cannot be easily accessed by conventional detection techniques (Fechner *et al.*, 2016; Fu, 2015; Kuramoto *et al.*, 2009; Santini *et al.*, 2009; Witczak-Krempa *et al.*, 2013). In this review, we take examples from the latter category and discuss their microscopic nature revealed using optical SHG based techniques.

In a correlated electron system, the interplay among electronic degrees of freedom often leads to a macroscopic realization of the system's ground state that involves spontaneous symmetry breaking and the emergence of an order parameter. Often, the microscopic nature of the ordered state can be investigated through direct external probes. For example, the electric dipole arrangement in ferroelectrics can be studied by applying an external electric field. However, there are occasions where the ordering element is not of such conventional scalar or vector form as charge density or electric/magnetic dipole, but instead is of the tensor form. Such ordering parameters of the tensor character couple with the usual scalar or vector probe fields so weakly that they can be invisible to these conventional probes. As the most well known example, the hidden order phase in the heavy fermion compound URu_2Si_2 has been theoretically proposed to be a multipolar (tensor) order (Mydosh and Oppeneer, 2011, 2014; Shah *et al.*, 2000), but so far both the microscopic nature of the order parameter and the macroscopic symmetries of its long range ordered state remain mysterious with few direct experimental results.

In classical electrodynamics, the concept of electric and magnetic multipolar expansions is introduced to account for the spatial distribution of charge or magnetization density (Jackson, 1999). Electric multipolar expansions include terms of total charge Q (or monopole term), dipole moment p_i, quadrupole moment q_{ij}, and even higher multipolar moments where

$$Q = \int d\vec{r'}\,\rho(\vec{r'}) \tag{7}$$

$$p_i = \int d\vec{r'}\,r_i'\rho(\vec{r'}) \tag{8}$$

$$q_{ij} = \int d\vec{r'}\,(3r_i'r_j' - r'^2\delta_{ij})\rho(\vec{r'}) \tag{9}$$

Similarly, magnetic multipolar expansions consist the first term of magnetic moment m_i with

$$m_i = \int d\vec{r'}\,\mu_i(\vec{r'}) \tag{10}$$

the second term μ_{ij} with

$$\mu_{ij} = \int d\vec{r'} r'_i \mu_j(\vec{r'}) \tag{11}$$

that can be decomposed into three irreducible groups:

(1) the pseudo-scalar from the trace of the tensor that contains a single component,

$$a = \frac{1}{3}\mu_{ii} = \frac{1}{3}\int d\vec{r'} r'_i \mu_i(\vec{r'}) \tag{12}$$

(2) the toroidal moment pseudo-vector that is the anti-symmetric part of the tensor and contains three independent components,

$$T_i = \frac{1}{2}\varepsilon_{ijk}\mu_{jk} = \frac{1}{2}\int d\vec{r'} \varepsilon_{ijk} r'_j \mu_k(\vec{r'}) \tag{13}$$

(3) the quadrupole moment tensor that is the traceless symmetric part of $\overleftrightarrow{\mu}$ and contains five independent components,

$$\tilde{q}_{ij} = \frac{1}{2}\left(\mu_{ij} + \mu_{ji} - \frac{2}{3}\delta_{ij}\mu_{kk}\right) = \frac{1}{2}\int d\vec{r'}\left[r'_i\mu_j\left(\vec{r'}\right) + r'_j\mu_i\left(\vec{r'}\right) - \frac{2}{3}\delta_{ij}\,\overrightarrow{r'\cdot\mu\left(\vec{r'}\right)}\right] \tag{14}$$

where the toroidal moment t_i and the quadrupole moment q_{ij} couple to the curl and the gradient of the magnetic field respectively, while the pseudo-scalar a is coupled to the divergent of the magnetic field and therefore represents a monopole moment; and even higher order terms.

In condensed matter systems, the electric and magnetic multipolar order parameters can be defined in a similar way as above (Fechner *et al.*, 2016), given the integration volume in Eqs. (7–14) is now replaced by the unit cell of a material. Taking the second term of magnetic multipolar expansions as an example (**Fig. 6**), the monopole, toroidal and quadrupole terms can be realized by the arrangement of localized magnetic moment within the unit cell. Each of the patterns in **Fig. 6** has its time reversal symmetry related counterpart, and therefore its long range ordered state naturally has at least two degenerate ground states. Different patterns have different sets of symmetries, and consequently their long range ordered states have different point groups. For example, $\tilde{q}_{z^2}$ pattern breaks all three mirror planes m_{xy}, m_{xz} and m_{yz}, but obeys the symmetry operations of each mirror operation followed by a time reversal symmetry operation m'_{xy}, m'_{xz} and m'_{yz} with the prime standing for the time reversal symmetry operation. In contrast, $\tilde{q}_{xz}$ pattern only breaks m_{xz} while remaining symmetric under m'_{xz}, m_{xy} and m_{yz} operations. Therefore, assuming these two patterns ferroically order in a long range in a $4/mmm$ crystal lattice, the resulting magnetic point group will be $m'm'm'$ for the $\tilde{q}_{z^2}$ type quadrupole order, and $m'mm$ for the $\tilde{q}_{xz}$ type one.

In the end of this section, let us take the example of ferroic $\tilde{q}_{xz}$- type magnetic quadrupole order to briefly summarize the challenges of direct experimental detection of tensor (i.e., multipolar) orders. First, the summation of all the magnetic moment within a unit cell is zero, and therefore there is no net magnetization per unit cell that can couple directly with the external

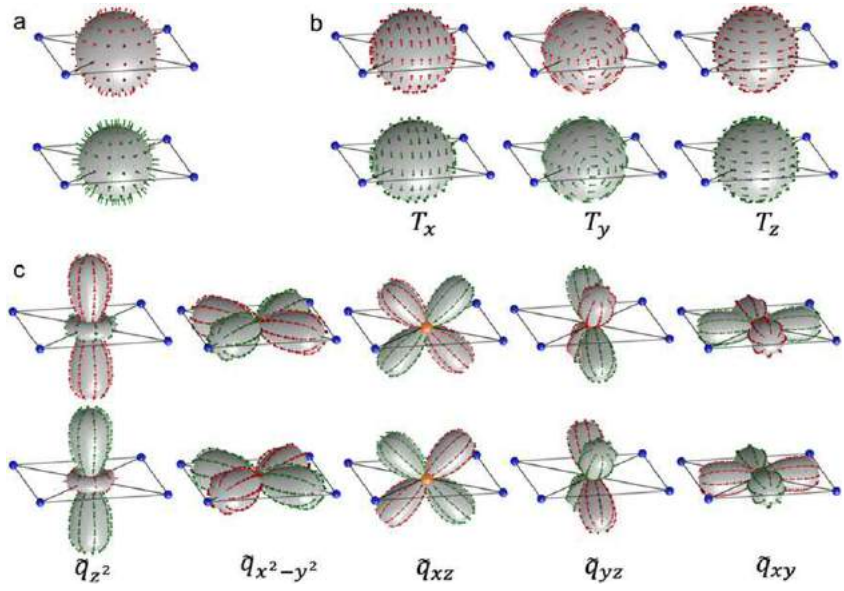

Fig. 6 Patterns of local magnetic dipole moments $\vec{\mu}$ within a unit cell of a square plaque that generates magnetic quadrupoles, created based on Fechner, M., *et al.*, 2016. Quasistatic magnetoelectric multipoles as order parameter for pseudogap phase in cuprate superconductors. Physical Review B 93(17), 174419. (a) Magnetic monopole component; (b) three toroidal components, T_x, T_y and T_z; (c) five magnetic quadrupole components, $\tilde{q}_{z^2}$, $\tilde{q}_{x^2-y^2}$, $\tilde{q}_{xz}$, $\tilde{q}_{yz}$, and $\tilde{q}_{xy}$. Patterns in the 2nd and 4th rows are the time-reversed counterparts of that in the 1st and 3rd rows.

magnetic field. Second, the order parameter here is described by a rank-2 tensor, so the conjugate field to couple with needs to be a tensor field. In principle, the multiple copies of electric and magnetic fields in the nonlinear optics can act as the tensor field that couples with the tensor order parameter. Third, the time reversal symmetry related two ground states lead to two domains in the ordered state. Therefore, the symmetry properties of ordered state can be easily concealed by the domain averaged measurements. Forth, the ferroic order preserves the translational symmetries of the underlying crystal lattice, which is the $\boldsymbol{q}=0$ problem known in diffraction measurements. As a result, the diffraction pattern from the ferroic tensor order overlays on top of the nuclear diffraction pattern, and requires extremely high resolution to resolve.

Hidden Phases Revealed by SHG-based Techniques

So far, SHG has been successfully exploited to reveal the otherwise hidden phases in several correlated electron systems because of its high sensitivity to the symmetry point groups, tensor field from multiple copies of electromagnetic fields, and spatial resolution of optical diffraction limit $\sim 1\mu m$.

A Ferrotoroidal Order in $LiCoPO_4$

$LiCoPO_4$ has the olivine crystal structure and is well characterized by the *mmm* symmetry point group in the paramagnetic state (Rabe, 2007; Spaldin *et al.*, 2008; Van Aken *et al.*, 2007; Zimmermann *et al.*, 2014). It contains planes of CoO_6 octahedra that are buckled by sharing of corners with PO_4 tetrahedra (gray) above (red Co^{2+}) and below (green Co^{2+}) the plane as shown in **Fig. 7 (a)**. Looking into the *ab*-plane, two pairs of Co^{2+} cations within one unit cell are separated into two sets with slightly different radius from their center (**Fig. 7(c)** left column without magnetic moments). At 21.8 K, the localized magnetic moments at Co^{2+} form a collinear antiferromagnetic order with alternating sign from row to row in the plane, and are oriented along a direction that is 4.6° away from the axis of the row, i.e., *b*-axis. Considering the clockwise ($+4.6^\circ$) and anti-clockwise (-4.6°) tilt of the Co^{2+} magnetic moment and their time reversal symmetry related counterparts, there are in total four distinct but energetically equivalent magnetic ground states, as sketched in **Fig. 7(b)**. In practice, each of the states can be split into two parts – one part has the *b*-axis components of the magnetic moments, and the other has the rest of the magnetic moments that are aligned along *a*-axis. The first

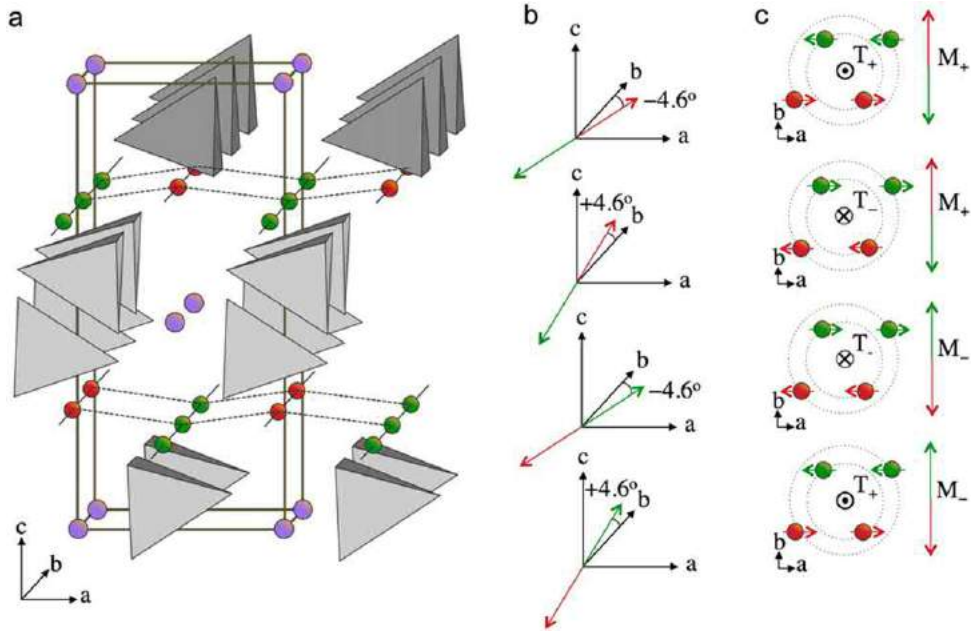

Fig. 7 Illustration of ferrotoroidicity in $LiCoPO_4$. (a) Crystal structure of $LiCoPO_4$, where cobalt ions (green and red spheres) form planar layers. These cobalt layers are buckled (red for the upward-buckled rows and green for the downward-buckled rows) by the presence of phosphorus-centered oxygen tetrahedra (gray) above and below them. And lithium ions (purple spheres) are located between the tetrahedra. (b) The magnetic moments of cobalt are collinearly aligned, but tilted at an angle of 4.6° with respect to the rows. There are in total four arrangements for the magnetic moments, two directions of the tilt relative to the rows and their time-reversed counterparts. (c) Decompositions of the magnetic moments into the a-axis component that gives rise to the non-zero toroidization out of (T^+) or into (T^-) the ab-plane (left column), and b-axis component that forms the antiferromagnetic order of the non-centrosymmetric *mmm*′ symmetry with two time-reversal related structures (M^- and M^+) (right column). Corresponding to the four arrangements in (b), there are four combinations between the toroidal moment and the magnetic moment, namely, T^+M^+, T^-M^+, T^-M^- and T^+M^-. Adapted from Rabe, K.M., 2007. Solid-state physics: Response with a twist. Nature 449 (7163), 674–675.

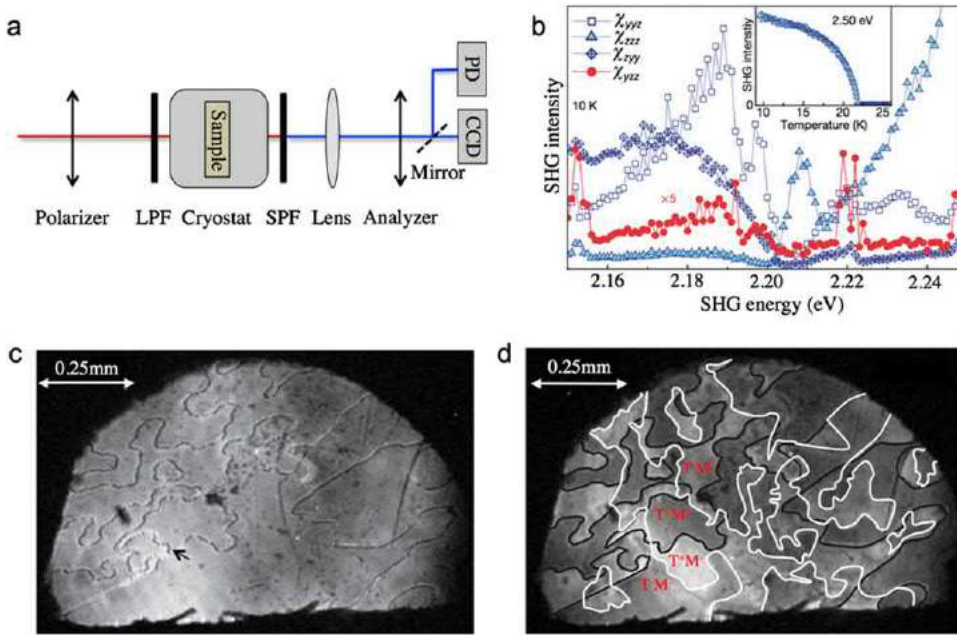

Fig. 8 Observation of ferrotoroidic domains by optical SHG. (a) Schematic of the transmission-based SHG setup. Incoming light passes through Polarizer and long pass filter (LPF) before normal incidence onto the sample in the cryostat. The transmitted light is filtered by the short pass filter (SPF) and the selected SHG light gets detected by either CCD camera (imaging) or photodiode (integrated intensity); (b) Energy dependent contributions from independent non-zero tensor components χ_{ijk} are plotted. The inset shows the temperature dependent SHG intensity from χ_{zzz} with an order parameter like onset at 21.8 K; (c) Antiferromagnetic domains that are obtained using light from χ_{yyz} are imaged. The dark lines (with dark arrow pointing one example) are the domain boundaries between the two time-reversal related antiferromagnetic domains. (d) Coexistence of antiferromagnetic and ferrotoroidic domains that are obtained using the interfering light from χ_{yyz} and χ_{zyy} is shown. The dark and white lines indicate the antiferromagnetic and ferromagnetic domain walls respectively. Examples of four areas are labeled with the order parameter combinations T^+M^+, T^+M^-, T^-M^+, and T^-M^-. Adapted from Van Aken, B.B., *et al.*, 2007. Observation of ferrotoroidic domains. Nature 449 (7163), 702–705 and Spaldin, N.A., M. Fiebig, Mostovoy, M., 2008. The toroidal moment in condensed-matter physics and its relation to the magnetoelectric effect. Journal of Physics: Condensed Matter 20(43), 434203 (All authors contributed equally to this work).

part forms an antiferromagnetic order that is characterized by the *mmm*′ symmetry point group, and has two degenerate ground states related by the time reversal symmetry operation (M_+ and M_- in the right column in **Fig. 7(c)**). The second part gives rise to a ferrotroidal order $T_i = \sum_{\alpha=1}^{N} \varepsilon_{ijk} r_{\alpha,j} \mu_{\alpha,k}$ that is characterized by 2′ symmetry point group with *a*-axis as the 2-fold rotation axis, and also has two time reversal symmetry related ground states that originate from the clockwise and anti-clockwise tilt of the magnetic moments (T_+ and T_- in the right column in **Fig. 6(c)**). Four combinations of the antiferromagnetic order and the ferrotoroidic order, i.e., M_+T_+, M_+T_-, M_-T_+ and M_-T_-, fully reproduce the four magnetic structures in **Fig. 7(b)**.

Van Aken and colleagues established (Van Aken *et al.*, 2007), for the first time, the presence of ferrotoroidic order by detecting domains of opposite toroidizations through the coherent interference between the SHG from the ferrotoroidic order and the antiferromagnetic order in $LiCoPO_4$. By measuring SHG response of a polished (100)-oriented $LiCoPO_4$ film with a thickness $122\mu m$ using a transmission SHG setup with a normal incidence geometry shown in **Fig. 8(a)**, independent nonzero electric dipole induced SHG susceptibility tensor components χ_{yyz}, χ_{zzz}, χ_{zyy} and χ_{yzz} shows up at the Néel temperature $T_N = 21.8$K (**Fig. 8(b)**), which is only compatible with the low symmetry point group 2′ for the proposed ferrotoroidic order. On the other hand, χ_{yyz} is also present in the electric dipole induced SHG susceptibility tensor under *mmm*′ symmetry point group for the antiferromagnetic order. SHG image on the $LiCoPO_4$ crystal with light from χ_{yyz} component shows dark lines on a constant-intensity background in **Fig. 8(c)**, where the dark lines are the domain boundaries between two 180°-rotated antiferromagnetic domains. The image of the interference between light from χ_{yyz} and χ_{zyy} components displays a patchwork with bright and dark areas in **Fig. 8(d)**, which indicates two SHG contributions from antiferromagnetic and ferrotoroidic order are present here and interfere constructively (bright) and destructively (dark). In **Fig. 8 (d)**, the black lines highlight the antiferromagnetic domain boundaries while the white lines outline the ferrotorodic domains. Examples of four antiferromagnetic and ferrotoroidic order combinations are also marked in **Fig. 8(d)**.

Following the work by Van Aken *et al*, Anne Zimmermann and colleagues further confirmed the ferrotoroidic nature of the ordered state in $LiCoPO_4$ below the Néel temperature by observing hysteretic poling of ferrotoroidic domains in the conjugate toroidal field $\vec{S} \propto \vec{E} \times \vec{B}$ (Zimmermann *et al.*, 2014). SHG images in **Fig. 9(a)** show that magnetic or electric field applied alone during cooling results in a multi-domain state while the product of magnetic and electric field promotes a ferrotoroidic single-domain state. A striking pronounced hysteresis is observed in the integrated SHG intensity in **Fig. 9(b)** when cycling the magnetic field within a ± 2T interval at a fixed electric field of 11 kV cm^{-1}. Furthermore, the SHG intensity v.s. $\mu_0 H \cdot E$ spectra from separate runs where the sign and amplitude of the fixed fields varied all collapse onto one another in **Fig. 9(c)**, confirming that solely the

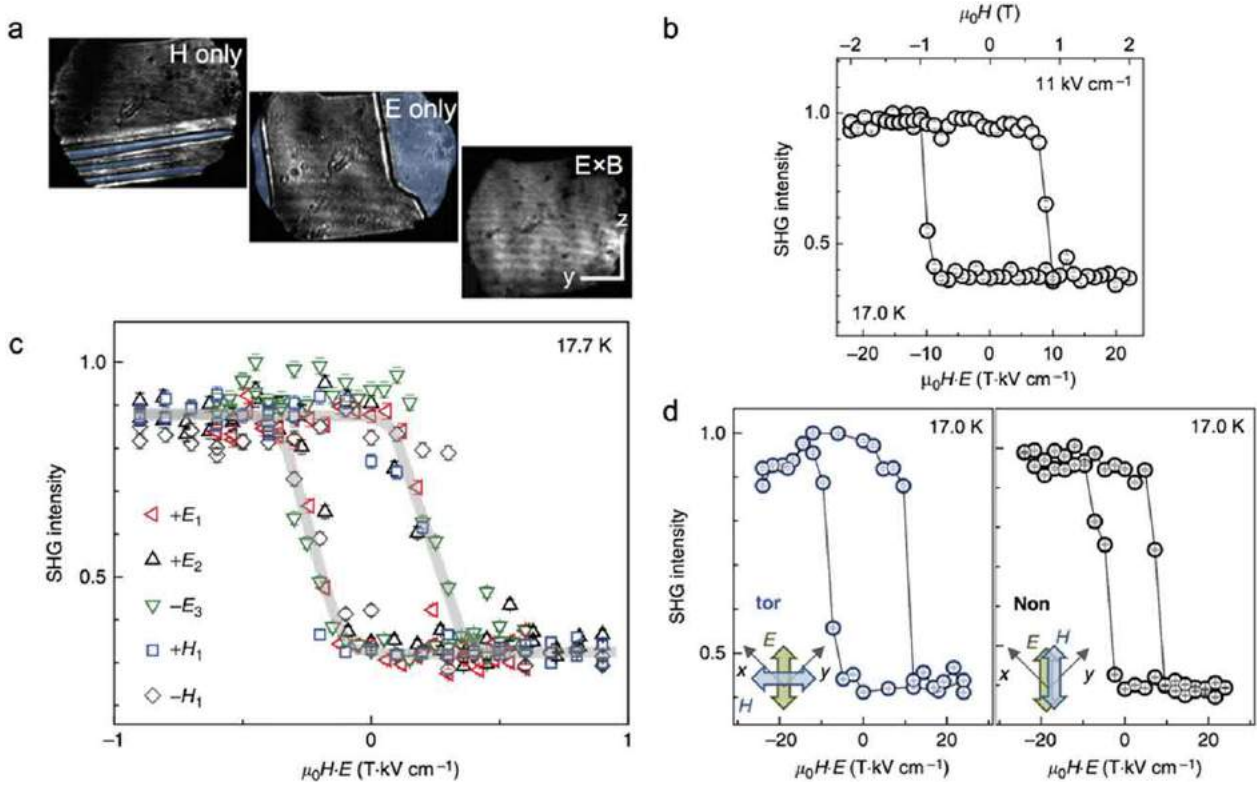

Fig. 9 Ferroic nature of magnetic toroidal order. (a) Multi-domain states (first two images) and single domain state (third image) emerge after cooling the $LiCoPO_4$ sample in a magnetic or electric field alone and in a toroidal field of orthogonal magnetic (H_x) and electric (E_y) fields respectively; (b) SHG intensity at 17.0 K as a function of a static magnetic field $\mu_0 H_x$ ranging from -2T to 2T in the presence of a static electric field E_y=11 kVcm^{-1} shows a hysteresis loop; (c) Hysteresis loops of SHG intensity collapse into one (gray guide-of-eye line) for various combinations in signs and amplitudes of magnetic and electric fields ($+E_1 = +1.2$ *kVcm*$^{-1}$, $+E_2 = +1.8$ *kVcm*$^{-1}$, $-E_3 = +1.0$ kVcm^{-1}, $+H_1 = +0.5T$, $-H_1 = -0.5T$) with the same $E_y \cdot H_x$ product; (d) Hysteresis loops in purely toroidal field and non-toroidal field with the same $E \cdot H$ product show different coercive fields. The insets show the corresponding experimental field configurations. Adapted from Zimmermann, A.S., D. Meier, D., Fiebig, M., 2014. Ferroic nature of magnetic toroidal order. Nature communications 5, 4796.

product of the magnetic and electric field determines the shape of hysteresis loop. Finally, the toroidal and non-toroidal poling are compared through measuring the coercive field under two field configurations as shown in **Fig. 9(d)**. The experimental value of $\frac{|EH|_{co.}^{tor}}{|EH|_{co.}^{non}} = 1.21 \pm 0.14$ significantly differs from the calculated value of 3 assuming equal efficiency between toroidal and non-toroidal poling, showing that toroidal field is much more effective in promoting the ferrotoroidic single-domain state.

A Parity-Broken Nematic Order in $Cd_2Re_2O_7$

The correlated metallic pyrochlore $Cd_2Re_2O_7$ undergoes a continuous phase transition from an inversion symmetric cubic structure (space group $Fd\bar{3}m$, **Fig. 10(a)**) to a parity-broken tetragonal structure (space group $I\bar{4}m2$, **Fig. 10(b)**) at T_s=200K, with the lattice distortion of the E_u symmetry (Castellan *et al.*, 2002; Kendziora *et al.*, 2005; Petersen *et al.*, 2006; Weller *et al.*, 2004; Yamaura and Hiroi, 2002). Recent theory predicts that spin-orbit coupled metal can host a variety of inversion asymmetric ordered phases resulting from Pomeranchuk type instabilities in the spin channel (Fu, 2015). These phases preserve the translational symmetries but break the point group symmetries of underlying crystal lattice, and therefore regarded as new examples of electronic liquid crystals (i.e., electronic nematicity). Moreover, the spin-orbit coupling leads to spin-split Fermi surfaces with characteristic spin textures, and generally induces structural distortions from the electronic parity-breaking orders. In particular, $Cd_2Re_2O_7$ is one of the proposed pyrochlore oxide candidates hosting a primary multipolar ordered phase of either E_u or T_{2u} type order parameters that further induces the observed secondary structural phase transition.

Prior to 2006, the low temperature structural order parameters was not settled between the distinct but nearly degenerated crystallographic structures $I\bar{4}m2$ (**Fig. 10(b)**, $\chi = \{\chi_{xyz} = \chi_{xzy} = \chi_{yxz} = \chi_{yzx} = \chi_{zxy} = \chi_{zyx}\}$) and $I4_122$ (**Fig. 10(c)**, $\chi = \{\chi_{xyz} = \chi_{xzy} = -\chi_{yxz} = -\chi_{yzx}\}$) (Castellan *et al.*, 2002; Kendziora *et al.*, 2005; Weller *et al.*, 2004; Yamaura and Hiroi, 2002). Jesse C. Petersen and colleagues resolved this mystery by using SHG ellipsometry, and identified the structural symmetry below T_s=200 K to be $I\bar{4}m2$ (Petersen *et al.*, 2006). The SHG response of the cubic (111) oriented $Cd_2Re_2O_7$ were measured on a reflection SHG setup of the normal incidence geometry, with independent control on the incident and reflected polarizations, a_ω and $a_{2\omega}$, by a half-wave plate ($\lambda/2$) and a crystal polarizer (P) respectively (**Fig. 11(a)**). In order to minimize the effects from linear birefringence,

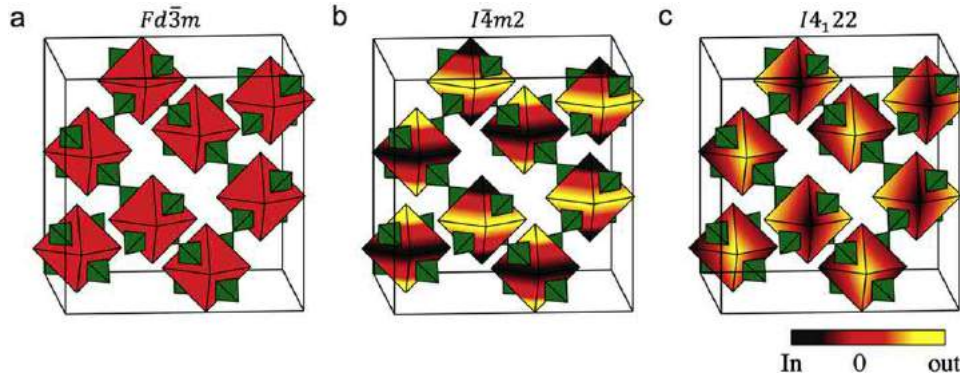

Fig. 10 The structural phase transition in $Cd_2Re_2O_7$. (a) Cubic pyrochlore unit cell with $Fd\bar{3}m$ symmetry. The vertices of green tetrahedral indicate Re sites and red octahedra indicate O_1 sites. Cd and O_2 sublattice are not shown here. (b-c) Pseudocubic unit cells of the distorted lattices with $I\bar{4}m2$ and $I4_122$ symmetry respectively. The neighboring octahedra have displacements with opposite sign, thus breaking inversion symmetry. Adapted from Petersen, J.C., *et al.*, 2006. Nonlinear optical signatures of the tensor order in $Cd_2Re_2O_7$. Nature Physics 2(9), 605–608.

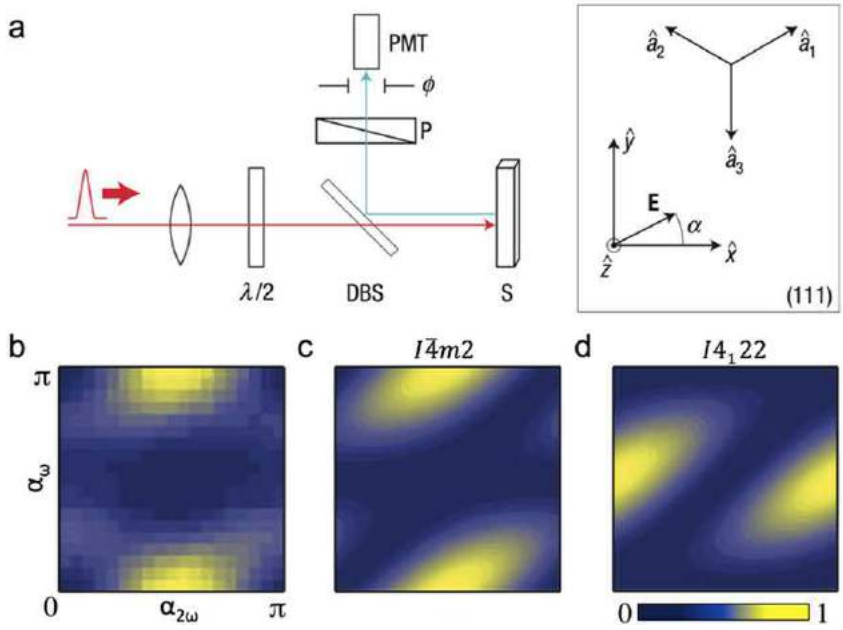

Fig. 11 Identification of crystal structure below the transition temperature $T_s = 200$K. (a) (Left) Schematic of the SHG ellipsometry setup, where femtosecond laser pulses were focused through a dichroic beamsplitter (DBS) to a cubic (111) surface of $Cd_2Re_2O_7$ and the reflected SHG was detected by a photomultiplier tube (PMT) after passing through an iris (Φ). Input and output polarizations are selected with a half-wave plate ($\lambda/2$) and a crystal polarizer (P), respectively. (Right) Cubic coordinate $(\hat{a}_1, \hat{a}_2, \hat{a}_3)$ for the high-temperature phase, and optical coordinate $(\hat{x}, \hat{y}, \hat{z})$, with $\hat{z}$-axis along the optical axis. (b) Normalized SHG intensity map as a function of input and output polarizations taken at 4 K. (c–d) Simulations of SHG intensity map under $I\bar{4}m2$ and $I4_122$ symmetry respectively. Adapted from Petersen, J.C., *et al.*, 2006. Nonlinear optical signatures of the tensor order in $Cd_2Re_2O_7$. Nature Physics 2(9), 605–608.

crystal twinning, experimental misalignment etc. on the SHG signal variations, the authors obtained 2D data sets of SHG intensity in by varying a_ω and $a_{2\omega}$. Comparing the experimental map (**Fig. 11(b)**) with the simulated maps under $I\bar{4}m2$ (**Fig. 11(c)**) and $I4_122$ (**Fig. 11(d)**) symmetries, it clearly identifies $I\bar{4}m2$ as the proper crystallographic structure for $Cd_2Re_2O_7$ at $T_s = 200$K.

John Harter and colleagues recently revisited $Cd_2Re_2O_7$ to identify the electronic origin of the primary order parameter (Harter *et al.*, 2017).

An Odd-Parity Hidden Order in Sr_2IrO_4

The single layer perovskite iridate Sr_2IrO_4 realizes a novel spin orbit coupled $J_{eff} = 1/2$ mott insulating ground state, and transits into a canted antiferromagnetic ordered state that lowers system symmetry from tetragonal ($4/mmm$ point group, **Fig. 12(a)**) to orthorhombic ($mmm1'$ point group, **Fig. 12(b)**) (Kim *et al.*, 2008, 2009). Despite the striking similarities in crystallographic,

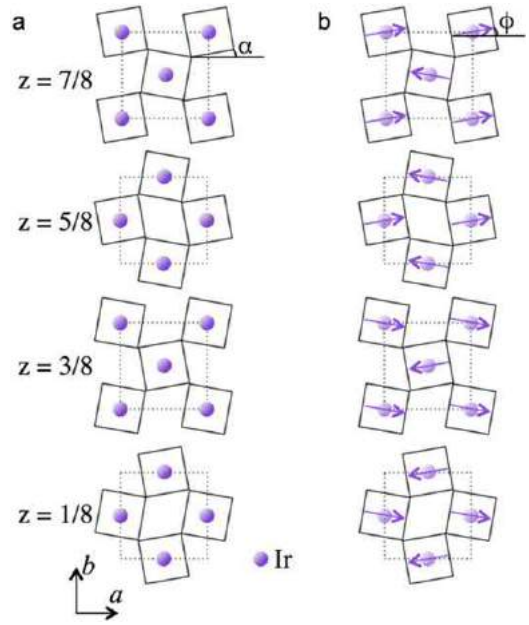

Fig. 12 Crystal structure and antiferromagnetic order in Sr_2IrO_4. (a) Crystal structure of Sr_2IrO_4, consisting four layers of IrO_6 octahedra (square in ab-plane) per unit cell. There are two sets of IrO_6 octahedra sublattices, clockwise and anti-clockwise rotated for $\alpha = \sim 12^\circ$ about c-axis. (b) Canted antiferromagnetic ordering of $J_{eff}=1/2$ moments (purple arrows) within the ab-plane and their stacking pattern along c-axis. The magnetic moments perfectly lock with the IrO_6 octahedra edges ($\phi/\alpha = 1$).

magnetic structures and electronic ground state between Sr_2IrO_4 and the cuprate parent compound La_2CuO_4, recent experimental realizations of pseudogap and D-wave gap behaviors in the doped Sr_2IrO_4 systems further bridge the analogy between Sr_2IrO_4 and cuprates. More recently, the observation of an odd-parity multipolar order in Sr_2IrO_4 and the Rh-doped counterpart is one more surprising reproduction of cuprate phenology in this iridate system.

It was a mystery why the magnetic moment at the Ir^{4+} locks perfectly with the IrO_6 octahedron orientation ($\phi/a = 1$ in **Fig. 12 (b)**), despite there are two opposite canting directions. Darius H Torchinsky and colleagues resolved a fine staggered tetragonal distortion of IrO_6 octahedra that enhances the magnetoelastic coupling to lock the magnetic moment with octrahedra rotation perfectly (Torchinsky *et al.*, 2015), using first version of rotating-scattering-plane-based RA integer harmonic generation setup (**Fig. 4** in Section "Generations of Rotational Anisotropy Experimental Designs"). The third harmonic generation (THG) RA patterns (**Fig. 13(a)**) were observed to rotate away from the crystal axes a and b for an angle of $\sim 11.5^\circ$, and the two in $P_{in} - S_{out}$ and $S_{in} - P_{out}$ geometries show alternating big-small lobes. Both observations prove broken mirror planes m_{ac} and m_{bc} in the data, therefore in the crystal as well. The largest subgroup of previously assigned structural symmetry point group $4/mmm$ is the centrosymmetric $4/m$, and the THG RA patterns are well fit by the bulk electric dipole induced THG under $4/m$. Microscopically, the loss of the two mirror planes originates from a staggered tetragonal distortion of the crystal lattice where the two sets of IrO_6 octahedral sublattices have unequal tetragonal splitting (Δ_1 and Δ_2 shown in **Fig. 13(b)**). In particular, theoretical calculations found that the magnetoelastic locking is remarkably insensitive (i.e., $\phi/a = 1$) to both the tetragonal distortion size and spin orbit coupling strength when the signs of the tetragonal distortion is staggered between sublattices ($\Delta_1 = -\Delta_2$), shown in **Fig. 13(c)**. In contrast, calculations on the uniform tetragonal distortion model ($\Delta_1 = \Delta_2$) shows a sharp decrease of the locking when the tetragonal distortion size increases at a given spin orbit coupling strength (**Fig. 13(d)**). The comparison between the two models suggests that the magnitude of the locking is more strongly influenced by the spatially averaged value of the tetragonal distortion rather than its local value on an individual IrO_6 octahedron, allowing the existence of large local tetragonal distortion to be reconciled with the observation of perfect magnetoelastic locking.

Cooling Sr_2IrO_4 down to the cryogenic temperature, Liuyan Zhao and colleagues revealed an odd-parity order that breaks the global inversion symmetry (Zhao *et al.*, 2016b), using the SHG RA setup in **Fig. 4**. At room temperature, the SHG RA pattern, in the $P_{in} - S_{out}$ polarization geometry in **Fig. 14(a)**, results from the bulk electric quadrupole induced SHG from the centrosymmetric crystal structure with $4/m$ symmetry. In contrast, the low temperature SHG RA pattern, $P_{in} - S_{out}$ in **Fig. 14(b)**, clearly breaks rotational symmetries and inversion symmetry, which is fitted by an interference between the structural contribution of bulk electric quadrupole induced SHG under $4/m$ and the emergent order contribution of bulk electric dipole induced SHG under $2'/m$ (or $m1'$). Here, it is worth noting that this new order is most obvious in the $P_{in} - S_{out}$ geometry while barely notable in the $S_{in} - P_{out}$ channel (**Fig. 14(c)** and **(d)**). Take this hidden order as an example, if we measure with the setup in **Fig. 3(b)** that only has cross and parallel polarized channels, the $P_{in} - S_{out}$ and $S_{in} - P_{out}$ signals above will be combined in the cross polarized channel. As a result, the small signal from the new order will be a tiny change on top of a large background from the structural signal, and therefore can be easily concealed, just as in the case of $S_{in} - P_{out}$ above.

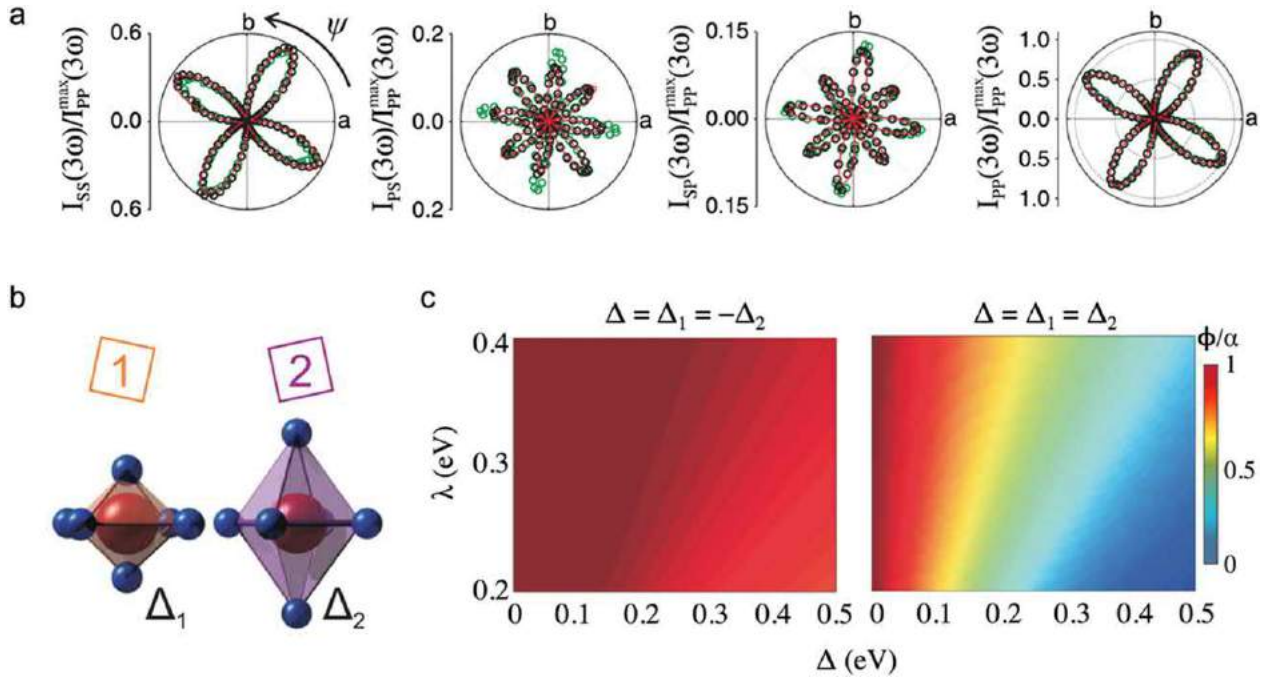

Fig. 13 Staggered tetragonal distortions in Sr_2IrO_4 revealed by THG. (a) THG RA patterns in four polarization geometries, $S_{in} - S_{out}$(SS), $P_{in} - S_{out}$(PS), $S_{in} - P_{out}$(SP) and $P_{in} - P_{out}$(PP), with THG intensity scaled to the PP maxima. Black circles are taken at 295K while green are at 180K, and red lines are best fits to the 295K data using bulk electric dipole induced THG RA patterns calculated under centrosymmetric 4/*m* point group. (b) An unequal tetragonal distortion (Δ_1 and Δ_2) on the two sublattices as required by the 4/*m* point group. (c–d) The calculated ratio of ϕ/a as a function of both λ and Δ for the case of staggered ($\Delta = \Delta_1 = -\Delta_2$) and uniform ($\Delta = \Delta_1 = \Delta_2$) distortion respectively. Adapted from Torchinsky, D.H., *et al.* 2015. Structural distortion-induced magnetoelastic locking in Sr_2IrO_4 revealed through nonlinear optical harmonic generation. Physical Review Letters 114(9), 096404.

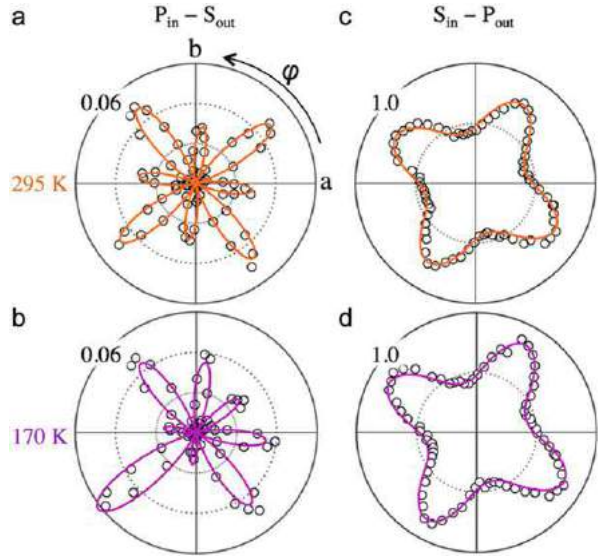

Fig. 14 Symmetry of the hidden order in Sr_2IrO_4. (a–b) SHG RA data collected in the $P_{in} - S_{out}$ polarization geometry at $T = 295$K and $T = 170$K, respectively. (*c*–d) Analogous data taken in the $S_{in} - P_{out}$ polarization geometry. The orange lines are best fits to the data at 295K using electric quadrupole induced SHG under the centrosymmetric 4/*m* point group, and the purple lines are best fits to the data at 170K using a coherent interference between electric quadrupole (centrosymmetric 4/*m* point group) and electric dipole (non-centrosymmetric 2′/*m* or *m*1′ point group) induced SHG. Adapted from Zhao, L., *et al.* 2016b. Evidence of an odd-parity hidden order in a spin-orbit coupled correlated iridate. Nature Physics 12(1), 32–36.

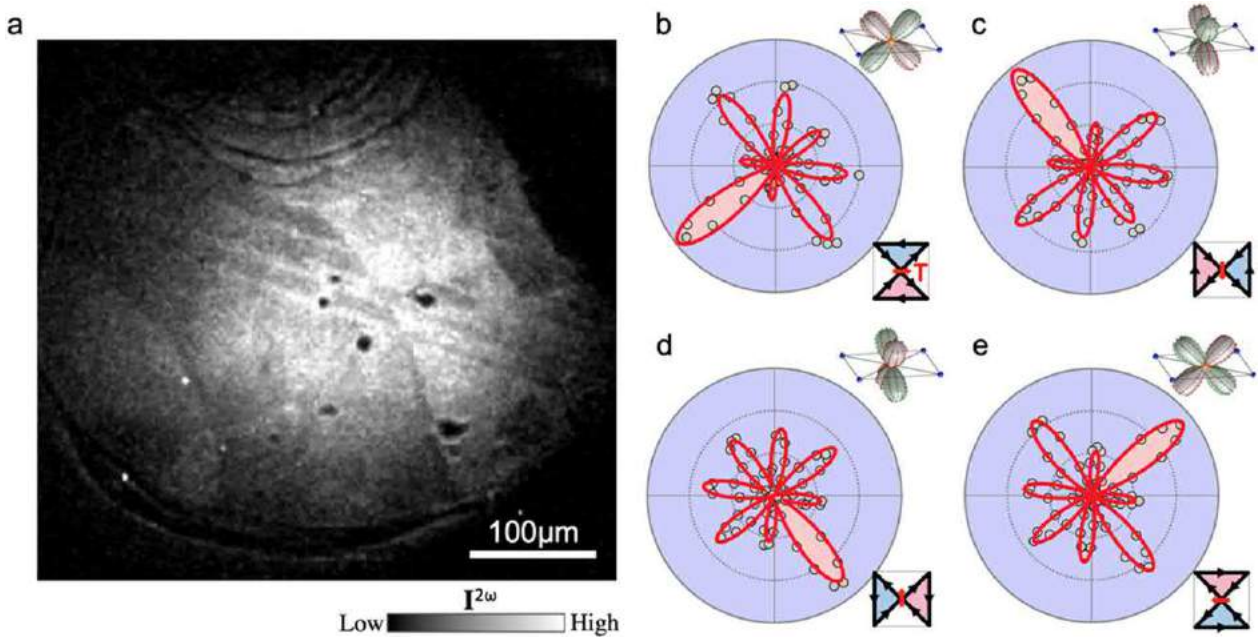

Fig. 15 Degenerate ground states of the hidden order in Sr_2IrO_4. (a) Wide-field reflection SHG image of Sr_2IrO_4 (001) plane measured under $P_{in} - S_{out}$ polarization geometry at T=175K. The brighter and darker patchwork in the image arise from the domains of the hidden order. (b–e) Four different types of SHG RA patterns found by performing local measurements within all the domains in (a). The largest lobes are shaded in pink to highlight the orientation of the patterns that are rotated every 90°. Schematics of the four degenerate magnetic quadrupole configurations are in the upper right insets while that of magneto-electric loop-current order are shown in the lower right insets. Adapted fromZhao, L., *et al.* 2016b. Evidence of an odd-parity hidden order in a spin-orbit coupled correlated iridate. Nature Physics 12(1), 32–36.

Wide field SHG image at 175K in **Fig. 15(a)** shows bright and dark patches that are separated by domain boundaries straight over a length scale of tens of mircrometers. Local RA SHG measurements within each of the patches across the crystal reveal a total of four distinct of SHG RA patterns in **Fig. 15(b–e)**, which are exactly 0°, 90°, 180° and 270°-rotated copies of that shown in **Fig. 14(b)**. These results suggest four degenerate ground states of the noncentrosymmetric ordered state, and are consistent with the microscopic model of magneto-electric loop-current order and magnetic quadrupole order. On one hand, the magneto-electric loop-current order, consisting a pair of counter-circulating current loops in each IrO_2 plaquet (Kung *et al.*, 2014; Orenstein, 2011; Varma, 1997; Weber *et al.*, 2009; Yakovenko, 2015), satisfies $2'/m$ point group and can be described by a toroidal pseudovector order parameter defined as $\vec{T} = \sum \vec{r}_i \times \vec{\mu}_i$, where $\vec{r}_i$ is the location of the orbital magnetic moment $\vec{\mu}_i$ inside the plaquette. Four degenerate configurations exist in a tetragonal lattice because the two intra-unit-cell current loops can lie along either of the two diagonals in the square plaquet and can have either of two time-reversed configurations, which correspond to four 90°-rotated directions of the pseudovector T as shown in the lower right inset of **Fig. 15(b–e)**. On the other hand, because of the tetragonal distortion of IrO_6 octahedron, $\tilde{q}_{xy}$ and $\tilde{q}_{yz}$ magnetic quadrupoles remain degenerate but split away from $\tilde{q}_{xy}$. $\tilde{q}_{xz}$, $\tilde{q}_{yz}$ and their time-reversed counterparts form the four degenerate configurations shown in the upper right inset of **Fig. 15(b–e)**, and each of their ferroic orders satisfies $2'/m$ point group. SHG RA data so far cannot differentiate between these two models, and certainly cannot rule out any other microscopic models that obey the same set of symmetries.

The temperature and doping evolution of the odd-parity hidden order in $Sr_2Ir_{1-x}Rh_xO_4$ is summarized in the phase diagram in **Fig. 16**. The hidden phase transition observed in the parent Sr_2IrO_4 persist in the hole doped $Sr_2Ir_{1-x}Rh_xO_4$, even though the transition temperature T_Ω is suppressed with increasing x. Remarkably, the splitting between T_Ω and the Néel temperature T_N grows monotonically from approximately 2 to 75K between $x=0$ and $x\sim0.11$, indicating that the Néel and hidden order are not trivially tied, but are independent and distinct electronic phase. But the relationship between this hidden phase and the pseudogap phenomena remains uncertain in Rh-doped Sr_2IrO_4 system because of lacking of the pseudogap onset temperature here.

A Magnetic Quadrupole Order in the Pseudogap of $YBa_2Cu_3O_y$

An enigmatic pseudogap region in the phase diagram of cuprates, characterized by a partial suppression of low energy electronic excitations, has been subjected to the debate of its nature for almost 30 years. Taking $YBa_2Cu_3O_y$ as an example, its parent $YBa_2Cu_3O_6$ has a tetragonal crystal structure (point group 4/*mmm*, **Fig. 17(a)**) while its detwinned hole doped counterparts $6<y\leq7$ have an orthorhombic crystal structure because of the O occupation in the CuO chains along *b*-axis (point group *mmm*, **Fig. 17(b)**) (Shaked, 1994). Recently, polarized neutron diffraction (Fauqué *et al.*, 2006; Mangin-Thro *et al.*, 2015; Mook *et al.*, 2008), Nernst effect (Daou *et al.*, 2010), THz polarimetry (Lubashevsky *et al.*, 2014) and ultrasound (Shekhter *et al.*, 2013)

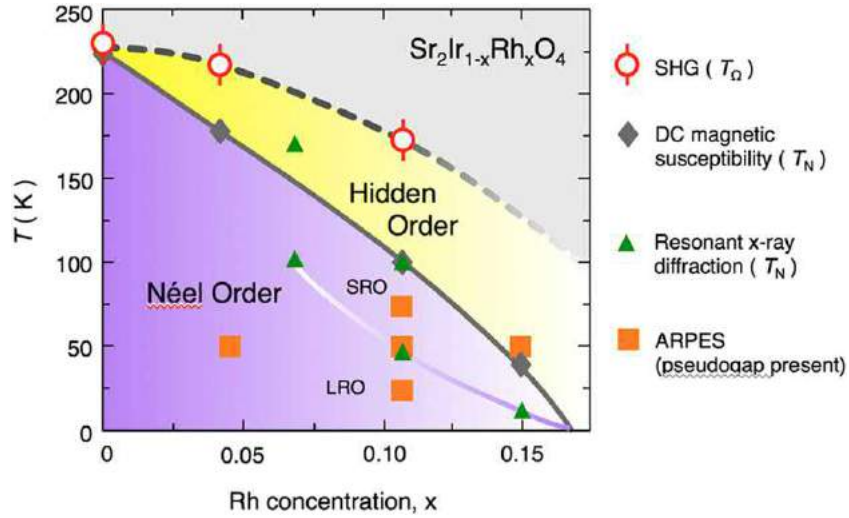

Fig. 16 Temperature versus doping phase diagram of $Sr_2Ir_{1-x}Rh_xO_4$. Onset temperatures of the hidden order (red circle), long range (LRO, green triangle) and short range (SRO, green triangle) (Clancy *et al.*, 2014) antiferromagnetic order (gray diamond) (Cao *et al.*, 2014; Qi *et al.*, 2012) are marked at multiple Rh doping levels. Points where a pseudogap (Cao *et al.*, 2014) (orange square) present are also marked, even though a pseudogap onset temperature is not known yet. Adapted from Zhao, L., *et al.* 2016b. Evidence of an odd-parity hidden order in a spin-orbit coupled correlated iridate. Nature Physics 12(1), 32–36.

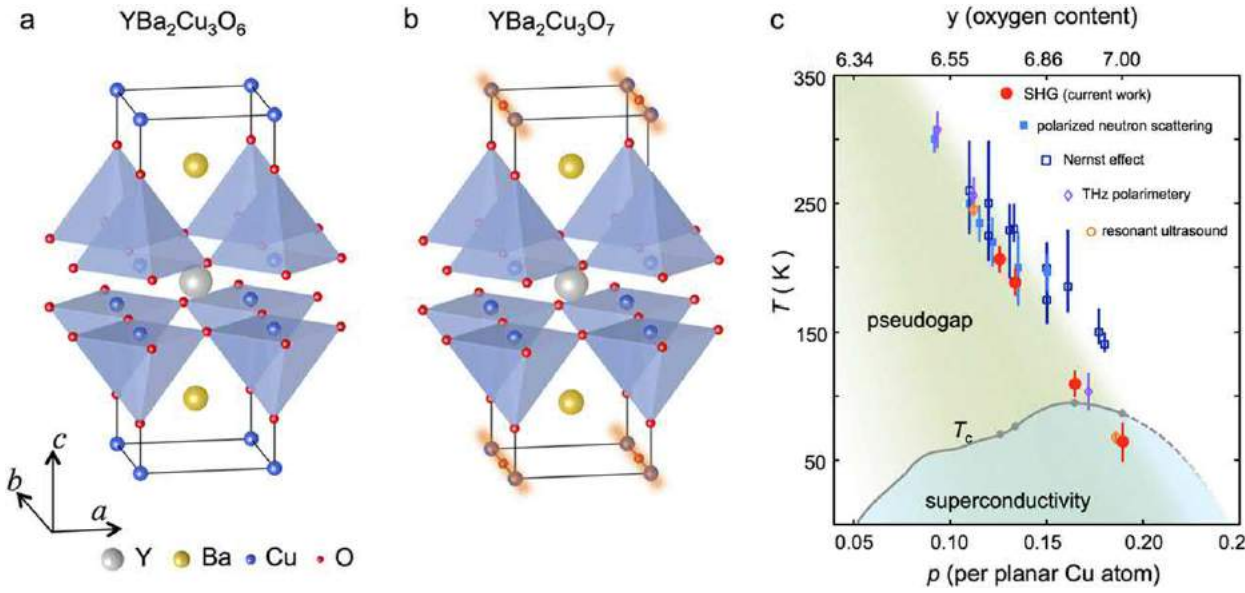

Fig. 17 Structure and phase diagram of $YBa_2Cu_3O_y$. (a–b) Tetragonal 4/*mmm* crystal structure for $YBa_2Cu_3O_6$ and orthorhombic *mmm* crystal structure for $YBa_2Cu_3O_7$, where the extra one oxygen per unit cell fully occupies the Cu-O chain along b-axis (orange shaded chains) in $YBa_2Cu_3O_7$. (c) Temperature versus oxygen doping (or hole doping) phase diagram of $YBa_2Cu_3O_y$ showing broken symmetry phases at the boundary of pseudogap region detected by SHG (red circles), polarized neutron scattering (blue squares) (Fauqué *et al.*, 2006; Mangin-Thro *et al.*, 2015; Mook *et al.*, 2008), Nernst effect (open navy squares) (Daou *et al.*, 2010), THz polarimetry (purple diamonds) (Lubashevsky *et al.*, 2014) and resonant ultrasound (orange circles) (Shekhter *et al.*, 2013) measurements. Adapted from Zhao, L., *et al.* 2016a. A global inversion-symmetry-broken phase inside the pseudogap region of $YBa_2Cu_3O_y$ Nature Physics 13(3), 250–254.

measurements on $YBa_2Cu_3O_y$ suggest that the pseudogap onset temperature T^* coincides with a phase transition that breaks time-reversal, four-fold rotation, and mirror symmetries respectively (**Fig. 17(c)**). However, the microscopic nature of the pseudogap and its fate in the overdoped region remains poorly understood.

Zhao and colleagues approaches the nature of pseudogap by tracking its full symmetry evolution over a wide range of doping levels from underdoped ($y=6.67,6.75$) to optimal doped ($y=6.92$) to overdoped ($y=7.0$) (Zhao *et al.*, 2016a), using the RA setup described in **Fig. 4**. Above T^*, linear RA patterns from $YBa_2Cu_3O_y$ measured in $S_{in}-S_{out}$ geometry (**Fig. 18(a–d)** insets) exhibits not

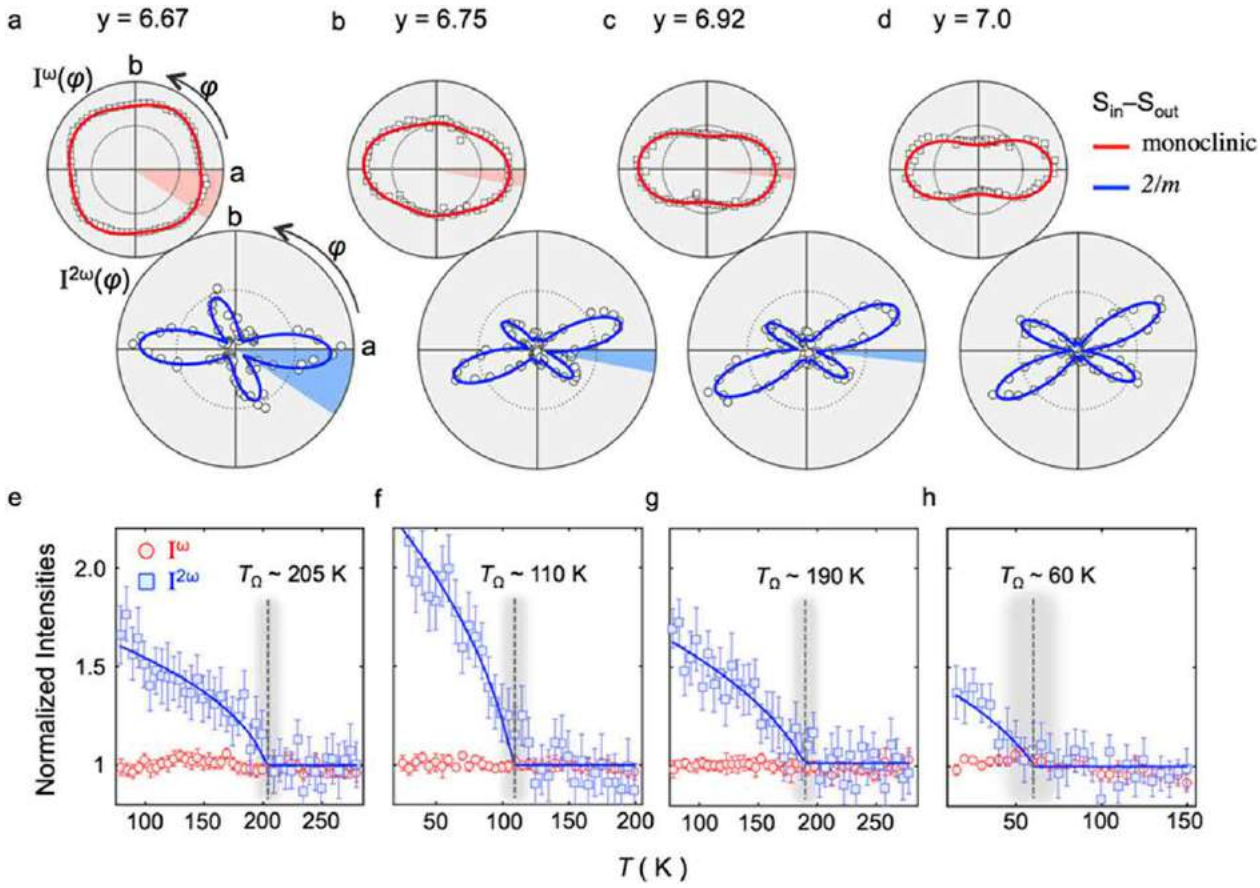

Fig. 18 Doping dependence of crystal symmetry and multipolar order in $YBa_2Cu_3O_y$. (a–d) Polar plots of linear (upper left insets) and SHG RA patterns measured at $T=295$K in $S_{in}-S_{out}$ polarization geometry from $YBa_2Cu_3O_y$ with y=6.67, 6.75, 6.92 and 7.0 respectively. The former are fit to electric dipole induced linear response of monoclinic crystal system (red lines) and the latter are fit to electric quadrupole induced SH response of the centrosymmetry 2/*m* point group (blue line). The angular deviations of the maxima of linear RA patterns and lobe bisector of SHG RA patterns away from the a-axis are shaded with red and blue respectively. (e–h) Temperature dependence of linear and SH response of $YBa_2Cu_3O_y$ with y=6.67, 6.75, 6.92 and 7.0 scaled to their room temperature values. Blue curves overlaid on SH data sets are guides to the eye. The inversion symmetry breaking transition temperature T_{Ω} determined from SH data are marked with dashed gray lines. The width of the shaded gray intervals represents the uncertain of T_{Ω}. Adapted from Zhao, L., *et al.* 2016a. A global inversion-symmetry-broken phase inside the pseudogap region of $YBa_2Cu_3O_y$ Nature Physics 13(3), 250–254.

only more pronounced anisotropy with increasing hole doping as expected from the filling of CuO chains, but also a rotation of the intensity maxima and minima away from the *a* and *b*-axes indicating an absence of m_{ac} and m_{bc} symmetries consistent with a monoclinic distortion. SHG RA data (Fig. 18(a–d)) further narrows down the symmetry point group to be the centrosymmetric 2/*m* out of the monoclinic class, and the source of SHG response therefore to be bulk electric quadrupole. The degree of monoclinicity, qualitatively tracked by via the angular deviations of the RA patterns away from *a*-axis, decreases monotonically over the range between $y=6.67$ and $y=7$, suggesting that this structural refinement from *mmm* to 2/*m* originates from the vacancy induced monoclinic distortions in the oxygen sublattice.

For all four doping levels shown in Fig. 18(e–h), the linear responses have no observable temperature dependence while, in sharp contrast, SHG responses exhibit a significant order parameter like upturn below a doping dependent critical temperature T_{Ω}. This dichotomy is naturally reconciled if the bulk inversion symmetry is broken below T_{Ω}, and turns on a stronger electric dipole induced SHG radiation on top of the pre-existing electric quadrupole response. More importantly, the onset of this inversion symmetry broken order T_{Ω} coincides with the pseudogap onset T^{*} in the underdoped and optimal doped region, and the boundary of this order extends into the superconducting dome in the overdoped region as shown in Fig. 17(c). The absence of any measurable anomalies in the SHG intensity across either the superconducting or the charge order temperature indicates that this non-centrosymmetric phase is independent of and coexists with both of them.

In fact, the enhanced SHG signal in $S_{in}-S_{out}$ geometry below T_{Ω} rules out the possibility of two-fold rotation symmetry C_2 present in the inversion symmetry broken phase, because the ED induced SHG inferred above is strictly forbidden by symmetry in $S_{in}-S_{out}$ geometry for any point that contains C_2. As a result, the phase below T_{Ω} breaks both inversion symmetry and C_2, and has the symmetry of either of the two magnetic point group $2'/m$ and $m1'$, or their subgroups. Therefore, the SHG RA pattern below T_{Ω}

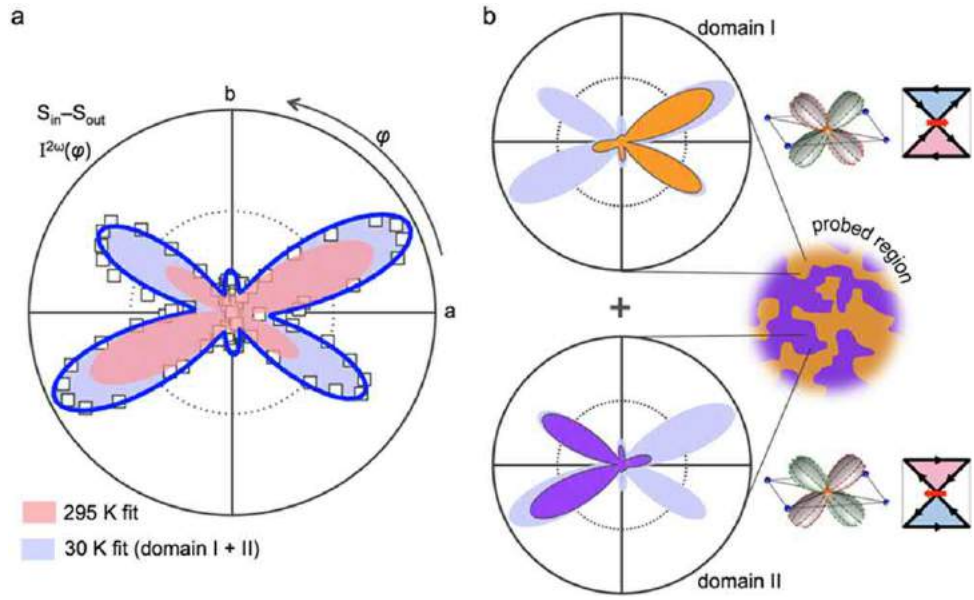

Fig. 19 Point group symmetry in the pseudogap region of $YBa_2Cu_3O_y$. (a) Polar plot of SH RA pattern measured at $T=30K$ in $S_{in}-S_{out}$ polarization geometry from $YBa_2Cu_3O_{6.92}$ (squares). The blue curve is a best fit to the average of two 180° rotated domains with $2'/m$ point group symmetry. The shaded pink area is the fit to the SH RA pattern at $T=295K$ that is overlaid for comparison. (b) Decomposition of the fit to the $R=30K$ data (shaded blue area) into its individual domain contributions (shaded orange and purple areas), with the schematics of two degenerate magnetic quadrupole or magneto-electric loop current configurations shown on the right side correspondingly. A schematic of the domain structure inside our beam spot is shown to the right. Adapted from Zhao, L., *et al.* 2016a. A global inversion-symmetry-broken phase inside the pseudogap region of $YBa_2Cu_3O_y$ Nature Physics 13(3), 250–254.

($S_{in}-S_{out}$ geometry, **Fig. 19(a)**) seemingly preserves two-fold rotation symmetry because of spatial averaging over domains of two degenerate ground states whose order parameters are related by 180°-rotation about *c*-axis (**Fig. 19(b)**). This implies a characteristic domains size that is much smaller than the laser spot size. Moreover, the low symmetry ($2'/m$ or $m1'$) underlying the pseudogap cannot be explained by stripe or nematic type orders alone, but instead suggest the presence of an odd-parity magnetic order parameter, which is consistent with theoretical proposals involving a ferroic ordering of local Cu-site magnetic-quadrupoles (inset 1 of **Fig. 19(b)**) (Fechner *et al.*, 2016; Lovesey *et al.*, 2015), current loops circulating within Cu-O octahedra (inset 2 of **Fig. 19(b)**) (Simon and Varma, 2002, 2003; Varma, 1997), O-site moments (Moskvin, 2012) or magneto-electric multipoles (Fechner *et al.*, 2016).

Although RA measurements above cannot identify the specific microscopic origin out of a number of candidates above, the results suggest a boundary of an odd-parity magnetic phase for the pseudogap region that extends inside the superconducting dome slightly beyond optimal doping. It is worth noting that this phase is similar to the one discussed in the pseudogap region of Sr_2IrO_4, a cuprate analogue, in part iii above, which indicates a robust connection between pseudogap and this odd-parity hidden order.

Conclusions and Prospects

SHG RA experiments in correlated electron systems show that nonlinear optics is not only a powerful complement to existing diffraction techniques to refine subtle structural distortions, but also a unique probe for detecting multipolar orders hidden to most existing techniques. While the elementary symmetry point group analysis can already tell a lot about the symmetry properties of the multipolar order, such as the odd-parity magnetic order in both Sr_2IrO_4 and $YBa_2Cu_3O_y$, a more sophisticated group theory analysis together with Landau theory can advance the understanding on the orders and their couplings to a significant depth, such as the primary odd-parity nematic order and the secondary structural order in $Cd_2Re_2O_7$.

The relative SHG amplitude of the hidden phase contributions to the crystallographic structure background exceeds the linear counterpart by orders of magnitude. In the case of multipolar order, the vector field of linear optics fails to couple with the tensor order parameter effectively, so that the nonlinear optical techniques become especially powerful. For example, in $YBa_2Cu_3O_y$, linear optical response can only determine the monoclinic crystal class for the room temperature structure, and shows no observable changes upon the emergence of the odd-parity magnetic order and pseudogap. On the other hand, SHG refines the structure to the centrosymmetric $2/m$ point group, and further resolves the broken inversion and two-fold rotation symmetries in the pseudogap region. Such advantages of SHG not only comes from its great sensitivity to symmetries, in particular, inversion symmetry, but also benefits from the tensor fields formed by more than one copy of magneto-electric fields.

So far, we have only focused on SHG, and would like to briefly discuss some possibilities of other second order nonlinear optics including SFG and DFG in this paragraph. First, the degenerate nature of SHG enforces the permutation symmetry of the last two indices of electric dipole induced SHG susceptibility tensor, i.e., $\chi_{ijk}=\chi_{ikj}$, which can forbid SHG response from some non-centrosymmetric system (Boyd, 2003). For example, $\chi=\{\chi_{xyz}=-\chi_{xzy}=\chi_{yzx}=-\chi_{yxz}=\chi_{zxy}=-\chi_{zyx}\}$ for cubic point group 432, and it will turn into zero after the SHG permutation symmetry $\{\chi_{xyz}=\chi_{xzy},\chi_{yzx}=\chi_{yxz},\chi_{zxy}=\chi_{zyx}\}$. In order to probe systems with symmetry such as 432, it is essential to apply SFG or DFG, rather than SHG. Second, as we have discussed in Sections "Generations of Rotational Anisotropy Experimental Designs and An Odd-Parity Hidden Order in Sr_2IrO_4", the more control we have on selecting tensor elements, the higher chance we have to capture hidden phases. However, in SHG measurements, there is no independent control over the two copies of incoming electromagnetic fields, so that it is in general hard to select individual tensor elements. Alternatively, SFG and DFG overcome this challenge well thanks to their two independent incoming fields. Third, in correlated electron systems, the impact of emergent orders on the electronic states usually happens to the ones within a couple of hundreds meV around the Fermi level whose resonant energy is much smaller than optical photon energy and hardly accessible by optical SHG or SFG. On the other hand, DFG, with the photon energy difference between the two incoming fields matching that of the changing states, can be especially sensitive to such phase transitions.

In this review, we have discussed only a few examples that demonstrate the possibilities of SHG RA measurements probing hidden multipolar orders in correlated electron system. We believe that the great potential of nonlinear optics and second order nonlinear optics in particular are far from being well explored, and foresee its power in revealing unconventional symmetry broken phases in correlated materials. Besides static measurements that we discussed through the review, the ultrashort laser pulses meant for strong magnetoelectric fields to enhance nonlinear responses are naturally compatible with time-resolved pump-and-non-linear-probe measurements, to explore the dynamics of the hidden orders, disentangle their couplings with coexisting orders, or even probe the transient photo-induced symmetry breaking phases.

References

Birss, R.R., 1964. Symmetry and Magnetism, 863. Amsterdam: North-Holland.

Boyd, R.W., 2003. Nonlinear optics. New York, NY: Academic Press.

Cao, Y., Wang, Q., Waugh, J.A., *et al.* 2016. Hallmarks of the Mott-Metal crossover in the hole doped $J=1/2$ Mott insulator Sr_2IrO_4. Nature communications 7, 11367.

Castellan, J.P., Gaulin, B.D., Van Dujin, J., *et al.*, 2002. Structural ordering and symmetry breaking in $Cd_2Re_2O_7$. Physical Review B 66 (13), 134528.

Clancy, J.P., Lupascu, A., Gretarsson, H., *et al.*, 2014. Dilute magnetism and spin-orbital percolation effects in $Sr_2Ir_{1-x}Rh_xO_4$. Physical Review B 89 (5), 054409.

Clark, D.J., Senthikumar, V., Le, C.T., *et al.*, 2014. Strong optical nonlinearity of CVD-grown MoS_2 monolayer as probed by wavelength-dependent second-harmonic generation. Physical Review B 90 (12), 121409.

Dähn, A., Hübner, W., Bennemann, K.H., 1996. Symmetry analysis of the nonlinear optical response: Second harmonic generation at surfaces of antiferromagnets. Physical Review Letters 77 (18), 3929.

Daou, R., Chang, J., LeBoeuf, D., *et al.*, 2010. Broken rotational symmetry in the pseudogap phase of a high-T_c superconductor. Nature 463 (7280), 519–522.

Fauqué, B., Sidis, Y., Hinkov, V., *et al.*, 2006. Magnetic order in the pseudogap phase of high-T_c superconductors. Physical Review Letters 96 (19), 197001.

Fechner, M., Fierz, M. J. A., Thole, F., *et al.*, 2016. Quasistatic magnetoelectric multipoles as order parameter for pseudogap phase in cuprate superconductors. Physical Review B 93 (17), 174419.

Fiebig, M., Fröhlich, D., Krichevtsov, B.B., *et al.*, 1994. Second harmonic generation and magnetic-dipole-electric-dipole interference in antiferromagnetic Cr_2O_3. Physical Review Letters 73 (15), 2127.

Fiebig, M., Fröhlich, D., Kohn, K., *et al.*, 2000. Determination of the magnetic symmetry of hexagonal manganites by second harmonic generation. Physical Review Letters 84 (24), 5620.

Fiebig, M., Pavlov, V.V., Pisarev, R.V., 2005. Second-harmonic generation as a tool for studying electronic and magnetic structures of crystals: Review. JOSA B 22 (1), 96–118.

Franken, P.A., Hill, A.E., Peters, C.W., *et al.*, 1961. Generation of optical harmonics. Physical Review Letters 7 (4), 118.

Fu, L., 2015. Parity-breaking phases of spin-orbit-coupled metals with gyrotropic, ferroelectric, and multipolar orders. Physical Review Letters 115 (2), 026401.

Gridnev, V.N., Pavlov, V.V., Pisarev, R.V., *et al.*, 2001. Second harmonic generation in anisotropic magnetic films. Physical Review B 63 (18), 184407.

Hamh, S.Y., Park, S.H., Jerng, S.K., *et al.*, 2016. Surface and interface states of Bi_2Se_3 thin films investigated by optical second-harmonic generation and terahertz emission. Applied Physics Letters 108 (5), 051609.

Harter, J.W., Niu, L., Woss, A.J., *et al.*, 2015. High-speed measurement of rotational anisotropy nonlinear optical harmonic generation using position-sensitive detection. Optics Letters 40 (20), 4671–4674.

Harter, J.W., Chu, H., Jiang, S., *et al.*, 2016. Nonlinear and time-resolved optical study of the 112-type iron-based superconductor parent $Ca_{1-x}La_x$ $FeAs_2$ across its structural phase transition. Physical Review B 93 (10), 104506.

Harter, J.W., Zhao, Z.Y., Yan, J.Q., Mandrus, D., Hsieh, D., 2017. A parity-breaking electronic nematic phase transition in the spin-orbit coupled correlated metal $Cd_2Re_2O_7$. Science 356 (6335), 295.

Heinz, T.F., Loy, M.M.T., Thompson, W.A., 1985. Study of Si (111) surfaces by optical second-harmonic generation: reconstruction and surface phase transformation. Physical Review Letters 54 (1), 63.

Hogan, T., Bjaalie, L., Zhao, L., *et al.*, 2016. Structural investigation of the bilayer iridate $Sr_3Ir_2O_7$. Physical Review B 93 (13), 134110.

Hsieh, D., McIver, J.W., Torchinsky, D.H., *et al.*, 2011. Nonlinear optical probe of tunable surface electrons on a topological insulator. Physical Review Letters 106 (5), 057401.

Huai, P., Nasu, K., 2002. Difference between photoinduced phase and thermally excited phase. Journal of the Physical Society of Japan 71 (4), 1182–1188.

Ichikawa, H., Nozawa, S., Sato, T., *et al.*, 2011. Transient photoinduced 'hidden'phase in a manganite. Nature Materials 10 (2), 101–105.

Jackson, J.D., 1999. Classical Electrodynamics. New York, NY: Wiley.

Janisch, C., Wang, Y., Ma, D., *et al.*, 2014. Extraordinary second harmonic generation in tungsten disulfide monolayers. Scientific Reports 4, 5530.

Kendziora, C.A., Sergienko, I.A., Jin, R., *et al.*, 2005. Goldstone-mode phonon dynamics in the pyrochlore $Cd_2Re_2O_7$. Physical Review Letters 95 (12), 125503.

Kim, B.J., Jin, H., Moon, S.J., *et al.*, 2008. Novel Jeff$=1/2$ Mott state induced by relativistic spin-orbit coupling in Sr_2IrO_4. Physical Review Letters 101 (7), 076402.

Kim, B.J., Ohsumi, H., Komesu, T., *et al.*, 2009. Phase-sensitive observation of a spin-orbital Mott state in Sr2IrO4. Science 323 (5919), 1329–1332.

Kim, D.H., Lim, D., 2015. Optical second-harmonic generation in few-layer $MoSe_2$. Journal of the Korean Physical Society 66 (5), 816–820.

Kirilyuk, A., Rasing, T., 2005. Magnetization-induced-second-harmonic generation from surfaces and interfaces. JOSA B 22 (1), 148–167.
Koshihara, S.-y., Adachi, S.-i., 2006. Photo-induced phase transition in an electron–lattice correlated system – future role of a time-resolved X-ray measurement for materials science. Journal of the physical Society of Japan 75 (1), 011005.
Kumar, A., Rai, R.C., Podraza, N.J., *et al.*, 2008. Linear and nonlinear optical properties of $BiFeO_3$. Applied Physics Letters 92 (12), 121915.
Kumar, N., Najmaei, S., Cui, Q., *et al.*, 2013. Second harmonic microscopy of monolayer MoS_2. Physical Review B 87 (16), 161403.
Kung, Y.F., Chen, C.C., Moritz, B., *et al.*, 2014. Numerical exploration of spontaneous broken symmetries in multiorbital Hubbard models. Physical Review B 90 (22), 224507.
Kuramoto, Y., Kusunose, H., Kiss, A., 2009. Multipole orders and fluctuations in strongly correlated electron systems. Journal of the Physical Society of Japan 78 (7), 072001.
Lafrentz, M., Brunne, D., Kaminski, B., *et al.*, 2012. Optical third harmonic generation in the magnetic semiconductor EuSe. Physical Review B 85 (3), 035206.
Landau, L., 1936. The theory of phase transitions. Nature 138, 840–841.
Landau, L.D., 1937. On the theory of phase transitions. I. Zhurnal Eksperimentalnoi Teoreticheskoi Fiziki 11, 19.
Li, Y., Rao, Y., Mak, K.F., *et al.*, 2013. Probing symmetry properties of few-layer MoS_2 and h-BN by optical second-harmonic generation. Nano Letters 13 (7), 3329–3333.
Lovesey, S.W., Khalyavin, D.D., Staub, U., 2015. Ferro-type order of magneto-electric quadrupoles as an order-parameter for the pseudo-gap phase of a cuprate superconductor. Journal of Physics: Condensed Matter 27 (29), 292201.
Lubashevsky, Y., Pan, L., Kirzhner, T., *et al.*, 2014. Optical birefringence and dichroism of cuprate superconductors in the THz regime. Physical Review Letters 112 (14), 147001.
Malard, L.M., Alencar, T.V., Barboza, A.P.M., *et al.*, 2013. Observation of intense second harmonic generation from MoS_2 atomic crystals. Physical Review B 87 (20), 201401.
Mangin-Thro, L., Sidis, Y., Wildes, A., Bourges, P., 2015. Intra-unit-cell magnetic correlations near optimal doping in $YBa_2Cu_3O_{6.85}$ Nature Communications 6., 7705.
McIver, J.W., Hsieh, D., Drapcho, S.G., *et al.*, 2012. Theoretical and experimental study of second harmonic generation from the surface of the topological insulator Bi_2Se_3. Physical Review B 86 (3), 035327.
Mook, H.A., Sidis, Y., Fauque, B., *et al.*, 2008. Observation of magnetic order in a superconducting $YBa_2Cu_3O_{6.6}$ single crystal using polarized neutron scattering. Physical Review B 78 (2), 020506.
Moskvin, A.S., 2012. Pseudogap phase in cuprates: Oxygen orbital moments instead of circulating currents. JETP Letters 96 (6), 385–390.
Murakami, Y., Hill, J.P., Gibbs, D., *et al.*, 1998. Resonant X-ray scattering from orbital ordering in $LaMnO_3$. Physical Review Letters 81 (3), 582.
Mydosh, J.A., Oppeneer, P.M., 2011. Colloquium: Hidden order, superconductivity, and magnetism: The unsolved case of URu_2Si_2. Reviews of Modern Physics 83 (4), 1301.
Mydosh, J.A., Oppeneer, P.M., 2014. Hidden order behaviour in URu_2Si_2 (A critical review of the status of hidden order in 2014). Philosophical Magazine 94 (32–33), 3642–3662.
Nasu, K., 2004. Photoinduced phase transitions. Singapore: World Scientific.
Nye, J.F., 1985. Physical properties of crystals: Their representation by tensors and matrices. Oxford: Oxford university press.
Nývlt, M., Bisio, F., Kirschner, J., 2008. Second harmonic generation study of the antiferromagnetic NiO (001) surface. Physical Review B 77 (1), 014435.
Orenstein, J., 2011. Optical nonreciprocity in magnetic structures related to high-T_c superconductors. Physical Review Letters 107 (6), 067002.
Pan, R.-P., Wei, H.D., Shen, Y.R., 1989. Optical second-harmonic generation from magnetized surfaces. Physical Review B 39 (2), 1229.
Petersen, J.C., Caswell, M.D., Dodge, J.S., *et al.*, 2006. Nonlinear optical signatures of the tensor order in $Cd_2Re_2O_7$. Nature Physics 2 (9), 605–608.
Qi, T.F., Korneta, O.B., Li, L., *et al.*, 2012. Spin-orbit tuned metal-insulator transitions in single-crystal $Sr_2Ir_{1-x}Rh_xO_4$ ($0 \leq x \leq 1$). Physical Review B 86 (12), 125105.
Rabe, K.M., 2007. Solid-state physics: Response with a twist. Nature 449 (7163), 674–675.
Reif, J., Zink, J.C., Schneider, C.M., *et al.*, 1991. Effects of surface magnetism on optical second harmonic generation. Physical Review Letters 67 (20), 2878.
Santini, P., Carretta, S., Amoretti, G., *et al.*, 2009. Multipolar interactions in f-electron systems: The paradigm of actinide dioxides. Reviews of Modern Physics 81 (2), 807.
Seyler, K.L., Schaibley, J.R., Gong, P., *et al.*, 2015. Electrical control of second-harmonic generation in a WSe_2 monolayer transistor. Nature Nanotechnology 10 (5), 407–411.
Shah, N., Chandra, P., Coleman, P., Mydosh, J.A., 2000. Hidden order in URu_2Si_2. Physical Review B 61 (1), 564.
Shaked, H., 1994. Crystal structures of the high-T_c superconducting copper-oxides. Elsevier Science Publishers BV.
Shannon, V.L., Koos, D.A., Richmond, G.L., 1987a. Changes in the second harmonic rotational anisotropy for a silver (111) electrode as a function of bias potential. Journal of Physical Chemistry 91 (22), 5548–5551.
Shannon, V.L., Koos, D.A., Richmond, G.L., 1987b. Second harmonic generation for in situ analysis of electrode surface structure. Applied Optics 26 (17), 3579–3583.
Shekhter, A., Ramshaw, B.J., Liang, R., *et al.*, 2013. Bounding the pseudogap with a line of phase transitions in $YBa_2Cu_3O_6 + [dgr]$. Nature 498 (7452), 75–77.
Shen, Y.-R., 1984. Principles of nonlinear optics.
Shen, Y.R., 1989. Optical second harmonic generation at interfaces. Annual Review of Physical Chemistry 40 (1), 327–350.
Simon, M.E., Varma, C.M., 2002. Detection and implications of a time-reversal breaking state in underdoped cuprates. Physical Review Letters 89 (24), 247003.
Simon, M.E., Varma, C.M., 2003. Symmetry considerations for the detection of second-harmonic generation in cuprates in the pseudogap phase. Physical Review B 67 (5), 054511.
Spaldin, N.A., Fiebig, M., Mostovoy, M., 2008. The toroidal moment in condensed-matter physics and its relation to the magnetoelectric effect. Journal of Physics: Condensed Matter 20 (43), 434203. (All authors contributed equally to this work).
Squires, G.L., 2012. Introduction to the theory of thermal neutron scattering. Cambridge: Cambridge University Press.
Tokura, Y., 2006. Photoinduced phase transition: A tool for generating a hidden state of matter. Journal of the Physical Society of Japan 75 (1), 011001.
Tom, H.W.K., Heinz, T.F., Shen, Y.R., 1983. Second-harmonic reflection from silicon surfaces and its relation to structural symmetry. Physical Review Letters 51 (21),
Torchinsky, D.H., Chu, H., Qi, T., *et al.*, 2014. A low temperature nonlinear optical rotational anisotropy spectrometer for the determination of crystallographic and electronic symmetries. Review of Scientific Instruments 85 (8), 083102.
Torchinsky, D.H., Chu, H., Zhao, L., *et al.*, 2015. Structural distortion-induced magnetoelastic locking in Sr_2IrO_4 revealed through nonlinear optical harmonic generation. Physical Review Letters 114 (9), 096404.
Van Aken, B.B., Rivera, J-P., Schmid, H., Fiebig, M., 2007. Observation of ferrotoroidic domains. Nature 449 (7163), 702–705.
Varma, C.M., 1997. Non-Fermi-liquid states and pairing instability of a general model of copper oxide metals. Physical Review B 55 (21), 14554.
Warren, B.E., 1969. X-ray Diffraction. Chelmsford, MA: Courier Corporation.
Weber, C., Lauchli, A., Mila, F., Giamarchi, T., 2009. Orbital currents in extended Hubbard models of high-T_c cuprate superconductors. Physical Review Letters 102 (1), 017005.
Weller, M.T., Hughes, R.W., Rooke, J., Knee, C.S., Reading, J., 2004. The pyrochlore family – a potential panacea for the frustrated perovskite chemist. Dalton Transactions 19, 3032–3041.
Witczak-Krempa, W., Chen, G., Kim, Y. B., Balents, L., 2014. Correlated quantum phenomena in the strong spin-orbit regime.Annual Review of Condensed Matter Physics 5 (1), 57.
Wu, L., Patankar, S., Morimoto, T., *et al.*, 2016. Giant anisotropic nonlinear optical response in transition metal monopnictide Weyl semimetals. Nature Physics 13, 350–355. (Advance online publication).
Wu, W., Wang, L., Li, Y., *et al.*, 2014. Piezoelectricity of single-atomic-layer MoS_2 for energy conversion and piezotronics. Nature 514 (7523), 470–474.
Yakovenko, V.M., 2015. Tilted loop currents in cuprate superconductors. Physica B: Condensed Matter 460, 159–164.

Yamada, C., Kimura, T., 1993. Anisotropy in second-harmonic generation from reconstructed surfaces of Gaas. Physical Review Letters 70 (15), 2344.
Yamaura, J.-I., Hiroi, Z., 2002. Low temperature symmetry of pyrochlore oxide $Cd_2Re_2O_7$. Journal of the Physical Society of Japan 71 (11), 2598–2600.
Yin, X., Ye, Z., Chenet, D.A., *et al*., 2014. Edge nonlinear optics on a MoS_2 atomic monolayer. Science 344 (6183), 488–490.
Zhang, J., Tan, X., Liu, M., *et al*., 2015. Cooperative photoinduced metastable phase control in strained manganite films. Nature Physics 15, 956.
Zhao. L., Belvin, C.A., Liang, R., *et al*., 2016a. A global inversion-symmetry-broken phase inside the pseudogap region of $YBa_2Cu_3O_y$. Nature Physics 13 (3), 250–250.
Zhao. L., Torchinsky, D., Chu, H., *et al*., 2016b. Evidence of an odd-parity hidden order in a spin-orbit coupled correlated iridate. Nature Physics 12 (1), 32–36.
Zimmermann, A.S., Meier, D., Fiebig, M., 2014. Ferroic nature of magnetic toroidal order. Nature Communications 5. 4796, 2014.
Zou, X., Hovmöller, S., Oleynikov, P., 2011. Electron Crystallography: Electron Microscopy and Electron Diffraction, 16. Oxford: Oxford University Press.

Saturated Absorption Spectroscopy for Diode Laser Locking

Bachana Lomsadze, University of Michigan, Ann Arbor, MI, United States

Introduction

Since their development about three decades ago, external-cavity diode lasers (ECDLs) have become very attractive coherent light sources for many applications. They are typically used for atomic and molecular spectroscopy, laser cooling and trapping, optical telecommunications, etc. The features that make ECDLs attractive are: their narrow linewidth, broadly tunable wavelength, low cost, compactness, and relative ease of construction.

Using external-cavity feedback configurations, one is able to achieve laser linewidths of hundreds of kHz. However, temperature fluctuations, vibrations, electronic noise, change in atmospheric pressure, etc., will broaden the laser linewidth and shift the center frequency. For some applications, for example, cooling and trapping, a drift of the center frequency of a few MHz for even a short period of time can be problematic. For these applications laser frequency locking is required.

Over the years, multiple laser frequency locking techniques have been developed. One of the most commonly used methods is the Pound–Drever–Hall (PDH) method which locks the frequency of a laser to a stable Fabry–Perot cavity reference. Alternatively, if the laser frequency must be locked to some atomic resonance then a setup using saturated absorption spectroscopy can be employed. This method has many different variations itself.

In this article the technique of locking a laser to a Rb hyperfine transition using saturated absorption spectroscopy with a room temperature Rb atomic vapor cell as a reference will be discussed. But, the method described here can be used with any reference cell.

This article is divided into four sections in which (1) the external-cavity feedback configurations are reviewed, (2) discusses the principles of saturation absorption spectroscopy, (3) shows the experimental arrangement used for laser locking, and (4) concludes the article.

External-Cavity Configurations

Two ECDL configurations are widely used today: the Littrow and Littman–Metcalf configurations. Simplified schematic diagrams of these configurations are shown in Fig. 1. They consist of a diode laser (injection current diode) with a highly reflective ($>90\%$) rear facet and antireflection coated front facet, a collimating lens, and a dispersive element for wavelength tuning (typically a diffraction grating). In both configurations the 1st-order diffracted beam is used for laser wavelength tuning.

For the Littrow configuration (Fig. 1(a)), the 1st-order diffracted beam from the grating is retro-reflected back into the laser diode for feedback and the laser wavelength tuning is accomplished by rotating the grating. For some applications, for example, when broad wavelength tuning is required, this configuration is not preferred because a change in the grating orientation causes a change in the laser output beam (0th order) pointing. However, this problem can be resolved with a pointing stabilization feedback arrangement.

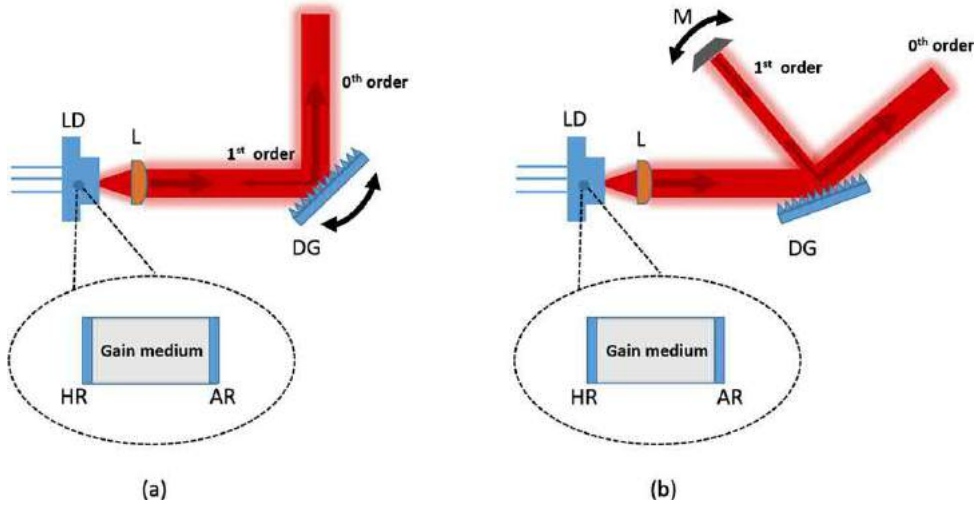

Fig. 1 Schematic diagram of Littrow (a) and Littman–Metcalf (b) external-cavity configurations. AR, antireflective coated facet; DG, diffraction grating; HR, high reflective facet; L, collimating Lens; LD, laser diode; M, mirror.

For the Littman–Metcalf configuration (Fig. 1(b)), the grating position is fixed and the wavelength is tuned by rotating a mirror. This configuration usually generates an output beam with a narrower linewidth because the 1st order beam refracts twice from the grating.

Principles of Saturated Absorption Spectroscopy

Fig. 2 shows the hyperfine energy level diagrams of Rubidium atoms' D_2 lines for both isotopes (^{87}Rb and ^{85}Rb).

The natural linewidth of the hyperfine lines (full-width at half maximum) is about 6 MHz. However, the measured transmission spectrum of a room temperature Rb vapor cell shows broadening of these lines up to 500 MHz (see Fig. 3) which is due to the velocity distribution of atoms in the reference cell and the Doppler Effect. This data was obtained by measuring the trans-

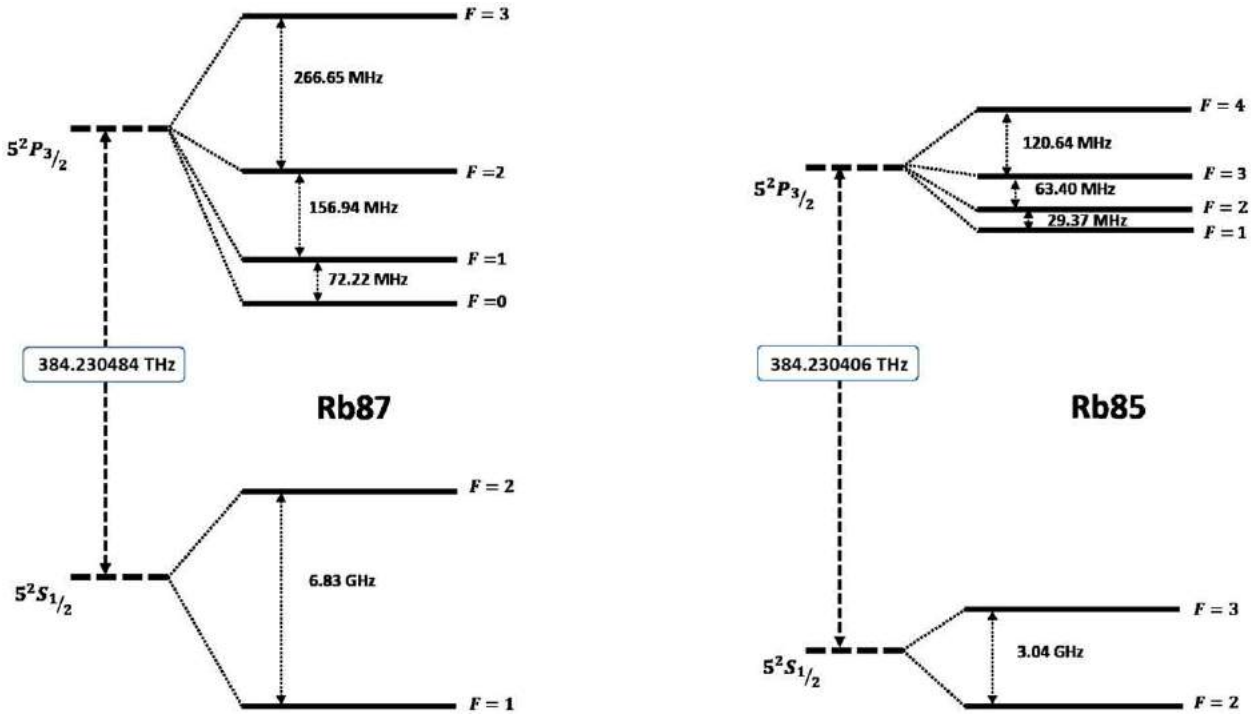

Fig. 2 Hyperfine energy level diagram for ^{87}Rb (left) and ^{85}Rb (right) D_2 lines (not to scale).

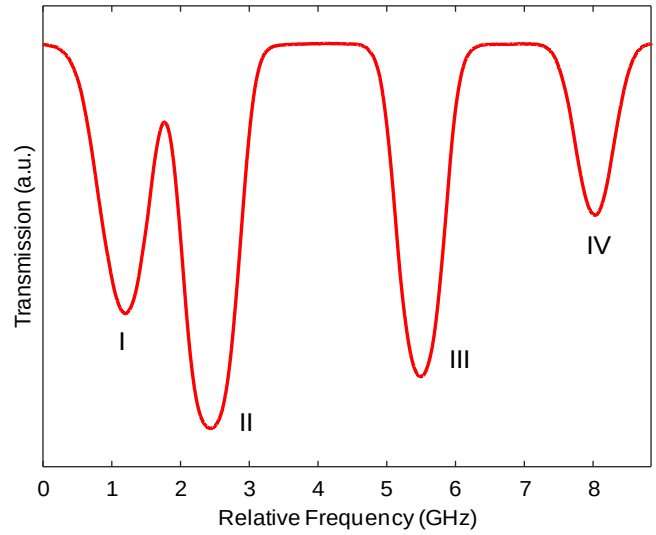

Fig. 3 Transmission profile of a room temperature Rb cell near 780.24 nm.

mission of a beam of laser light from an ECDL through the vapor cell while scanning its center frequency over a 9 GHz range centered near 780.24 nm (384.23 THz).

I and IV dips correspond to ^{87}Rb transitions from $5S_{1/2}$ F=2 and $5S_{1/2}$ F=1 to $5P_{3/2}$ states, respectively. II and III dips correspond to ^{85}Rb transitions from $5S_{1/2}$ F=3 and $5S_{1/2}$ F=2 to $5P_{3/2}$ states, respectively. Obviously, none of these broad absorption dips can be used to lock a laser onto a hyperfine resonance line.

One way to eliminate Doppler broadening is to cool the atoms down to hundreds of microkelvin temperatures using a complex optical cooling and trapping experimental setup.

Doppler Free Spectroscopy

Doppler broadening can also be eliminated without having to cool the atoms by using the simple arrangement shown in Fig. 4. Both strong (pump) and weak (probe) beams are generated from the same laser source (hence they have the same wavelength) and counter-propagate in the Rb cell.

The measured transmission spectrum of the probe beam after overlapping with the pump beam in the reference cell is shown in Fig. 5(I). The structure shown represent all dipole allowed transitions between hyperfine energy levels of both Rb isotopes. The spectrum, showing only the $5S_{1/2}$ (F=2) to $5P_{3/2}$ (all allowed hyperfine states) transitions for ^{87}Rb is plotted in Fig. 5(II). The elimination of Doppler broadening by counter propagating beams can be explained using a simple 2 level system.

Fig. 6(a) shows a two level system and the number density function of ground state Rb atoms as a function of their velocities (Maxwell–Boltzmann distribution) along the probe beam propagation axis at room temperature. Positive and negative velocities correspond to atoms moving in the same and opposite directions as the probe beam, respectively. When the laser frequency is detuned from the resonance frequency ($\omega_L < \omega$ or $\omega_L > \omega$), then both pump and probe beams interact with the same magnitude but the opposite sign velocity group atoms (due to the Doppler Effect). This interaction brings atoms from the ground state ($|g\rangle$) into the excited state ($|e\rangle$) and hence modifies the density function (Fig. 6(b) and (c)). Clearly the effect from the pump beam is stronger because of its high intensity. This effect is called the "hole burning." When the laser frequency is close to the atomic

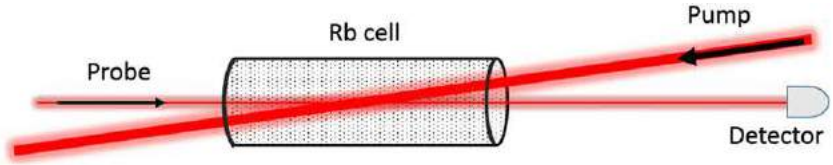

Fig. 4 Schematic diagram of counter propagating pump and probe beams overlapped inside the Rb cell.

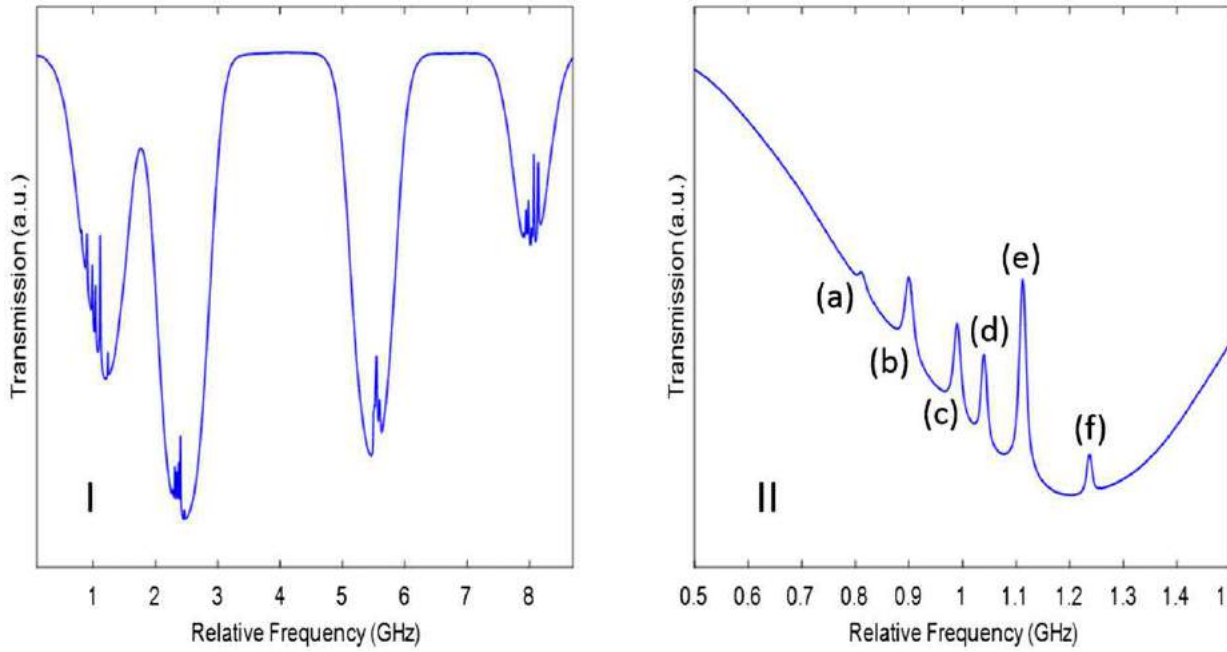

Fig. 5 (I) Transmission profile of the probe beam overlapped with the pump beam in a room temperature Rb reference cell. (II) Hyperfine structure of the $5S_{1/2}$ to $5P_{3/2}$ transitions for ^{87}Rb isotope (a) F=2 to F=1, (c) F=2 to F=2, (f) F=2 to F=3, (b) "crossover" resonance between (a) and (c), (d) "crossover" resonance between (a) and (f), (e) "crossover" resonance between (c) and (f).

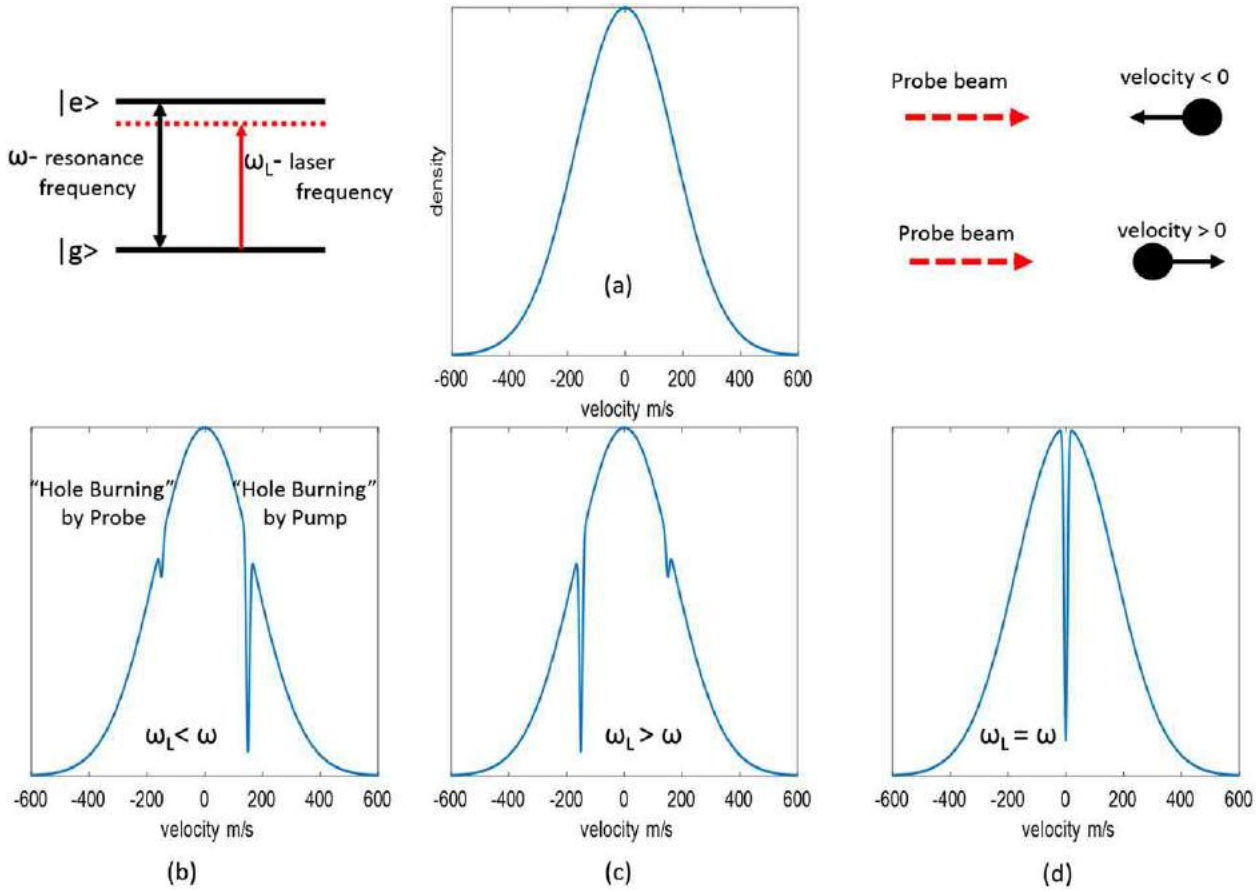

Fig. 6 (a) Two level system and the number density function of ground state Rb atoms as a function of their velocity. (b) "Hole burning" when the laser frequency is less than the resonance frequency. (c) Laser frequency is higher than the resonance frequency. (d) Laser frequency is equal to the resonance frequency.

resonance frequency, then both pump and probe beams interact with the zero-velocity group of atoms (Fig. 6(d)). In this case the strong pump beam depletes the ground state population and will even saturate the absorption if intense enough. Consequently, the probe beam experiences modified (decreased) absorption. The hyperfine lines shown in (a), (c), and (f) on Fig. 5 represent decreased absorption of the probe beam in the presence of the strong pump beam.

Crossover Resonances

Fig. 5 also shows additional peaks (b), (d) and (e) that do not directly correspond to any atomic resonances. These peaks are referred to as "crossover" resonances. They emerge when a system has multiple excited states (e.g., Rb atoms – see Fig. 2) that are within the Doppler width and share the same ground state. For simplicity I will consider a 3 level "V" system that has one ground state ($|g\rangle$) and two excited states ($|e1\rangle$ and $|e2\rangle$) such as that shown in Fig. 7. When the laser frequency is in between these two resonances, then there are two different velocity groups of atoms that the pump beam interacts with: one that is redshifted (dip 2 which corresponds to $|g\rangle$–$|e1\rangle$ transition) and one that is blue shifted (dip 4 that corresponds to $|g\rangle$–$|e2\rangle$ transition). At the same time the probe beam interacts with exactly the opposite sign velocity groups of atoms. In Fig. 7 they correspond to dip 1 ($|g\rangle$–$|e2\rangle$ transition) and dip 3 ($|g\rangle$–$|e1\rangle$ transition). When the laser frequency is exactly halfway between these two resonances, the absorption dips from pump and probe beams overlap (1 with 2 and 3 with 4). In the presence of the strong pump field this causes reduction of the probe beam absorption. The peaks (b), (d), and (e) in Fig. 5 correspond to all possible crossover resonances (please see the caption for details).

Experimental Setup

Schematic diagram of an experimental setup that is typically used for laser locking is shown in Fig. 8.

A beam from an ECDL is divided using a polarizing beam splitter. Most of the laser power is sent to an experiment (e.g., cooling and trapping) whereas only a small portion (less than 1 mW) is used for generating a correction signal for laser

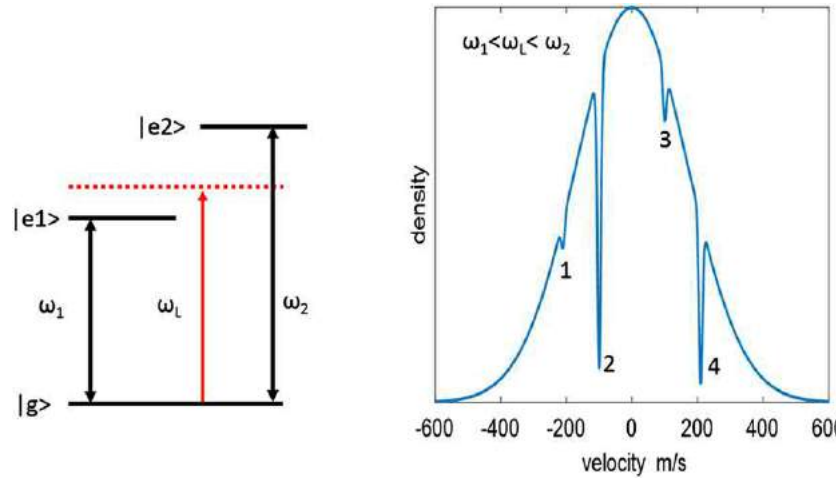

Fig. 7 Three level "**V**" system. "Hole burning" when the laser frequency is in between two resonance frequencies.

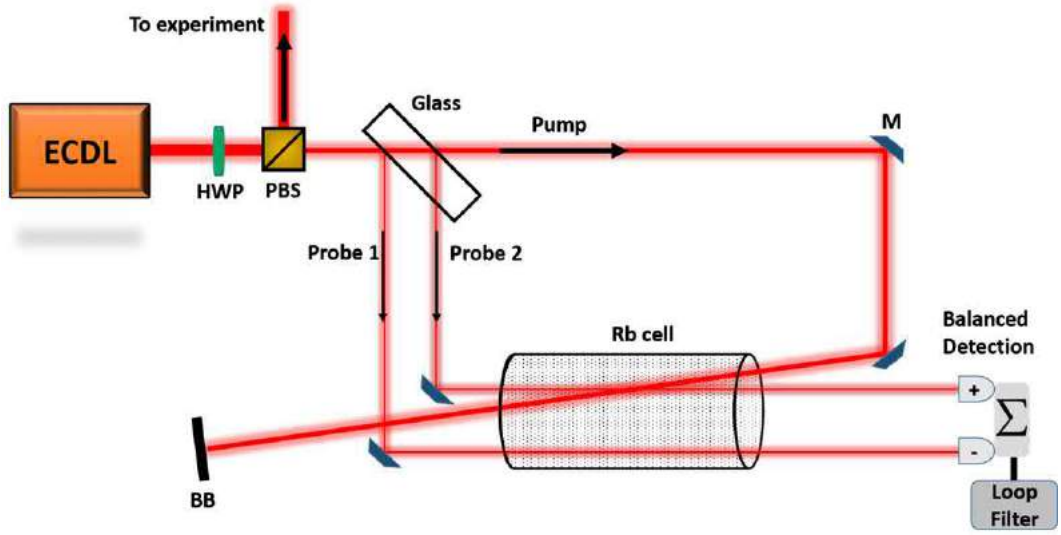

Fig. 8 Experimental setup for saturated absorption spectroscopy. BB, beam block; ECDL, external cavity diode laser; HWP, half wave plate; M, mirror; PBS, polarizing beam splitter.

locking. Using a thick glass this beam is split again into 3 different beams: 2 probe (weak) beams reflected from the front and rear surfaces of the glass and one (strong) transmitted pump beam. The pump beam is overlapped with one of the probe beams in the reference cell and both probe beams are monitored using a balanced photodetector. The signal from the balanced detector is sent to a loop filter that generates a correction signal for the laser.

Balanced detection has the following advantages:

1. As shown in **Fig. 6(a)** the measured transmission profile of the probe beam overlapped by the pump beam has the Doppler broadened background. In **Fig. 8** that corresponds to the transmission of the probe 2 beam. The transmission profile of the probe 1 beam is shown in **Fig. 3** and is pure Doppler broadened. Balanced detection takes the difference between these spectrums and the resulting spectrum is Doppler background free (**Fig. 9**).
2. Balanced detection helps to eliminate any intensity fluctuations that the ECDL might have. This is especially important if, for example, the laser must be locked on the side of one of the hyperfine lines. Without the balanced detection, any unwanted intensity fluctuation would cause the transmission spectrum and hence the correction signal to fluctuate as well. This would result in broadening of the laser frequency.

If the laser frequency must be locked to the peak of an absorption line then current or phase modulation techniques can be used. Using these methods one obtains a spectrum that is the derivative of the saturated absorption spectrum in which zero crossings correspond to the peaks of the absorption lines.

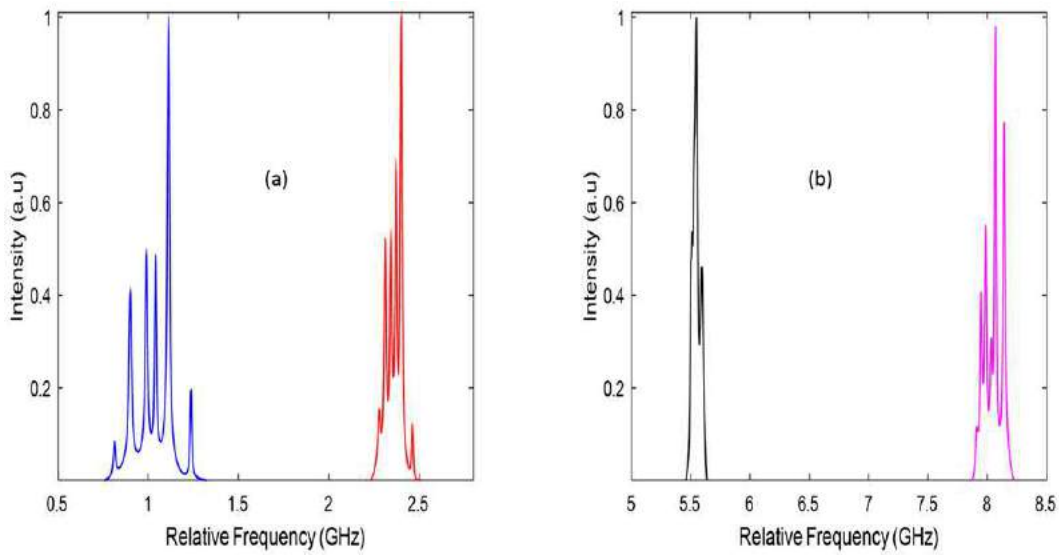

Fig. 9 Doppler background free saturated absorption spectra of Rb atoms. (a) 0–2.7 GHz range. (b) 5–8.5 GHz range.

Conclusions

In this article the use of saturated absorption spectroscopy for laser locking was reviewed. I have discussed the basic principles of saturation absorption spectroscopy and reviewed how room temperature reference cells can provide Doppler free absorption lines. I have explained the appearance of crossover resonances in multilevel systems and reviewed an experimental technique that is commonly used for generating a correction signal for laser locking.

Acknowledgment

I would like to thank Brad Smith for obtaining the saturated absorption spectra.

Further Reading

Arnold, A.S., Wilson, J.S., Boshier, M.G., 1998. A simple extended-cavity diode laser. Review of Scientific Instruments 69 (3). doi:10.1063/1.1148756.
Black, E.D., 2001. An introduction to Pound–Drever–Hall laser frequency stabilization. American Journal of Physics 69 (1). doi:10.1119/1.1286663.
MacAdam, K.B., Steinbach, A., Wieman, C., 1992. A narrow-band tunable diode laser system with grating feedback and a saturated absorption spectrometer for Cs and Rb. American Journal of Physics 60 (12). doi:10.1119/1.16955.

Attosecond Spectroscopy

Agnieszka Jaroń-Becker and Andreas Becker, University of Colorado, Boulder, CO, United States

Introduction

The quest for studying dynamics in matter on ultrashort timescales has driven the development of a variety of technologies in ultrafast science. The most prominent among these are ultrashort electron pulses, femtosecond laser pulses, X-ray free electron lasers and, currently the shortest off all these probes, attosecond light pulses. The atomic unit of time is 24.2 as$=24.2\times10^{-18}$ s, corresponding to the period of an electron circling in the first Bohr orbit divided by 2π. Attosecond pulse technology therefore allows for resolution of dynamical quantum processes on the timescales of electron motion in atoms and other forms of matter, as compared to the observation of atomic motion using femtosecond laser pulses.

The goal of this brief review is to first present the concepts and physical understanding behind the generation of attosecond optical pulses via the currently most common source, the highly nonlinear optical high harmonic generation process driven by strong laser pulses. In the second part of the review a number of spectroscopic techniques and applications for the temporal resolution on the attosecond timescale are discussed. For a more detailed presentation of the topics we refer to a number of recent books and review articles (Kapteyn *et al.*, 2005; Corkum and Krausz, 2007; Krausz and Ivanov, 2009; Chang, 2011; Popmintchev *et al.*, 2011; Gallmann *et al.*, 2012; Dahlströhm *et al.*, 2012; Vrakking, 2014; Leone *et al.*, 2014; Peng *et al.*, 2015; Pazourek *et al.*, 2015; Ramasesha *et al.*, 2016; Calegari *et al.*, 2006; Wu *et al.*, 2016; Leone and Neumark, 2016) as well as the original work cited therein.

In this review we use, if not stated differently, Hartree atomic units, which is an unit set particularly useful in atomic, molecular and optical physics. In this unit set the values for the electron mass m_e, the elementary charge e, Planck's constant $\hbar$; and the Coulomb's constant $k_e=1/4\pi\varepsilon_0$ are unity, i.e., $m_e=e=\hbar=k_e=1$.

Generation of Attosecond Optical Pulses

The progress of ultrashort laser pulse technology since the invention of mode-locking is summarized in **Fig. 1**. After a continuous decrease of the pulse duration until the mid 1980s there was only a marginal progress for more than a decade. In the late 1980s, femtosecond laser pulses with a duration of a few cycles and high peak intensities became available due to techniques for pulse compression and chirped pulse amplification. Such pulses approached a single optical cycle with an optical period around 2.7 fs at a central wavelength of 800 nm.

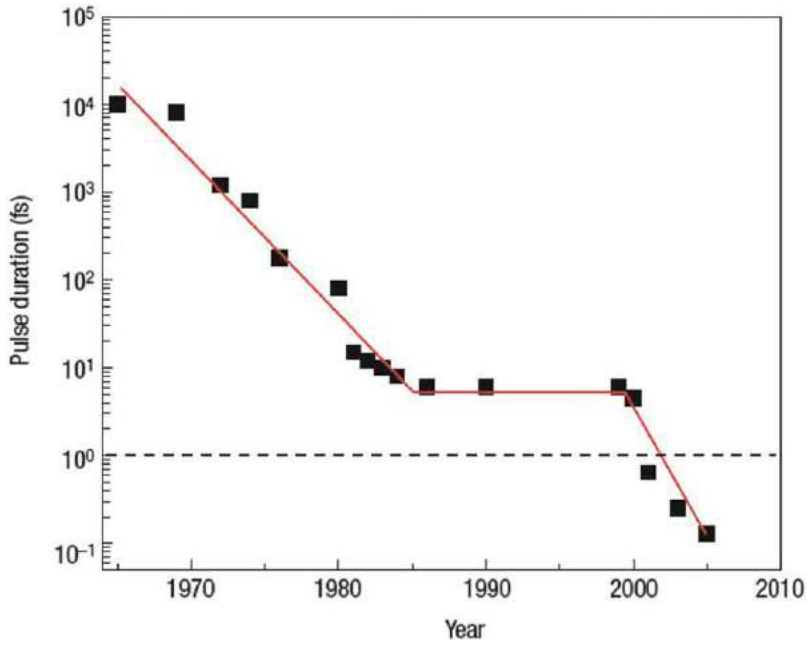

Fig. 1 Development of the laser pulse duration (in femtoseconds): The minimum duration first continuously decreased until the mid of the 1980s. Following marginal progress for more than a decade, the femtosecond barrier (dashed line) was broken in 2001. Adapted with permission from Mcmillan Publishers Ltd., Corkum, P.B., Krausz, F., 2007. Attosecond science. Nat. Phys. 3, 381, copyright (2007).

The basic idea towards engineering sub-femtosecond pulses, proposed in the early 1990s and first realized in 2001, is similar to the principle of mode-locking used for femtosecond lasers. If a range of phase-locked frequencies is combined via Fourier synthesis, the temporal profile is a sequence of pulses with a duration inverse proportional to the number of frequencies involved in the synthesis. Consequently, the generation of attosecond optical pulses relies on the availability of a comb of coherent light at equidistant frequencies in the ultraviolet, extreme ultraviolet or even the soft X-ray spectral regime. Before reviewing the physics underlying high-order harmonic generation as the most common source of attosecond pulse generation we briefly recap basic concepts and mathematical expressions used for the description of ultrashort laser pulses via their electric fields.

Electric Laser Field in Time and Frequency Domain

High-order harmonic and attosecond pulse generation is primarily focused on linearly polarized radiation. Since the wavelength of the involved radiation is, in general, much larger than the size of the target, typically an atom, the electric dipole approximation can be used and a linearly polarized pulse can be expressed by its time-dependent electric field:

$$\mathbf{E}(t) = \varepsilon_0(t)\cos(\omega_0 t + \phi_{CE})\hat{\varepsilon} \tag{1}$$

where $\varepsilon_0(t)$ is the time-dependent envelope function, ω_0 the central angular frequency, related to the central wavelength $\lambda_0 = 2\pi c/\omega_0$, $(\varphi)_{CE}$ is the carrier-(to-)envelope phase of the electric field and $\hat{\varepsilon}$ represents the polarization direction of the field.

For Gaussian pulses the envelope function is given by

$$\varepsilon_0(t) = \varepsilon_0 \exp\left[-2\ln 2\left(\frac{t^2}{\Delta t^2}\right)\right] \tag{2}$$

where ε_0 is the peak amplitude. The intensity profile of the pulse is then

$$I(t) = I_0 \exp\left[-4\ln 2\left(\frac{t^2}{\Delta t^2}\right)\right] \tag{3}$$

with I_0 is the peak intensity and Δt is the full-width-half-maximum (FWHM) pulse duration. Via Fourier transform the electric field of a laser pulse can be also described in the frequency domain. For a Gaussian pulse in the time domain the spectral amplitude is also a Gaussian function:

$$\tilde{\varepsilon}(\omega) = \tilde{\varepsilon}_0 \exp\left[-2\ln 2\left(\frac{(\omega-\omega_0)^2}{\Delta\omega^2}\right)\right] \tag{4}$$

with full-width half-maximum bandwidth $\Delta\omega$.

High-Order Harmonic Generation

The currently most commonly used method for the generation of coherent light in form of attosecond pulses is the physical process of high-order harmonic generation. Harmonic generation is a nonlinear frequency up conversion effect occurring in the interaction of laser fields with matter. Observation of the second harmonic of laser light in 1961, immediately after the invention of the laser, marked the emergence of the field of perturbative nonlinear optics. In the perturbative interaction regime, the intensities of generated higher harmonics decrease rapidly due to the weak perturbation of matter by the laser field.

The generation of high-order harmonics was discovered in 1987 via the observation of a radiation spectrum characterized by a large number of odd harmonics of the fundamental driving laser frequency with intensities that did not decrease significantly, as it was expected in perturbative nonlinear optics. In the respective experiments, atoms interacted with femtosecond laser pulses having peak intensities of the order of $10^{13} - 10^{14}$ W cm^{-2}. At these intensities the electric field strength of the laser pulse becomes comparable to those of the fields due to the Coulomb interaction between charged particles in atoms and molecules. Consequently, the electric field cannot be considered as a perturbation anymore. In such intense laser pulses, matter is ionized even if the photon energy of the field is only a small fraction of the ionization potential. This strong-field regime of laser-matter interaction is restricted by an upper limit at intensities leading to the onset of relativistic effects and those due to the influence of the magnetic field of the laser.

Microscopic description

The power spectrum of the radiation emitted by an accelerating charge is proportional to the magnitude of the acceleration. Assuming that the electric field of a strong laser pulse interacts with a single electron in an atom, a time-dependent dipole is created by the electron wavepacket and the parent ion given by

$$\mathrm{d}(t) = \langle \Psi(\mathbf{r};t)|\mathbf{r}|\Psi(\mathbf{r};t)\rangle \tag{5}$$

where $\Psi(\mathbf{r};t)$ represents the electron wave function, resulting from the solution of the nonrelativisitc Schrödinger equation describing the interaction between the atom (in the single-active-electron approximation) with the electric field of an intense laser

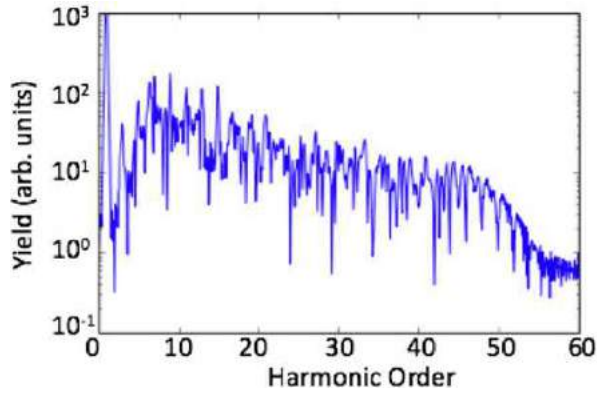

Fig. 2 Exemplary result of numerical simulations of the single-atom high harmonic spectrum for atomic hydrogen interacting with a laser pulse with central wavelength $\lambda_0 = 800$ nm and peak intensity $I_0 = 3 \times 10^{14}$ W cm^{-2}. From Miller, M.R., 2016. PhD thesis, JILA, University of Colorado, Boulder.

pulse, and **r** is the space vector. The dipole acceleration can be either determined via the second time derivative of the dipole

$$\mathbf{a}(t) = \frac{d^2}{\mathrm{d}t^2}\langle \Psi(\mathbf{r};t)|\mathbf{r}|\Psi(\mathbf{r};t)\rangle \tag{6}$$

or using the Ehrenfest theorem as

$$\mathrm{a}(t) = \langle \Psi(\mathbf{r};t)| - \nabla V(\mathbf{r}) + \mathbf{E}(t)|\Psi(\mathbf{r};t)\rangle \tag{7}$$

where $V(\mathbf{r})$ is the Coulomb potential between the propagating electron and the parent ion. The high-order harmonic spectrum is then obtained as the Fourier transform of the dipole acceleration.

An example of a high-order harmonic spectrum is shown in Fig. 2 for a hydrogen atom driven by a linearly polarized electric field at a central wavelength of 800 nm and a peak intensity of 3×10^{14} W cm^{-2}. The structure of the radiation spectrum possesses the following characteristic features of high-order harmonic generation: strong emission at the photon energy of the driving field, followed by a quick drop off due the exponential decay for perturbative low-order harmonics, subsequent plateau of harmonic generation at constant strength and rapid decrease of radiation efficiency beyond a cut-off energy.

Three-step model of HHG

An intuitive physical picture of HHG emerges from an analysis known as the three-step model, depicted graphically in Fig. 3(a). The strong electric field suppresses the Coulomb barrier binding the electron to the atomic core, which permits the tunneling of an electron wave packet. The wave packet then propagates in the oscillating electric field and, depending on the time of release, it can be steered back to the parent ion. With some probability, the wave packet can recombine with the parent ion to the ground state. The energy absorbed by the electron during the propagation is emitted in form of a high energy photon. In a quantum mechanical model description of HHG this picture is reflected via a product of three probability amplitudes, accounting for ionization, propagation and recombination of the electron wave packet.

Due to the quasiperiodic repetition of this process every half cycle in a multicycle laser field, the resulting dipole emission spectrum is discrete, consisting of odd harmonics of the central driver laser frequency. In an ultrashort few-cycle driver pulse, the discrete structure becomes less and less pronounced and eventually disappears in particular in the cut-off region of the spectrum. Predictions of this quantum mechanical model are in qualitative agreement, in particular concerning the form and extension of the plateau, with experimental observation and the results of ab-initio numerical calculations, based on Eq. (6).

Some basic features of the HHG process and spectrum can be understood from a simplified classical analysis, based on the Newton's equation in one dimension, for propagation of a point-like electron in an oscillating linearly polarized electric field:

$$\frac{d^2x}{dt^2} = -\varepsilon(t)x \tag{8}$$

with initial conditions $\dot{x} = 0$ and $x = 0$. Predictions of the corresponding classical trajectory analysis, emerging during a center cycle of the field used to generate the HHG spectrum shown in Fig. 2, are shown in Fig. 3(b). Each color corresponds to a different ionization time and represents trajectories that return to the parent ion and, hence, contribute to HHG; while gray lines indicate ionization events that do not lead to return of the electron. The classical analysis predicts a cutoff in the HHG spectrum at a maximum energy of

$$E_{\max} = I_p + 3.17U_p \approx I\lambda_0^2 \tag{9}$$

where I_p is the ionization potential of the atom, $U_p = I_0/4\omega_0^2$ is the ponderomotive potential. This prediction is in good agreement with experimental observations of HHG spectra in multicycle driver laser pulses, whereas for few-cycle pulses the cutoff can be shifted to slightly higher harmonics. The excursion distance for the classical trajectory corresponding to the cut-off harmonic

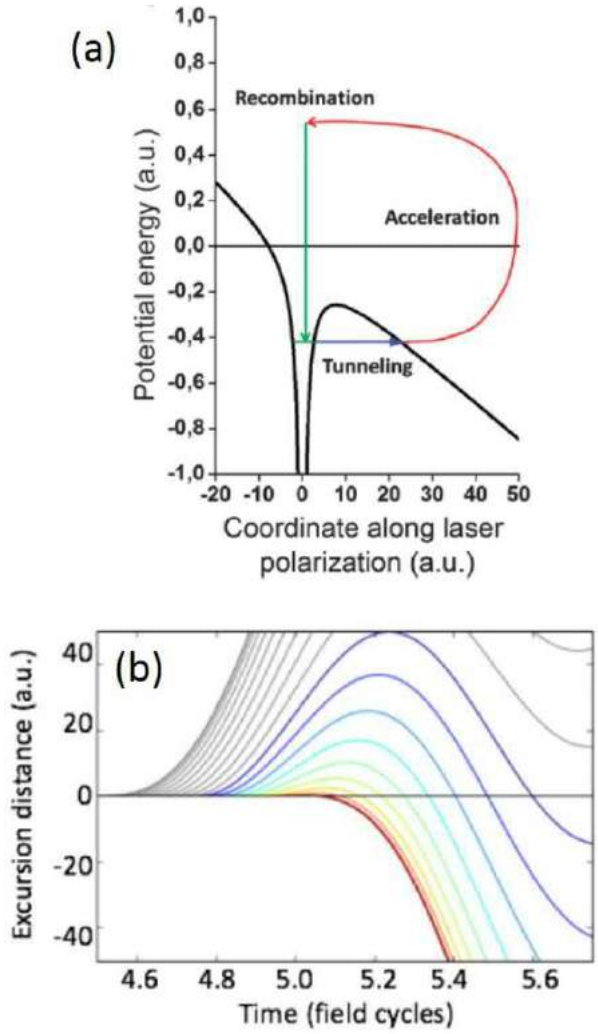

Fig. 3 (a) Conceptual picture of the three-step model of high harmonic generation (Reproduced from Vrakking, M.J.J., 2014. Attosecond imaging. Phys. Chem. Chem. Phys. 16, 2775 with permission of the PCCP Owner Societies.); (b) Classical trajectories of a point-like electron propagating in a strong laser field. Laser parameters are the same as for the HHG spectrum in **Fig. 2**. Reproduced from Miller, M.R., 2016. PhD thesis, JILA, University of Colorado, Boulder.

energy is given by the quiver radius $\alpha_0 = U_p/\omega_0^2$. Propagation of the electron via two different trajectories, following excursion distances either shorter or longer than the quiver radius, contribute to the harmonic generation process for each of the other photon energies.

The scaling of the harmonic cut-off energy with the intensity and the square of the driver wavelength (Eq. (9)) indicates that in principle the spectral bandwidth of the harmonic emission, which is important for the attosecond pulse generation (see Section Formation of Attosecond Pulses), can be extended by increasing the intensity or the wavelength of the driver laser. However, in practice HHG in a specific charge state of the single atom gets terminated by the depletion of the electron population in the ground state, which limits the applied laser intensity. Furthermore, the efficiency of the single-atom harmonic emission decreases rapidly with increase of the wavelength due to the diffusion of the electron wave packet along the longer and longer excursion distances in the continuum.

Macroscopic phase-matching aspects

In order to achieve efficient high harmonic emission and, hence, attosecond pulse generation the high harmonic signal from the many atoms in the generating medium must be phase matched. This means that the phase velocity of the driving laser field and the harmonic light fields must be matched in order to guarantee coherent addition of the harmonic emission from the individual atoms. Understanding of phase matching in the non-perturbative interaction regime is different to the concepts known from perturbative nonlinear optics. Besides factors arising due to the geometry of the HHG experiment and intrinsic atomic phases, the phase mismatch depends on the relation of neutral and ionic species in the focal region and also the pressure of the gaseous

medium. Accurate analysis of these effects recently enabled the extension of high harmonic generation up to the keV X-ray regime with ultraviolet and mid-infrared driving lasers.

Controlling polarization of high-order harmonics

High harmonic generation driven by an strong electric field is very sensitive to the ellipticity of the pump pulse, and the intensity of the harmonic signal decreases rapidly as the polarization moves away from linear. This can be easily understood within the classical picture, since in an elliptically polarized field the electron wave packet acquires a transverse component of the velocity, prohibiting its return to the core of the parent ion. Consequently, for research in HHG is mainly focused on the generation of linearly polarized high-order harmonics. The polarization of the harmonics can be controlled using two laser pulses. Setups using either collinear bichromatic beams or noncollinear beams in both cases with counter-rotating polarization have been demonstrated.

Formation of Attosecond Pulses

The temporal structure of high harmonic emission yields the basic for the generation of attosecond pulses, either in the form of a train of pulses or as isolated pulses. The generation of a train does not necessitate any particular requirement for harmonic driver pulse. For spectroscopic applications it can be however desirable to isolate a single pulse from the train by controlling the underlying harmonic emission process.

Attosecond pulse trains

A phase-locked periodic spectrum of equidistant frequencies results in a periodic intensity profile in the time domain, i.e. a comb or train of pulses. With a multi-cycle laser pulse a high harmonic spectrum with pronounced peaks at odd multiples of the fundamental frequency is generated. In the plateau regime of the spectrum the emission peaks are of similar strength and have a smooth relative phase relation. Consequently, the Fourier synthesis of this part of the spectrum leads to a train of pulses with ultrashort duration, which depending on the spectral bandwidth can be shorter than a femtosecond.

Alternatively, the generation of a train of attosecond pulses via HHG can be understood in the time domain directly: In a multi-cycle laser pulse the emission and recombination of the electron wave packet is repeated periodically in every half cycle of the electric field. Hence, a burst of harmonic radiation is emitted over a fraction of the femtosecond driving field cycle, forming each time an attosecond laser pulse. These pulses are separated by the half period of the driving field, and a train of attosecond pulses is generated. Since the harmonic generation process strongly depends on the electric field at the instant of emission of the electron wave packet and during its propagation, in a driving pulse the spectral content and the temporal structure of each attosecond pulse is however different. Furthermore, the electron wave packet returns from alternate side of the atom during subsequent half cycles resulting in attosecond pulses with alternating up- and down-chirp in frequency.

Isolated attosecond pulses

While attosecond pulse trains are used for spectroscopic purposes, a greater flexibility is gained through the generation of isolated attosecond pulses. The extraction or isolation of a single attosecond pulse from a train is too fast for usual electronic devices. However, understanding of the high-order harmonic process gives rise to techniques to select harmonic emission from one recombination event. An overview of the most common techniques is discussed below.

Spectral selection

According to the physical picture of high harmonic generation the maximum harmonic energies close to the cutoff of the spectrum are emitted during the most strongest electric field cycle near the peak of the driving pulse. The generation of an isolated attosecond pulse can be therefore realized by selecting photon energies that are emitted through just one recombination event. This can be achieved by filtering spectral components at lower-energy harmonics, that arise from multiple recombination events produced during other cycles than the central field cycle.

Such a spectral selection scheme has been implemented using a few-cycle laser pulse, in which the electric field strength and hence the extension of the generated harmonic spectrum varies greatly from cycle to cycle. Due to the sensitivity of the high harmonic process to the actual field strength, a control of the carrier-envelope phase is essential as well. The spectral selection method has been also demonstrated in laser pulses, in which the atom is fully ionized during the rising part of the pulse and, hence, the harmonic process is terminated at a certain field cycle.

Temporal gating

Another set of methods for isolated attosecond pulse generation relies on the restriction of the HHG process to just one single half cycle of the electric field. Currently, the most effective technique (polarization gating, **Fig. 4**) makes use of the strong sensitivity of the harmonic process on the ellipticity of the pulse. Since the efficiency of the harmonic emission process decreases quickly by varying the polarization from linear, an isolated attosecond pulse can be produced with a driver laser pulse that is linearly polarized during a single half cycle of the field, while it is elliptically polarized during the remainder of the pulse.

Such a driving pulse with variation of its polarization can be produced by superimposing two counter-rotating circularly polarized laser pulse. By time delaying two such few-cycle laser pulses, the resulting pulse is linearly polarized over the central half

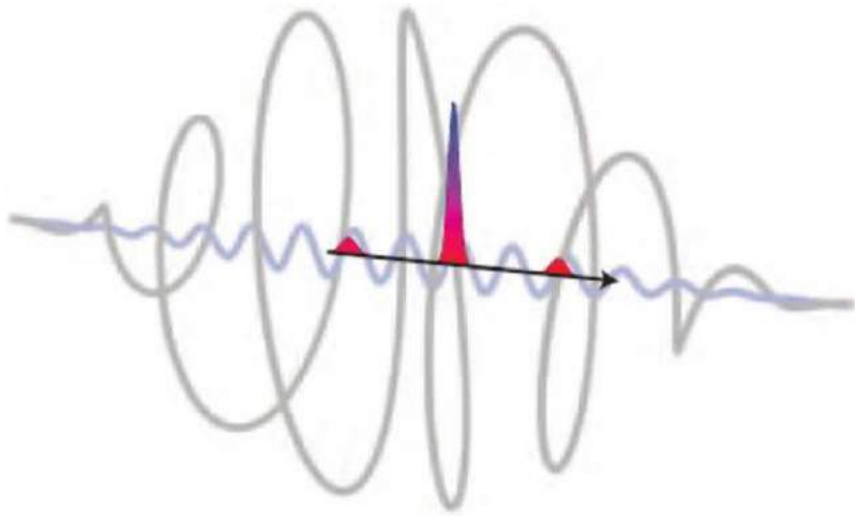

Fig. 4 Principle of polarization gating. A femtosecond driver laser pulse with time-dependent polarization is used. The pulse is linearly polarized during a single half-cycle of the pulse, leading to the restriction of efficient attosecond pulse generation to an isolated event. Reprinted with permission from Mcmillan Publishers Ltd., Chini, M., *et al.*, 2014. Nat. Photonics 8, 178, copyright (2014).

field cycle only, creating the desired temporal gate for efficient harmonic emission and isolated attosecond pulse generation. This concept was further extended by combining a driver pulse with time-dependent ellipticity and a linearly polarized second harmonic pulse. In this technique, known as double optical gating, an asymmetric electric field with time-varying polarization is created. As a consequence, the high harmonic and attosecond pulse generation occurs only once per optical cycle, instead of once per half cycle, providing more flexibility for the application in the polarization gating technique.

Attosecond light house

A further route to the generation of isolated attosecond pulses is provided by a technique, dubbed as the attosecond light house effect (**Fig. 5**). Usually, in a HHG process the attosecond pulses are emitted in a spatially confined beam, having a divergence much lower than the driving laser beam. By introducing a time-dependent rotation of the wavefront, the propagation direction of each harmonic emission event varies depending on the wavefront direction at the moment of the HHG process. While a train of attosecond pulses is generated in this technique, the pulses in the train are angularly separated. If the rotation of the wavefront is larger than the divergence of the attosecond pulses, an isolated attosecond pulse can be spatially selected in the far field. The method has been demonstrated using plasma mirrors as well as gas targets.

Phase matching gating

Each of the previous methods achieves isolation of a single attosecond pulse via control of the HHG process on the microscopic (single-atom) level. Harmonic emission and attosecond pulse production however also depend on the phase matching of the radiation generated from the many atoms in the medium. An essential part of the phase-matching conditions is the degree of ionization of the medium, which increases during the interaction with a strong laser pulse. Consequently, on the leading edge of the pulse the laser phase velocity in the medium may be lower than the speed of light, c, due to the dominance of neutral atoms in the medium. In contrast, due to the higher fraction of ions in the medium in the trailing edge of the pulse, the laser phase velocity may be larger than c. As a result, the phase matching conditions are satisfied only during a temporal window within the laser pulse (**Fig. 6**) and efficient harmonic emission and attosecond pulse generation is restricted to the recombination events falling in this phase matching window. For midinfrared driver lasers the phase matching window is found to be much shorter than for Ti: sapphire lasers. This is, in particular, due to the fact that at longer wavelengths optimal phase conditions are reached at higher gas pressures and ionization densities. As practical consequence, the temporal phase matching window can be narrowed to a single half cycle leading to robust isolated attosecond pulse generation.

Spectroscopy Techniques and Applications

Measurements on the attosecond timescale are still in its infancy. One of the open challenges is the realization of a conventional pump-probe spectroscopy set-up using two isolated attosecond laser pulses, in which time resolution is achieved by the delay and durations of the two pulses. So far, the intensity of attosecond laser pulses has however been too low to achieve significant attosecond pump-probe signals. But, using the knowledge how to manipulate and control electrons with laser pulses, a number of techniques have been developed in which attosecond measurements are performed either without the applications of attosecond

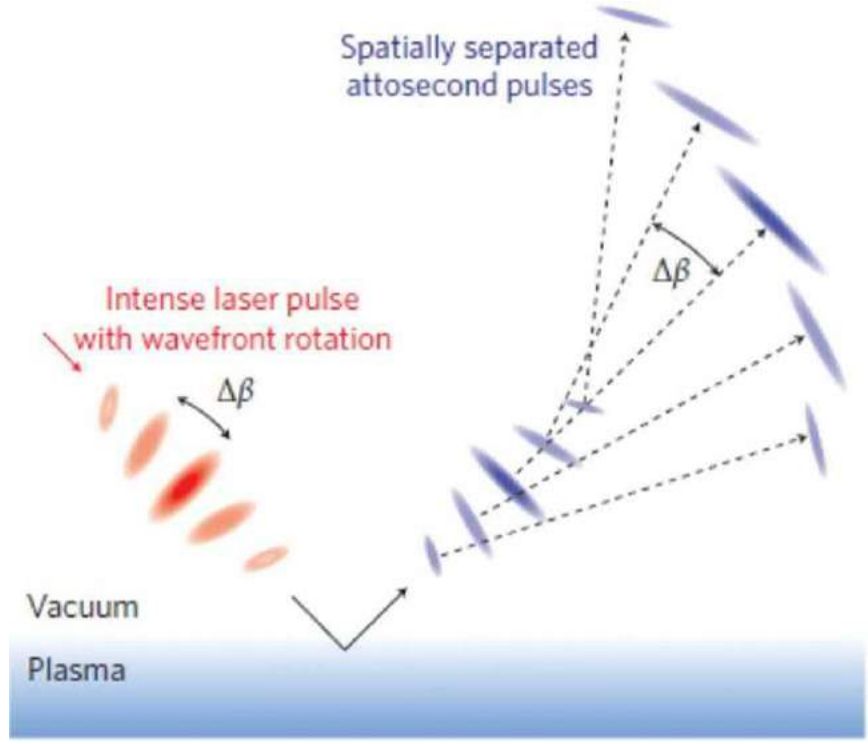

Fig. 5 Attosecond light house principle. Using a time-dependent rotation of the laser wavefront, the attosecond pulses in the train are emitted spatially in different directions and an isolated pulse can be selected. Reproduced with permission from Mcmillan Publishers Ltd., Wheeler, J.A., *et al.*, 2012. Nat. Photonics 6, 829, copyright (2012).

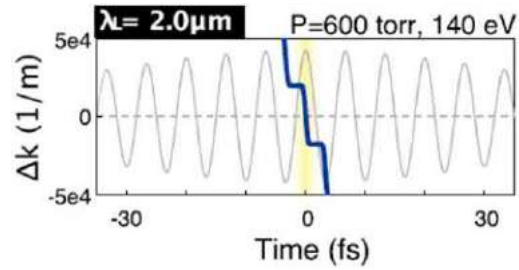

Fig. 6 Calculated phase mismatch Δk for a many-cycle long mid-infrared driver pulse. The phase matching window ($\Delta k \approx 0$, highlighted in yellow) is limited to one half-cycle of the driver pulse, restricting efficient build-up of harmonic emission and attosecond pulse generation to an isolated event. From Chen, M.C., *et al.*, 2014. Proc. Natl. Acad. Sci. USA 111, E2361.

pulses at all or by combining attosecond and femtosecond laser pulses. Applications of attosecond measurements using these spectroscopic techniques are made to reveal electron dynamics in atoms, molecules and solid state systems such as metals and semiconductors.

Attosecond Spectroscopy With Femtosecond Pulses

Recollision spectroscopy

The basic physical process of ionization and return of the electron wave packet in strong laser fields itself (see Section Three-step model of HHG) enables observations with attosecond time resolution. In the spirit of pump-probe spectroscopy, the emission of the electron wave packet may also initiate (pump) a process in the parent ion, e.g., the evolution in excited states or vibration/dissociation in case of a molecular system, while the return of the electron acts as a probe. In this scenario the time between emission and return of the wave packet can be considered as the pump-probe delay. Within a field cycle, the return of the electron wave packet with different energies occur at different time delays (Fig. 7), enabling simultaneous snapshots of the system within one physical process. The variation of the time delays is in the attosecond range for laser wavelengths in the near-infrared and mid-infrared regime. Further control of the delay, within certain limits, can be achieved via tuning of the wavelength and hence the duration of the field cycle of the driving laser field.

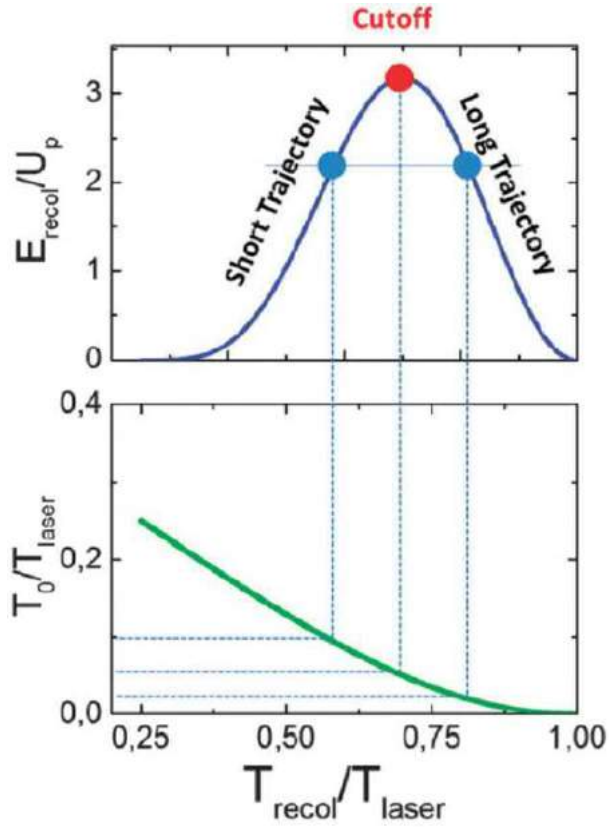

Fig. 7 Relation between the electron return energy and the return time (upper panel) and the times of ionization and return of the electron (lower panel), according to the classical rescattering model (Eq. (8)). Reproduced from Vrakking, M.J.J., 2014. Attosecond imaging. Phys. Chem. Chem. Phys. 16, 2775 with permission of the PCCP Owner Societies.

On its return the electron wave packet can recombine with, diffract on or further ionize the parent ion. The resulting harmonic or photoelectron spectra contain information about changes in the structure of the ion as well as the electronic distribution inside the ion with an attosecond time resolution. For the analysis and interpretation of the spectra it is often assumed that the amplitude for the full process can be factorized in one for the propagation of the electron wave packet and one for its (field-free) scattering or recombination.

Attoclock: Angular streaking

Another way to realize attosecond time resolution in the interaction of a femtosecond laser pulse with matter relies on the application of an ultrashort two-cycle nearly circularly polarized laser pulse. Here, the resolution is achieved because of the rotation of the electric field vector in the polarization plane, which serves as the hand of a clock. If a target atom or molecule gets ionized the final emission angle of the electron wave packet is related to the instant of ionization. Thus, the observed photoelectron angular momentum distribution contains time information about the ionization process. Peak positions in such distributions can be measured with an angular resolution of about 1°, which corresponds to a temporal resolution of about 7.4 as at a center wavelength of 800 nm. Further calibration of this attoclock concept also relies on the accurate analysis of the propagation of the electron wave packet in the combined fields of the circularly polarized field and the Coulomb field between the electron and the parent ion.

Such angular-resolved photoelectron measurements with ultrashort nearly circularly polarized femtosecond laser pulses have been applied to assess the time for tunneling of an electron wave packet through the strong field suppressed Coulomb barrier, the time delay in the successive steps of electron emission in sequential strong-field double ionization and the emission time of an electron wave packet from a molecule. Besides the challenges in the accurate analysis of the results, it remains technically difficult to generate circularly polarized pulses in the few-cycle regime. A residual amount of ellipticity in the polarization results in small oscillations of the electric field magnitude on the sub-cycle timescale and may distort the resulting momentum distributions.

Photoelectron and Ion Spectroscopy With Attosecond Pulses

Although pump-probe spectroscopy with attosecond pulses in both steps could not be realized up to date, in experiments employing different two-color schemes with a femtosecond (near-infrared) pulse and isolated attosecond (XUV) pulse or attosecond pulse trains high time resolution in measurements were achieved. As a common principle the attosecond time resolution is

achieved by controlling the application of the attosecond pulse over the optical field cycle of the near-infrared pulse. Since the near-infrared pulse is usually a replica of the pulse used to generate the attosecond pulses, perfect synchronization and control of the time delay between the pulses can be assured. In general, the use even of a moderately strong femtosecond pulse however complicates the analysis and interpretation of the results, as the impact of the near-infrared field on the observed dynamics needs to be taken into account.

Attosecond streak camera

One set of attosecond measurements makes use of the streak camera principle by applying a linearly polarized isolated attosecond pulse along with a moderately strong femtosecond near-infrared laser pulse in interaction with an atom. A photoelectron, which is instantaneously liberated into the continuum by the attosecond pulse, experiences a momentum shift due to the presence of the near-infrared laser pulse (Fig. 8). The magnitude of the shift depends on the vector potential at the time of ionization and the subsequent propagation of the electron wave packet in the combined fields of the Coulombic interaction between the electron and the ion as well as with the near-infrared streaking pulse. By recording the momentum (or energy) shift as a function of the time delay between the two pulses a streaking trace is obtained.

The attosecond streak camera concept was initially used to characterize isolated attosecond pulses, but soon after it was applied to obtain dynamic information about processes in atoms and solids, in which usually the streaking traces for two events at different photoelectron energies are compared with each other. Proof-of-principle experiment of this kind was a measurement of the lifetime of an inner-shell vacancy in an atom, which was produced by excitement of a core electron by the attosecond XUV pulse. Sampling of the emission of the Auger electron, following the filling of the hole from higher energy level, by the femtosecond near-infrared field provides access to time resolution of this inner-shell process. Later, observation of a relative shift between streaking traces for photoelectrons emitted from different levels in atoms and solids was interpreted as due to the difference in the time delays acquired during the emission and propagation of the respective electron wave packets.

Reconstruction of attosecond beating by interference of two-photon transitions

Using a train of attosecond pulses, instead of an isolated attosecond pulse, in conjunction with a long weak near-infrared pulse leads to a technique referred to as reconstruction of attosecond beating by interference of two-photon transitions (RABBITT). Assuming that the spectral bandwidth of the attosecond pulses in the train spans a number of harmonics, ionization of the target in the combined fields leads to a series of main bands (corresponding to the photoelectron energies upon absorption of one of harmonic photons) and side bands (corresponding to the additional absorption or emission of a near-infrared photon) in the photoelectron energy spectrum (Fig. 9). Since there are two pathways leading to the same side band, there appears an interference pattern in the spectrum as a function of the time the attosecond pulse is applied over the optical half cycle of the near-infrared field.

The time resolution in the RABBITT method is related to the modulation of the probability of the side band peaks, that depends on the phase difference between consecutive harmonics (or the time of application of the attosecond pulses) and an

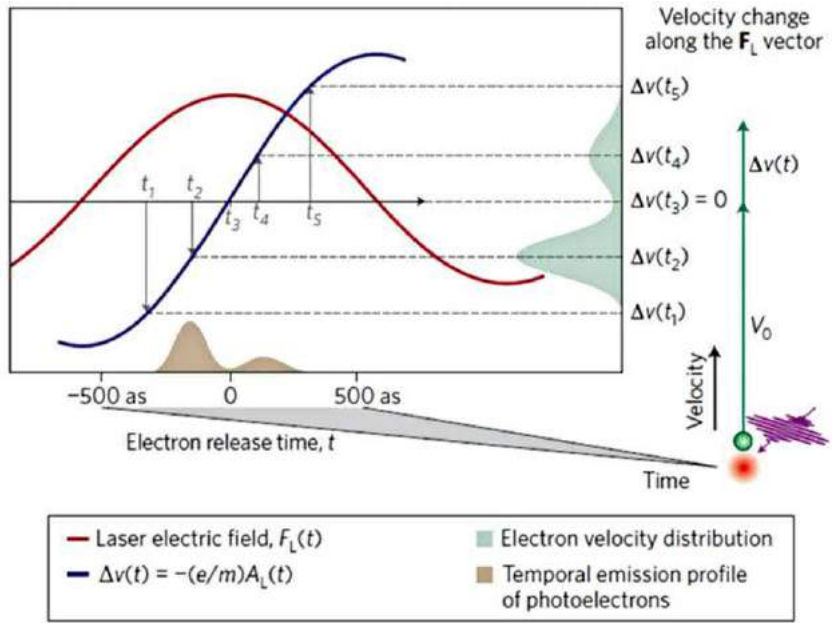

Fig. 8 Concept of attosecond streak camera: Photoelectrons emitted by an attosecond pulse are probed by the presence of a femtosecond (streaking) pulse. The final photoelectron momentum is related to the vector potential of the streaking pulse at the instant of release of the electron wave packet. Reprinted with permission from Mcmillan Publishers Ltd., Krausz, F., Stockmann, M.I., 2014. Nat. Photonics 8, 205, copyright (2014).

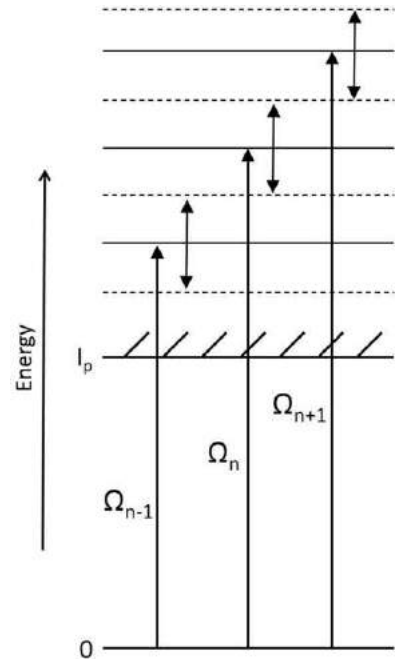

Fig. 9 Illustration of energy levels with main bands (solid lines) and side bands (dashed lines) in the continuum of the spectrum, as well as transitions involved in RABBITT principle.

intrinsic phase related to the target. In combination with measurement of the amplitudes of the harmonics the phase information can be used to reconstruct the temporal profile of the attosecond pulses and the electron wave packets initiated by these pulses. In the ideal (limiting) case of a CW monochromatic field and a periodic repetition of identical attosecond pulses in the train, the RABBITT scheme provides equivalent information as the attosecond streak camera technique, discussed in Section Spectral selection. Since it is usually less demanding experimentally to generate attosecond pulse trains and the probe infrared field can be weaker than in the case of streaking, the RABBITT is often considered more practical. Accordingly, applications of the RABBITT technique include delay measurements in atomic and condensed matter systems.

Similar kind of photoelectron spectroscopy has been performed using stronger femtosecond laser pulses as well, which induce higher-order near-infrared photon transitions in the continuum (or, equivalently larger photoelectron momentum shifts). Consequently, interferences between the main bands in the angular resolved photoelectron energy spectrum appear and information about photoelectron emission and recollision can be retrieved.

Ion spectroscopy

As an alternative to observation of the photoelectron, the population of ionic states and migration and localization of charges in the course of molecular dissociation or fragmentation can be detected to observe attosecond electron dynamics. In this set-up the attosecond pulse is used to ionize and/or excite the molecule and induces an electron wave packet dynamics between different electronic states on the attosecond timescale. The electric field of the femtosecond laser pulse, applied at a variable time delay, probes this dynamics either via ionization of the electron or by driving it inside the dissociating molecular system. The population in the ion channels as well as the localization of the electron charges on the fragments depends on the time delay between the two pulses and contains time information about the attosecond electron dynamics inside the molecule. Analysis of the attosecond dynamics can be complicated by correlation between the photoelectron and the residual molecular ion as well as nonadiabatic response of the electron wave packet to the driving oscillating electric field.

Absorption Spectroscopy With Attosecond Pulses

In another class of attosecond measurements time-resolved spectral information is obtained by applying a femtosecond near-infrared laser pulse either before or after the attosecond pulse. In these pump-probe techniques, which are complementary to those discussed in Section Photoelectron and Ion Spectroscopy With Attosecond Pulses, the time information is contained in the radiation transmitted through the gas medium, liquid or solid.

Attosecond transient absorption spectroscopy

In transient absorption spectroscopy the absorption by a medium following the interaction with a pump and a probe pulses, that are time-delayed with respect to each other, is measured. In the application to attosecond measurements a weak attosecond XUV

pulse is used together with an intense femtosecond near-infrared pulse. The technique can be applied for either sequence of the pulses, where the near-infrared pulse either precedes (conventional pulse sequence) or follows (unconventional pulse sequence) the XUV pulse. Although the time-integrated absorption spectrum is observed, fast transient dynamics may be imprinted onto the spectrum. The response of the medium depends on the interaction at the single-atom level, which is proportional to the imaginary part of the product of Fourier transforms of the one-electron dipole moment (Eq. (5)) and the two-color electric field, and the collective response from all atoms via the macroscopic polarization. Temporal resolution in such a transient absorption experiment is achieved by recording the spectrum as a function of the delay between the two pulses.

In the conventional sequence the attosecond pulse probes the dynamics initiated by the near-infrared pump pulse. With this set-up electron wave packet dynamics in the valence shells of atomic systems coherently excited by an intense femtosecond pulse has been measured. In another application the time dependence of excitations in the conduction band of silicon dioxide have been observed. Using the opposite pulse sequence, in which the attosecond pulse comes first as a pump, spectral features corresponding to light-induced states in the continuum, in particular the lifetimes and lineshapes related to autoionizing states, have been analyzed.

Attosecond four-wave mixing spectroscopy

Very recently, earlier proposals to extend the concepts of 2D Fourier transform and four-wave mixing spectroscopy by using attosecond pulses have been implemented in the experiment. To this end, the spectrum of the attosecond pulses were comprised to just one harmonic, which is resonant with a transition to an excited state in the atomic target. Further coupling of states in the atom by a near-infrared pulse, leads to the population of a state via the absorption of a combination of one XUV and two near-infrared photons. The decay of the state back to the ground state induces the emission of a (fourth) photon at a specific energy. Detection of the emitted radiation as a function of the time delay between the attosecond XUV pulse train and the femtosecond near-infrared pulse allows to retrieve time-resolved information about the dynamics in the coupled states.

References

Calegari, F., Sansone, G., Stagira, S., Vozzi, C., Nisoli, M., 2006. Advances in attosecond science. J. Phys. B: At., Mol. Opt. Phys. 49, 062001.

Chang, Z., 2011. Fundamentals of Attosecond Optics. CRC Press, Taylor & Francis Group.

Corkum, P.B., Krausz, F., 2007. Attosecond science. Nat. Phys. 3, 381–387.

Dahlström, J.M., L'Huillier, A., Maquet, A., 2012. Introduction to attosecond delays in photoionization. J. Phys. B: At. Mol. Opt. Phys. 45, 183001.

Gallmann, L., Cirelli, C., Keller, U., 2012. Attosecond science: recent highlights and future trends. Annu. Rev. Phys. Chem. 63, 447–469.

Kapteyn, H.C., Murnane, M.M., Christov, I.P., 2005. Extreme nonlinear optics: coherent X rays from lasers. Phys. Today 59 (3), 39–44.

Krausz, F., Ivanov, M., 2009. Attosecond physics. Rev. Mod. Phys. 81, 163–234.

Leone, S.R., McCurdy, C.W., Burgdörfer, J., *et al.*, 2014. What will it take to observe processes in 'real time'? Nat. Photonics 8, 162–166.

Leone, S.R., Neumark, D.M., 2016. Attosecond science in atomic, molecular, and condensed matter physics. Faraday Discuss. 194, 15–39.

Pazourek, R., Nagele, S., Burgdörfer, J., 2015. Attosecond chronoscopy of photoemission. Rev. Mod. Phys. 87, 765–802.

Peng, L.-Y., Jiang, W.-C., Geng, J.-W., Xiong, W.-H., Gong, Q., 2015. Tracing and controlling electronic dynamics in atoms and molecules by attosecond pulses. Phys. Rep. 575, 1–71.

Popmintchev, T., Chen, M.-C., Arpin, P., Murnane, M.M., Kapteyn, H.C., 2011. The attosecond nonlinear optics of bright coherent X-ray generation. Nat. Photonics 4, 822–832.

Ramasesha, K., Leone, S.R., Neumark, D.M., 2016. Real-time probing of electron dynamics using attosecond time-resolved spectroscopy. Annu. Phys. Rev. Chem. 67, 41–63.

Vrakking, M.J.J., 2014. Attosecond imaging. Phys. Chem. Chem. Phys. 16, 2775–2789.

Wu, M., Chen, S., Camp, S., Schafer, K.J., Gaarde, M.B., 2016. Theory of strong-field attosecond transient absorption spectroscopy. J. Phys. B: At. Mol. Opt. Phys. 49, 062003.

Nonlinear Spectroscopies

SR Meech, University of East Anglia, Norwich, UK

Introduction

Chemistry is concerned with the induction and observation of changes in matter, where the changes are to be understood at the molecular level. Spectroscopy is the principal experimental tool for connecting the macroscopic world of matter with the microscopic world of the molecule, and is, therefore, of central importance in chemistry. Since its invention, the laser has greatly expanded the capabilities of the spectroscopist. In linear spectroscopy, the monochromaticity, coherence, high intensity, and high degree of polarization of laser radiation are ideally suited to high-resolution spectroscopic investigations of even the weakest transitions. The same properties allowed, for the first time, the investigation of the nonlinear optical response of a medium to intense radiation. Shortly after the foundations of nonlinear optics were laid, it became apparent that these nonlinear optical signals could be exploited in molecular spectroscopy, and since then a considerable number of nonlinear optical spectroscopies have been developed. This short article is not a comprehensive review of all these methods. Rather, it is a discussion of some of the key areas in the development of the subject, and indicates some current directions in this increasingly diverse area.

Nonlinear Optics for Spectroscopy

The foundations of nonlinear optics are described in detail elsewhere in the encyclopedia, and in some of the texts listed in the Further Reading section at the end of this article. The starting point is usually the nonlinearity in the polarization, P_i, induced in the sample when the applied electric field, E, is large:

$$P_i = \varepsilon_0 \left[\chi^{(1)}_{ij} E_j + \frac{1}{2}\chi^{(2)}_{ijk} E_j E_k + \frac{1}{4}\chi^{(3)}_{ijkl} E_j E_k E_l + \ldots \right] \tag{1}$$

where ε_0 is the vacuum permittivity, $\chi^{(n)}$ is the nth order nonlinear susceptibility, the indices represent directions in space, and the implied summation over repeated indices convention is used. The signal field, resulting from the nonlinear polarization, is calculated by substituting it as the source polarization in Maxwell's equations and converting the resulting field to the observable, which is the optical intensity.

The nonlinear polarization itself arises from the anharmonic motion of electrons under the influence of the oscillating electric field of the radiation. Thus, there is a microscopic analog of **Eq. (1)** for the induced molecular dipole moment, μ_i:

$$\mu_i = \left[\varepsilon_0 \alpha_{ij} d_j + \varepsilon_0^{-2} \beta_{ijk} d_j d_k + \varepsilon_0^{-3} \gamma_{ijkl} d_j d_k d_l + \ldots \right] \tag{2}$$

in which α is the polarizability, β the first molecular hyperpolarizability, γ the second, etc. The power series is expressed in terms of the displacement field **d** rather than **E** to account for local field effects. This somewhat complicates the relationship between the molecular hyperpolarizabilities, for example, γ_{ijkl} and the corresponding macroscopic susceptibility, $\chi^{(3)}_{ijkl}$, but it is nevertheless generally true that a molecule exhibiting a high value of γ_{ijkl} will also yield a large third-order nonlinear polarization. This relationship between the macroscopic and microscopic parameters is the basis for an area of chemistry which has influenced nonlinear optics (rather than the inverse). The synthesis of molecules which have giant molecular hyperpolarizabilities has been an active area, because of their potential applications in lasers and electro-optics technology. Such molecules commonly exhibit a number of properties, including an extended linear π electron conjugation and a large dipole moment, which changes between the ground and excited electronic states. These properties are in accord with the predictions of theory and of quantum chemical calculations of molecular hyperpolarizabilities.

Nonlinear optical spectroscopy in the frequency domain is carried out by measuring the nonlinear signal intensity as a function of the frequency (or frequencies) of the incident radiation. Spectroscopic information is accessible because the molecular hyperpolarizability, and therefore the nonlinear susceptibility, exhibits resonances: the signal is enhanced when one or more of the incident or generated frequencies are resonant with the frequency of a molecular transition. The rich array of nonlinear optical spectroscopies arises in part from the fact that with more input fields there are more accessible resonances than is the case with linear spectroscopy. As an example we consider the third-order susceptibility, $\chi^{(3)}_{ijkl} E_j E_k E_l$, in the practically important case of two incident fields at the same frequency, ω_1, and a third at ω_2. The difference between the two frequencies is close to the frequency of a Raman active vibrational mode, Ω_{rg} (**Fig. 1**). The resulting susceptibility can be calculated to have the form:

$$\chi^{(3)}_{ijkl}(-2\omega_1 + \omega_2; \omega_1, \omega_1, -\omega_2) = \frac{N\Delta\rho_{rg}\Omega_{rg}\left(\alpha^{\mathrm{R}}_{ij}\alpha^{\mathrm{R}}_{kl} + \alpha^{\mathrm{R}}_{ik}\alpha^{\mathrm{R}}_{jl}\right)}{12\hbar[\Omega^2_{rg} - (\omega_1 - \omega_2)^2 + \Gamma^2 - 2\mathrm{i}(\omega_1 - \omega_2)\Gamma]} \tag{3}$$

 doi:10.1016/B978-0-12-809283-5.00956-3

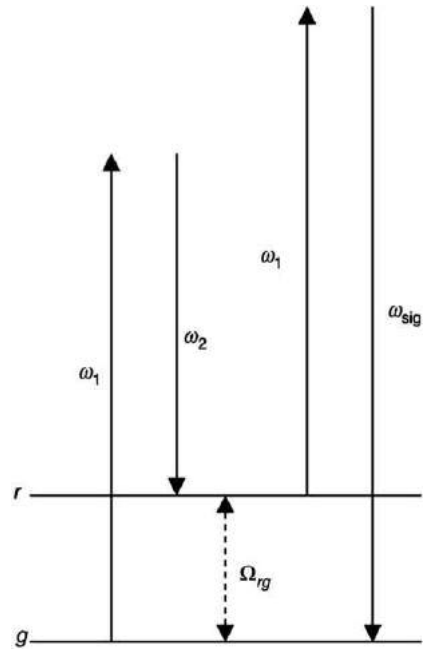

Fig. 1 An illustration of the resonance enhancement of the CARS signal, $\omega_{sig}=2\omega_1-\omega_2$ when $\omega_1-\omega_2=\Omega_{rg}$.

in which the α^R are elements of the Raman susceptibility tensor, Γ is the homogeneous linewidth of the Raman transition, $\Delta\rho_{rg}$ a population difference, and N the number density. A diagram illustrating this process is shown in Fig. 1, where the signal field is at the anti-Stokes Raman frequency $(2\omega_1-\omega_2)$. The spectroscopic method which employs this scheme is called coherent anti-Stokes Raman spectroscopy (CARS) and is one of the most widely applied nonlinear optical spectroscopies (see below). Clearly, from Eq. (3) we can see that the signal will be enhanced when the difference frequency is resonant with the Raman transition frequency.

With essentially the same experimental geometry there will also be nonlinear signals generated at both the Stokes frequency $(2\omega_2-\omega_1)$ and at the lower of the incident frequencies, ω_2. These signals have distinct phase matching conditions (see below), so they can easily be discriminated from one another, both spatially and energetically. Additional resonance enhancements are possible if either of the individual frequencies is resonant with an electronic transition of the molecule, in which case information on Raman active modes in the excited state is also accessible.

It is worthwhile noting here that there is an equivalent representation of the nth order nonlinear susceptibility tensor $\boldsymbol{\chi}^{(n)}(\omega_{sig}:\omega_1,\ldots,\omega_n)$ as a time domain response function, $\mathbf{R}^{(n)}(\tau_1,\ldots,\tau_n)$. While it is possible to freely transform between them, the frequency domain representation is the more commonly used. However, the response function approach is increasingly applied in time domain nonlinear optical spectroscopy when optical pulses shorter than the homogeneous relaxation time are used. In that case, the time ordering of the incident fields, as well as their frequencies, is of importance in defining an experiment.

An important property of many nonlinear optical spectroscopies is the directional nature of the signal, illustrated in Fig. 2. The directional nonlinear signal arises from the coherent oscillation of induced dipoles in the sample. Constructive interference can lead to a large enhancement of the signal strength. For this to occur the individual induced dipole moments must oscillate in phase – the signal must be phase matched. This requires the incident and the generated frequencies to travel in the sample with the same phase velocity, $k_\omega/\omega=c/n_\omega$ where n_ω is the index of refraction and k_ω is the wave propagation constant at frequency ω. For the simplest case of second-harmonic generation (SHG), in which the incident field oscillates at ω and the generated field at 2ω, phase matching requires $2k_\omega=k_{2\omega}$. The phase-matching condition for the most efficient generation of the second harmonic is when the phase mismatch, $\Delta k=2k_\omega-k_{2\omega}=0$, but this is not generally fulfilled due to the dispersion of the medium, $n_\omega<n_{2\omega}$. In the case of SHG, a coherence length L can be defined as the distance traveled before the two waves are 180° out of phase, $L=|\pi/\Delta k|$. For cases in which the input frequencies also have different directions, as is often the case when laser beams at two frequencies are combined (e.g., CARS), the phase-matching condition must be expressed as a vectorial relationship, hence, for CARS, $\Delta\mathbf{k}=\mathbf{k}_{AS}-2\mathbf{k}_1+\mathbf{k}_2\approx 0$, where $\mathbf{k}_{AS}$ is the wavevector of the signal at the anti-Stokes frequency (Fig. 2). Thus for known input wavevectors one can easily calculate the expected direction of the output signal, $\mathbf{k}_s$. This is illustrated for a number of important cases in Fig. 2.

Nonlinear Optical Spectroscopy in Chemistry

As already noted, there is a vast array of nonlinear optical spectroscopies, so it is clear some means of classification will be required. For coarse graining the order of the nonlinear process is very helpful, and that is the scheme we will follow here.

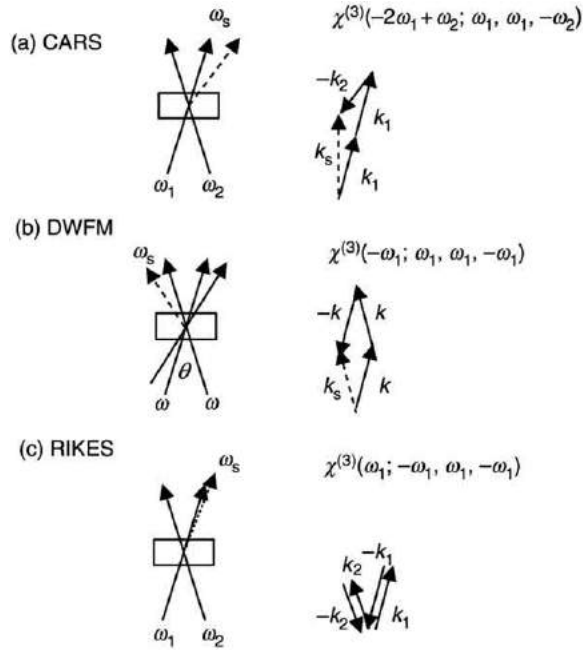

Fig. 2 Experimental geometries and corresponding phase matching diagrams for (a) CARS (b) DFWM (c) RIKES. Note that only the RIKES geometry is fully phase matched.

Second-Order Spectroscopies

Inspection of **Eq. (1)** shows that in gases, isotropic liquids, and solids where the symmetry point group contains a center of inversion, $\chi^{(2)}$ is necessarily zero. This is required to satisfy the symmetry requirement that polarization must change sign when the direction of the field is inverted, yet for a quadratic, or any even order dependence on the field strength, it must remain positive. Thus second-order nonlinearities might not appear very promising for spectroscopy. However, there are two cases in which second-order nonlinear optical phenomena are of very great significance in molecular spectroscopy, harmonic generation and the spectroscopy of interfaces.

Harmonic conversion

Almost every laser spectroscopist will have made use of second-harmonic or sum frequency generation for frequency conversion of laser radiation. Insertion of an oscillating field of frequency ω into **Eq. (1)** yields a second-order polarization oscillating of 2ω. If two different frequencies are input, the second-order polarization oscillates at their sum and difference frequencies. In either case, the second-order polarization acts as a source for the second-harmonic (or sum, or difference frequency) emission, provided $\chi^{(2)}$ is nonzero. The latter can be arranged by selecting a noncentrosymmetric medium for the interaction, the growth of such media being an important area of materials science. Optically robust and transparent materials with large values of $\chi^{(2)}$ are available for the generation of wavelengths shorter than 200 nm to longer than 5 μm. Since such media are birefringent by design, a judicious choice of angle and orientation of the crystal with respect to the input beams allows a degree of control over the refractive indices experienced by each beam. Under the correct phase matching conditions $n_\omega \approx n_{2\omega}$, and very long interaction lengths result, so the efficiency of signal generation is high.

Higher-order harmonic generation in gases is an area of growing importance for spectroscopy in the deep UV and X-ray region. The generation of such short wavelengths depends on the ability of amplified ultrafast solid state lasers to generate extremely high instantaneous intensities. The mechanism is somewhat different to the one outlined above. The intense pulse is injected into a capillary containing an appropriate gas. The high electric field of the laser causes ionization of atoms. The electrons generated begin to oscillate in the applied laser field. The driven recombination of the electrons with the atoms results in the generation of the high harmonic emission. Although the mechanism differs from the SHG case, questions of phase matching are still important. By containing the gas in a corrugated waveguide phase matching is achieved, considerably enhancing the intensity of the high harmonic. Photon energies of hundreds of electronvolts are possible using this technique. The generation of such high energies is not yet routine, but a number of potential applications are already apparent. A powerful coherent source of X-ray and vacuum UV pulses will certainly aid surface analysis techniques such as UV photo-emission and X-ray photoelectron spectroscopy. Much excitement is currently being generated by the possibility of using intense ultrashort X-ray pulses to record structural dynamics on an ultrafast timescale.

Surface second-order spectroscopy

At an interface inversion symmetry is absent by definition, so the second-order nonlinear susceptibility is finite. If the two bulk phase media are themselves isotropic, then even a weak second-order signal necessarily arises from the interface. This surface-specific

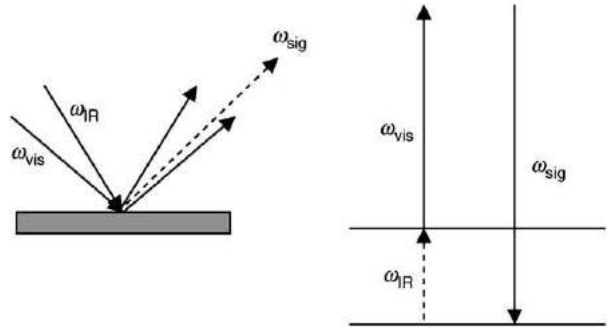

Fig. 3 The experimental geometry for SFG, and an illustration of the resonance enhancement of $\omega_{sig} = \omega_{IR} + \omega_{vis}$ at a Raman and IR allowed vibrational transition.

all-optical signal is unique, because it can be used to probe the interface between two condensed phases. This represents a great advantage over every other form of surface spectroscopy. In linear optical spectroscopy, the signal due to species at the interface are usually swamped by contributions from the bulk phase. Other surface-specific signals do exist, but they rely on the scattering of heavy particles (electrons, atoms) and so can only be applied to the solid vacuum interface. For this reason the techniques of surface SHG and SFG are widely applied in interface spectroscopy.

The most widely used method is sum frequency generation (SFG) between temporally overlapped tuneable infrared and fixed frequency visible lasers, to yield a sum frequency signal in the visible region of the spectrum. The principle of the method is shown in **Fig. 3**. The surface nonlinear susceptibility exhibits resonances at vibrational frequencies, which are detected as enhancements in the visible SFG intensity. Although the signal is weak, it is directional and background free, so relatively easily measured by photon counting techniques. Thus, SFG is used to measure the vibrational spectra of essentially any optically accessible interface. There are, however, some limitations on the method. The surface must exhibit a degree of order – if the distribution of adsorbate orientation is isotropic the signal again becomes zero by symmetry. Second, a significant enhancement at the vibrational frequency requires the transition to be both IR and Raman allowed, as suggested by the energy level diagram (**Fig. 3**).

The SHG signal can also be measured as a function of the frequency of the incident laser, to recover the electronic spectrum of the interface. This method has been used, particularly in the study of semiconductor surfaces, but generally the electronic spectra of adsorbates contain less information than their vibrational spectra. However, by measuring the SHG intensity as a function of time, information on adsorbate kinetics is obtained, provided some assumptions connecting the surface susceptibility to the molecular hyperpolarizability are made. Finally, using similar assumptions, it is possible to extract the orientational distribution of the adsorbate, by measuring the SHG intensity as a function of polarization of the input and output beams. For these reasons, SHG has been widely applied to analyze the structure and dynamics of interfaces.

Third-Order Spectroscopies

The third-order coherent Raman spectroscopies were introduced above. One great advantage of these methods over conventional Raman is that the signal is generated in a coherent beam, according to the appropriate phase matching relationship (**Fig. 2**). Thus, the coherent Raman signal can easily be distinguished from background radiation by spatial filtering. This has led to CARS finding widespread application in measuring spectra in (experimentally) hostile environments. CARS spectroscopy has been widely applied to record the vibrational spectra of flames. Such measurements would obviously be very challenging for linear Raman or IR, due to the strong emission from the flame itself. The directional CARS signal in contrast can be spatially filtered, minimizing this problem. CARS has been used both to identify unstable species formed in flames, and to probe the temperature of the flame (e.g., from measured population differences **Eq. (3)**). A second advantage of the technique is that the signal is only generated when the two input beams overlap in space. Thus, small volumes of the sample can be probed. By moving the overlap position around in the sample, spatially resolved information is recovered. Thus, it is possible to map the population of a particular transient species in a flame.

CARS is probably the most widely used of the coherent Raman methods, but it does have some disadvantages, particularly in solution phase studies. In that case the resonant $\chi^{(3)}$ signal (**Eq. (3)**) is accompanied by a nonresonant third-order background. The interference between these two components may result in unusual and difficult to interpret lineshapes. In this case, some other coherent Raman methods are more useful. The phase matching scheme for Raman-induced Kerr effect spectroscopy (RIKES) was shown in **Fig. 2**. The RIKES signal is always phase matched, which leads to a long interaction length. However, the signal is at ω_2 and in the direction of ω_2, which would appear to be a severe disadvantage in terms of detection. Fortunately, if the input polarizations are correctly chosen the signal can be isolated by its polarization. In an important geometry, the signal (ω_2) is isolated by transmission through a polarizer oriented at 45° to a linearly polarized pump (ω_1). The pump is overlapped in the sample with the probe (ω_2) linearly polarized at $-45°$. Thus the probe is blocked by the polarizer, but the signal is transmitted. This geometry may be viewed as pump-induced polarization of the isotropic medium to render it birefringent, thus inducing ellipticity in the transmitted probe, such that the signal leaks through the analyzing polarizer (hence the alternative name optical Kerr effect).

In this geometry, it is possible to greatly enhance signal-to-noise ratios by exploiting interference between the probe beam and the signal. Placing a quarterwave plate in the probe beam with its fast axis aligned with the probe polarization, and reorienting it slightly ($<1^\circ$) yields a slightly elliptically polarized probe before the sample. A fraction of the probe beam, the local oscillator (LO), then also leaks through the analyzing polarizer, temporally and spatially overlapped with the signal. Thus, the signal and LO fields are seen by the detector, which measures the intensity as:

$$I(t) = \frac{nc}{8\pi}|\mathbf{E}_{\mathrm{LO}}(t) + \mathbf{E}_{\mathrm{S}}(t)|^2 = I_{\mathrm{LO}}(t) + I_{\mathrm{S}}(t) + \frac{nc}{4\pi}\mathrm{Re}\left[\mathbf{E}_{\mathrm{S}}^{*}(t)\cdot\mathbf{E}_{\mathrm{LO}}(t)\right] \tag{4}$$

where the final term may be very much larger than the original signal, and is linear in $\chi^{(3)}$. This term is usually isolated from the strong I_{LO} by lock-in detection. This method is called optical heterodyne detection (OHD), and generally leads to excellent signal to noise. It can be employed with other coherent signals by artificially adding the LO to the signal, provided great care is taken to ensure a fixed phase relationship between LO and signal. In the RIKES experiment, however, the phase relationship is automatic. The arrangement described yields an out-of-phase LO, and measures the real part of $\chi^{(3)}$, the birefringence. Alternatively, the quarterwave plate is omitted, and the analyzing polarizer is slightly reoriented, to introduce an in-phase LO, which measures the imaginary part of $\chi^{(3)}$, the dichroism. This is particularly useful for absorbing media. The OHD-RIKES method has been applied to measure the spectroscopy of the condensed phase, and has found particularly widespread application in transient studies (below).

Degenerate four wave mixing (DFWM) spectroscopy is a simple and widely used third-order spectroscopic method. As the name implies, only a single frequency is required. The laser beam is split into three, and recombined in the sample, in the geometry shown in **Fig. 2**. The technique is also known as laser-induced grating scattering. The first two beams can be thought of as interfering in the sample to write a spatial grating, with fringe spacing dependent on the angle between them. The signal is then scattered from the third beam in the direction expected for diffraction from that grating. The DFWM experiment has been used to measure electronic spectra in hostile environments, by exploiting resonances with electronic transitions. It has also been popular in the determination of the numerical value of $\chi^{(3)}$, partly because it is an economical technique, requiring only a single laser, but also because different polarization combinations make it possible to access different elements of $\chi^{(3)}$. The technique has also been used in time resolved experiments, where the decay of the grating is monitored by diffraction intensity from a time delayed third pulse.

Two-photon or, more generally, multiphoton excitation has applications in both fundamental spectroscopy and analytical chemistry. Two relevant level schemes are shown in **Fig. 4**. Some property associated with the final state permits detection of the multiphoton absorption, for example, fluorescence in (a) and photocurrent in (b).

Excitation of two-photon transitions, as in **Fig. 4(a)**, is useful in spectroscopy because the selection rules are different to those for the corresponding one-photon transition. For example the change in angular momentum quantum number, ΔL, in a two-photon transition is 0, ± 2, so, for example, an atomic S to D transition can be observed. High spectroscopic resolution may be attained using Doppler free two-photon absorption spectroscopy. In this method, the excitation beams are arranged to be counter-propagating, so that the Doppler broadening is cancelled out in transitions where the two excitation photons arise from beams with opposing wavevectors. In this case, the spectroscopic linewidth is governed only by the homogeneous dephasing time.

The level scheme in **Fig. 4(b)** is also widely used in spectroscopy, but in this case the spectrum of the intermediate state is obtained by monitoring the photocurrent as a function of ω_1. The general technique is known as resonance enhanced multiphoton ionization (REMPI) and yields high-quality spectra of intermediate states which are not detectable by standard methods, such as fluorescence. The sensitivity of the method is high, and it is the basis of a number of analytical applications, often in combination with mass spectrometry.

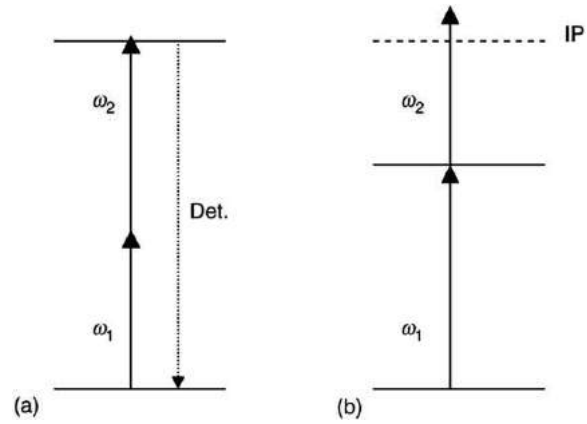

Fig. 4 Two cases of resonant two photon absorption. In (a) the excited state is two-photon resonant, and the process is detected by the emission of a photon. In (b) the intermediate state is resonant, and the final energy is above the ionization potential (IP) so that photocurrent or mass detection can be used.

Ultrafast Time Resolved Spectroscopy

The frequency and linewidth of a Raman transition may be extracted from the CARS measurement, typically by combining two narrow bandwidth pulsed lasers, and tuning one through the resonance while measuring the nonlinear signal intensity. The time resolved analogue requires two pulses, typically of a few picoseconds duration (and therefore a few wavenumbers bandwidth) at ω_1 and ω_2 to be incident on the sample. This pair coherently excites the Raman mode. A third pulse at ω_1 is incident on the sample a time t later, and stimulates the CARS signal at $2\omega_1 - \omega_2$ in the phase-matched direction. The decay rate of the vibrational coherence is measured from the CARS intensity as a function of the delay time. Thus, the frequency of the mode is measured in the frequency domain, but the linewidth is measured in the time domain. If very short pulses are used (such that the pulsewidth is shorter than the inverse frequency of the Raman active mode) the Raman transition is said to be impulsively excited, and the CARS signal scattered by the time delayed pulse reveals an oscillatory response at the frequency of the Raman active mode, superimposed on its decay. Thus, in this case, spectroscopic information is measured exclusively in the time domain. In the case of nonlinear Raman spectroscopy, similar information is available from the frequency and the time domain measurements, and the choice between them is essentially one of experimental convenience. For example, time domain CARS, RIKES, and DFWM spectroscopy have turned out to be particularly powerful routes to extracting low-frequency vibrational and orientational modes of liquids and solutions, thus providing detailed insights into molecular interactions and reaction dynamics in the condensed phase.

Other time domain experiments contain information that is not accessible in the frequency domain. This is particularly true of photon echo methods. The name suggests a close analogy with nuclear magnetic resonance (NMR) spectroscopy, and the (optical) Bloch vector approach may be used to describe both measurements, although the transition frequencies and time-scales involved differ by many orders of magnitude. In the photon echo experiment, two or three ultrafast pulses with carefully controlled interpulse delay times are resonant with an electronic transition of the solute. In the two-pulse echo, the echo signal is emitted in the phase match direction at twice the interpulse delay, and its intensity as a function of time yields the homogeneous dephasing time associated with the transition. In the three-pulse experiment the pulses are separated by two time delays. By measuring the intensity of the stimulated echo as a function of both delay times it is possible to separately determine the dephasing time and the population relaxation time associated with the resonant transition. Such information is not accessible from linear spectroscopy, and can be extracted only with difficulty in the frequency domain. The understanding of photon echo spectroscopy has expanded well beyond the simple description given here, and it now provides unprecedented insights into optical dynamics in solution, and thus informs greatly our understanding of chemistry in the condensed phase. The methods have recently been extended to the infra red, to study vibrational transitions.

Higher-Order and Multidimensional Spectroscopies

The characteristic feature of this family of spectroscopies is the excitation of multiple resonances, which may or may not require measurements at $\chi^{(n)}$ with $n > 3$. Such experiments require multiple frequencies, and may yield weak signals, so they only became experimentally viable upon the availability of stable and reliable solid state lasers and optical parametric generators. Measurements are made in either the time or the frequency domain, but in either case benefit from heterodyne detection.

One of the earliest examples was two-dimensional Raman spectroscopy, where multiple Raman active modes are successively excited by temporally delayed pulse pairs, to yield a fifth-order nonlinear signal. The signal intensity measured as a function of both delay times (corresponding to the two dimensions) allows separation of homogeneous and inhomogeneous contributions to the line shape. This prodigiously difficult $\chi^{(5)}$ experiment has been completed in a few cases, but is plagued by interference from third-order signals.

More widely applicable are multidimensional spectroscopies using infrared pulses or combinations of them with visible pulses. The level scheme for one such experiment is shown in **Fig. 5** (which is one of many possibilities). From the scheme, one can see that the nonlinear signal in the visible depends on two resonances, so both can be detected. This can be regarded as a multiply resonant nondegenerate four-wave mixing (FWM) experiment. In addition, if the two resonant transitions are coupled, optical excitation of one affects the other. Thus, by measuring the signal as a function of both frequencies, the couplings between transitions are observed. These appear as cross peaks when the intensity is plotted in the two frequency dimensions, very much as with 2D NMR. This technique is already providing novel information on molecular structure and structural dynamics in liquids, solutions, and proteins.

Spatially Resolved Spectroscopy

A recent innovation is nonlinear optical microscopy. The nonlinear dependence of signal strength on intensity means that nonlinear processes are localized at the focal point of a lens. When focusing is strong, such as in a microscope objective, spatial localization of the nonlinear signal can be dramatic. This is the basis of the two-photon fluorescence microscopy method, where a high repetition rate source of low-energy ultrafast pulses is focused by a microscope objective into a sample labeled with a fluorescent molecule, which has absorption at half the wavelength of the incident photons (**Fig. 4**). The fluorescence is necessarily localized at the focal point because of its dependence on the square of the incident intensity. By measuring intensity while scanning the position of the focal point in space, a 3D image of the distribution of the fluorophore is constructed. This technique turns out to have a number of advantages over one photon fluorescence microscopy, most notably in terms of ease of implementation, minimization of sample damage, and depth resolution. The technique is widely employed in cell biology.

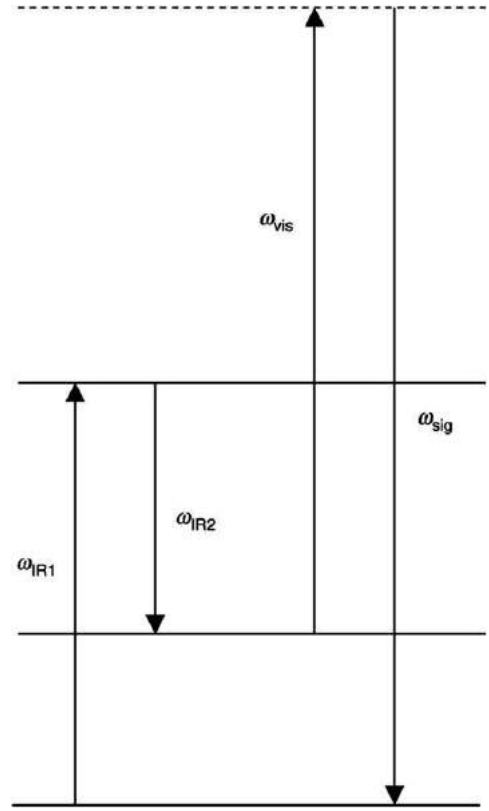

Fig. 5 Illustration of multiple resonance enhancements in a FWM geometry, from which 2D spectra may be generated.

Stimulated by the success of two-photon microscopy, further nonlinear microscopies have been developed, all relying on the spatial localization of the signal. CARS microscopy has been demonstrated to yield 3D images of the distribution of vibrations in living cells. It would be difficult to recover such data by linear optical microscopy. SHG has been applied in microscopy. In this case, by virtue of the symmetry selection rule referred to above, a 3D image of orientational order is recovered. Both these and other nonlinear signals provide important new information on complex heterogeneous samples, most especially living cells.

See also: Transient Holographic Grating Techniques in Chemical Dynamics

Further Reading

Andrews, D.L., Allcock, P., 2002. Optical Harmonics in Molecular Systems. Weinheim, Germany: Wiley-VCH.
Andrews, D.L., Demidov, A.A. (Eds.), 2002. An Introduction to Laser Spectroscopy. New York: Kluwer Academic.
Butcher, P.N., Cotter, D., 1990. The Elements of Nonlinear Optics. Cambridge, UK: Cambridge University Press.
Cheng, J.X., Xie, X.S., 2004. Coherent anti-Stokes Raman scattering microscopy: Instrumentation, theory, and applications. Journal of Physical Chemistry B 108, 827.
de Boeij, W.P., Pshenichnikov, M.S., Wiersma, D.A., 1998. Ultrafast solvation dynamics explored by femtosecond photon echo spectroscopies. Annual Review of Physical Chemistry 49, 99.
Eisenthal, K.B., 1992. Equilibrium and dynamic processes at interfaces by 2nd harmonic and sum frequency generation. Annual Review of Physical Chemistry 43, 627.
Fleming, G.R., 1986. Chemical Applications of Ultrafast Spectroscopy. Oxford, UK: Oxford University Press.
Fleming, G.R., Cho, M., 1996. Chromophore-solvent dynamics. Annual Review of Physical Chemistry 47, 109.
Fourkas, J.T., 2002. Higher-order optical correlation spectroscopy in liquids. Annual Review of Physical Chemistry 53, 17.
Hall, G., Whitaker, B.J., 1994. Laser-induced grating spectroscopy. Journal of Chemistry Society Faraday Transactions 90, 1.
Heinz, T.F., 1991. Second order nonlinear optical effects at surfaces and interfaces. In: Ponath, H.E., Stegeman, G.I. (Eds.), Nonlinear Surface Electromagnetic Phenomena, pp. 353.
Hesselink, W.H., Wiersma, D.A., 1983. Theory and experimental aspects of photon echoes in molecular solids. In: Hochstrasser, R.M., Agranovich, V.M. (Eds.), Spectroscopy and Excitation Dynamics of Condensed Molecular Systems. Amsterdam: North Holland (Chapter 6).
Levenson, M.D., Kano, S., 1988. Introduction to Nonlinear Laser Spectroscopy. San Diego, CA: Academic Press.
Meech, S.R., 1993. Kinetic application of surface nonlinear optical signals. In: Lin, S.H., Fujimura, Y., Villaeys, A. (Eds.), Advances in Multiphoton Processes and Spectroscopy, vol. 8. Singapore: World Scientific, pp. 281.
Mukamel, S., 1995. Principles of Nonlinear Optical Spectroscopy. Oxford: Oxford University Press.
Rector, K.D., Fayer, M.D., 1998. Vibrational echoes: a new approach to condensed-matter vibrational spectroscopy. International Review of Physical Chemistry 17, 261.
Richmond, G.L., 2001. Structure and bonding of molecules at aqueous surfaces. Annual Review of Physical Chemistry 52, 357.

Shen, Y.R., 1984. The Principles of Nonlinear Optics. New York: Wiley.
Smith, N.A., Meech, S.R., 2002. Optically heterodyne detected optical Kerr effect – Applications in condensed phase dynamics. International Review of Physical Chemistry 21, 75.
Tolles, W.M., Nibler, J.W., McDonald, J.R., Harvey, A.B., 1977. Review of theory and application of coherent anti-Stokes Raman spectroscopy (CARS). Applied Spectroscopy 31, 253.
Wright, J.C., 2002. Coherent multidimensional vibrational spectroscopy. International Review of Physical Chemistry 21, 185.
Zipfel, W.R., Williams, R.M., Webb, W.W., 2003. Nonlinear magic multiphoton microscopy in the biosciences. Nature Biotechnology 21, 1368.
Zyss, J. (Ed.), 1994. Molecular Nonlinear Optics. San Diego: Academic Press.

Alternative Plasmonic Materials

Gururaj V Naik, Rice University, Houston, TX, United States

The contents of this chapter are adapted from P. R. West *et al.*, *Laser & Photon. Rev.* **4**(6) 795–808 (2010), and G. V. Naik *et al.*, *Adv. Mater.* **25**(24) 3264–3294 (2013).

Introduction

Modern science has enabled techniques to precisely control the flow of heat, light, charge, spin and other quantities at nanometer scale. Each of these new capabilities have given rise to new technologies that have or on their way to revolutionize our society. For example, controlling the flow of charge at nanoscale has led to nanoelectronics – a technology that brought us into the information age. An impact of such large magnitude was possible only because scientists from many different disciplines worked together to solve major technological problems allowing us to build billions of transistors on a millimeter sized chip. Electronics in 1980s used only a few elements such as silicon, oxygen, boron, phosphorous, and aluminum to build the devices. However, within two decades, the number of elements expanded to cover more than half the periodic table. Many ground-breaking material innovations led to the use of new materials which enabled electronic devices to be simultaneously smaller and faster by nearly 3 orders. Such explosion of material innovation transformed electronics from a niche technology to a revolutionizing technology. Similar to nanoelectronics, other technologies – especially nanophotonics can make an impact of the same or even greater scale if only such big scale material innovation happens. Nanophotonics had an explosion of revolutionary concepts in the past decade with the advent of metamaterials and transformation optics. However, practical implementation of these concepts and their commercialization require material innovation. Thus, the promise of nanophotonics as the next generation technology relies upon material innovation.

In order to better appreciate the need for materials innovation in nanophotonics, it is useful to understand how nanophotonic devices operate and what they are made of. Nanophotonics enables confining light to nanoscale – a feat that is possible either by high refractive index or resonant materials. While materials with very high refractive index at optical frequencies are not available in nature, coupling light to electronic resonances in materials is a common approach. Free electron plasma as in metals can support electronic or plasmon-polaritonic resonance over a broad spectrum: infrared, visible and ultra-violet. Thus, metals are referred to as plasmonic materials and form one of the basic building blocks of nanophotonic devices. Since metals have high optical losses, nanophotonic devices often pick the metals with lowest ohmic losses – the noble metals. Thus, the material library of nanophotonics has been limited for a long time to only noble metals and some common dielectrics and semiconductors. Drawing analogy to nanoelectronics, the situation of nanophotonics resembled that of electronics in 1980s. But in the recent past, the quest for new materials in the nanophotonics library has intensified which has led to a new research direction – searching alternatives to metals. Here, in this article, the recent breakthroughs in materials innovation for nanophotonics – especially in searching alternatives to metals – will be discussed. In the next few sections, the need for alternative plasmonic materials, what they are, how to find them, and what applications do they benefit will be discussed. Given that the plasmonic material research is currently rapidly growing, this article has limited itself to summarizing the major breakthroughs in this emerging field of research.

A Case for Alternative Plasmonic Materials

Metals such as gold and silver are commonly used in plasmonic and optical metamaterial devices because of their small ohmic losses or high DC conductivity. However, at optical frequencies another loss mechanism namely interband transitions plays an important role in these metals. Loss arising from interband transitions occurs when a valence electron in a metal absorbs a photon to jump to the Fermi surface or an electron near the Fermi-surface absorbs a photon to jump to the next unoccupied conduction band. This is the loss mechanism that is responsible for the color of copper and gold. **Fig. 1** shows the imaginary part of permittivity of gold in the optical range adopted from Johnson and Christy (1972). The loss depicted by the imaginary part of permittivity can be split into two parts: interband and intraband losses. The intraband losses (or Drude losses) in gold are high in the near-infrared (NIR) and are lower for shorter wavelengths. On the other hand, interband losses in gold are high for the shorter wavelengths in the visible range. These additional losses at optical frequencies caused by interband transitions make metals such as gold unsuitable for many plasmonic and metamaterial devices.

For applications like transformation optics (TO) that require that the imaginary part of a metal's dielectric function be small, even if interband transitions are absent in the operating spectral range, intraband transitions and scattering losses are invariably present and often result in large overall losses. The high loss in conventional plasmonic materials is but one of the major disadvantages in using metals like gold and silver in plasmonics, optical metamaterials and TO. One issue is that the magnitude of the real part of the permittivity is very large in conventional metals. This is a problem in designing many TO devices because these devices often require meta-molecules with a nearly balanced polarization response (Cai and Shalaev, 2009). In other words, the polarization response from the metallic components should be of the same order as that from the dielectric components within

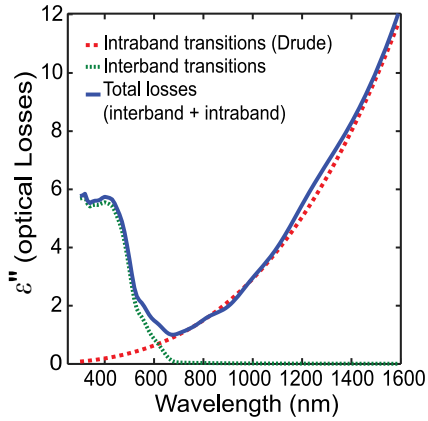

Fig. 1 The imaginary part of permittivity of gold in the optical range (solid line), data from Johnson, P.B., Christy, R.W., 1972. Phys. Rev. B 6, 4370. The individual contributions from free electron losses (intraband transitions) and interband transition losses are shown in dotted lines.

each meta-molecule. When the real parts of permittivity of the metal and dielectric are on the same order, the geometric fill fractions of the metal and dielectric can be readily tuned to match the design requirements. On the other hand, if the magnitude of the magnitude of the real part of permittivity of the metal is a few orders larger than that of the dielectric, the metal fill fraction in the meta-molecule will be a few orders smaller than that of the dielectric. This constraint would necessitate very tiny metal inclusions in the meta-molecule, which poses a number of problems especially in terms of successful nanofabrication. Thus, having smaller magnitudes of real permittivity for plasmonic materials would be advantageous in many applications. Since, the origin of the large magnitude of real permittivity in noble metals can be traced to their very large carrier concentrations, reducing the carrier concentration in metals would help applications such as TO metamaterials.

Aside from the issues of loss and not adjustable dielectric permittivity described above, metals also pose nanofabrication challenges, especially when grown as thin films. Metal thin films exhibit quite different morphologies when compared to bulk metal, which can lead to the degradation of optical properties. First of all, ultra-thin metal films deposited by common techniques such as evaporation or sputtering often grow as semi-continuous or discontinuous films. Metal films exhibit a percolation threshold in the film thickness when grown on commonly used substrates such as glass, quartz, sapphire and silicon. Overcoming this limit would require extra efforts such as using a wetting layer (Chen *et al.*, 2010) or a lattice-matched substrate (Pashley, 1959). However, these approaches have limitations in terms of design integration and scalability. Generally, thin metal films exhibit a structure composed of many small grains. In contrast, thick films are typically composed of large grains, and their optical properties resemble those of the bulk material. When the films are thin, the grainy structure causes additional grain-boundary scattering for free electrons and increases the losses in the metal. Another loss mechanism arising from the microstructure of thin metal films is related to surface roughness. Nanoscale patterning invariably results in rough surfaces and edges, which cause additional scattering and optical losses.

Losses in thin metal films can increase nearly threefold due to grain-boundary scattering. In order to avoid such additional losses in conventional plasmonic materials, single-crystal growth of noble metal films has been attempted (Wang *et al.*, 2015). The improvement in losses is evident, but it is not substantial due to the limitations that arise from the nanopatterning of the metal films.

Another important issue to consider in the context of realistic devices and integration is the chemical stability of the materials. The degradation of metals on exposure to air/oxygen or humidity would pose additional problems in fabrication and integration of devices. Among conventional plasmonic materials, silver and copper are well known to degrade in air, but gold is very stable in air. While copper forms a native oxide layer in air, silver is sensitive to sulfidation and tarnishes to form a layer of silver sulfide. The tarnishing of these metals has a direct consequence on their optical properties, and the optical losses increase, which in turn results in larger values of the imaginary part of the dielectric function.

Another important technological challenge associated with noble metals is that they are not compatible with standard silicon manufacturing processes. This precludes plasmonic and metamaterial devices from leveraging on standard nanofabrication technologies. This also diminishes the possibility of integrating plasmonic and metamaterial components with nanoelectronic components. The compatibility issue with noble metals arises from the fact that these metals can diffuse into silicon to form deep traps, which severely affects the performance of nanoelectronics devices. Hence the integration of noble metals into silicon manufacturing processes is a difficult challenge. Recently, copper has being incorporated into silicon processes, but additional, special processing steps are needed to create diffusion barriers between the silicon and the copper. Gold and silver still remain outside the realm of feasibility for silicon manufacturing processes.

Major drawback of metals is that their optical properties cannot be tuned or adjusted easily. For example, the carrier concentration of metals cannot be changed much with the application of moderate electric fields, optical fields, or temperature, etc. Hence, in applications where switching or modulation of the optical properties is essential, metals cannot be used as tunable elements.

With all the shortcomings of conventional plasmonic materials, researchers have been motivated to search for better alternatives. Many alternatives to metals have been proposed that overcome one or more of the drawbacks mentioned above. The significance of a particular alternative depends on the end application, but general criteria for the choice of an alternative plasmonic material can be outlined from the issues raised in the preceding discussions.

Optical Properties of Materials

Since the rest of the article focuses on optical constants of materials and their influence on device performance, a closer look at the optical response of materials is helpful. Since plasmonic materials support a free electron cloud/plasma, their free-electron response can be described by the Drude model (Dressel *et al.*, 2002) as shown in Eq. (1).

$$\varepsilon_{Drude}(\omega) = \varepsilon'_{Drude} + i\varepsilon''_{Drude} = \varepsilon_b - \frac{\omega_p^2}{(\omega^2+\gamma^2)} + i\frac{\omega_p^2\gamma}{(\omega^2+\gamma^2)\omega} \tag{1}$$

Here ε_b is the polarization response from the core electrons (background permittivity), ω_p is the plasma frequency and γ is the Drude relaxation rate. Relaxation rate is responsible for scattering/ohmic losses and scales directly with the imaginary part of the dielectric function. And, the square of the plasma frequency, ${\omega_p}^2$ – proportional to carrier concentration (n) – scales directly with the real and imaginary part of the permittivity. In optics, it has been noticed that the Drude scattering rate γ is a phenomenological parameter that is dependent the microstructure of the plasmonic structure. For example, the internal grain size of the metal film can increase the carrier scattering rate. When the grain size is large, as in a bulk metal, the relaxation rate is given by γ_0. When the grain size is small, as in the case of thin metal films, Eq. (2) describes the enhanced relaxation rate that arises from additional grain-boundary scattering (Kreibig and Vollmer, 1995).

$$\gamma = \gamma_0 + A\frac{v_F}{d} \tag{2}$$

where A is a dimensionless empirical constant, v_F is the Fermi-velocity of the electrons in the metal and d is the average grain size. Smaller grains result in a larger γ and, hence, higher losses.

Surface roughness is yet another factor which might lead to a larger effective γ. All such effects can be empirically described by a factor called *loss factor* as described in Eq. (3). Experimental findings suggest that a loss factor of 3 to 5 is common in nano-patterned gold and silver films.

$$\gamma = (Loss\ factor) \times \gamma_0 \tag{3}$$

While Drude model effectively captures the response from free carriers (or intraband transitions in quantum picture), the response from interband transitions of electrons are harder to model. Interband transitions form a significant loss mechanism in materials at optical frequencies, and occur when electrons jump from one energy band to a higher empty energy band by absorbing incident photons. In metals, when a bound electron absorbs an incident photon, the electron can shift from a lower energy band to the Fermi surface or from near the Fermi surface to the next higher empty energy band. Both of these processes – described as interband transitions – result in high loss at optical frequencies.

In optics, often the response of interband transitions is captured by a series of Lorentz oscillators. Lorentz oscillator is the basic model describing the light-matter interaction in a two-level system, and its form is as given by Eq. (4) (Dressel *et al.*, 2002):

$$\varepsilon_{lk} = \frac{f_{lk}\omega_{p,lk}^2}{\omega_{lk}^2 - \omega^2 - i\omega\Gamma_{lk}} \tag{4}$$

Here, f_{lk} corresponds to the strength of the oscillator at energy levels l and k, ω_{lk} is the resonant frequency corresponding to the difference between the energies of levels l and k, Γ_{lk} is the damping in the oscillator accounting for non-zero line-width of the peak, and $\omega_{p,lk}$ is similar to the plasma frequency given by Eq. (1) with the difference that n here refers to the density of states l or k. When there are many of such interacting energy levels, the effective permittivity can be expressed as a summation over all allowed Lorentzian terms. This is a popular approach referred to as Drude-Lorentz model (see Eq. (5)) to reasonably approximate the dielectric function of metals.

$$\varepsilon(\omega) = \varepsilon_{Drude\ (\omega)} + \sum_{lk}\varepsilon_{lk}(\omega) \tag{5}$$

In general, solids with a periodic lattice have electronic energy levels which exist as bands instead of discrete levels, requiring a Joint-Density-of-States (JDOS) description and integration over all the allowed transitions at any given photon energy (a more detailed discussion of this formulation is found in reference Dressel *et al.* (2002)).

Ideal Plasmonic Material

A material with a purely real and negative permittivity ($\varepsilon'' = 0$ and $\varepsilon' < 0$) would be an ideal candidate to replace metals in most of the plasmonic and metamaterial devices. Such a material produces a metallic response to light while exhibiting zero losses. However, it is impossible to have zero losses and negative permittivity simultaneously for all frequencies in any dispersive material

due to the causality condition. All is not lost, though – there can be a frequency interval in which the permittivity is purely real and negative. This is possible without violating causality only when there are large losses present at lower frequencies. A theoretical solution satisfying this requirement has been proposed by Khurgin and Sun (2010). The fact that none of the known metals have this specific band structure explains why natural metals are always with associated losses. In other words, there are no naturally occurring materials that are simultaneously lossless and have negative real permittivity in any part of spectrum. Reference Khurgin and Sun (2010) suggests a way to engineer sodium metal for zero losses in the NIR spectrum: stretching sodium lattice by nearly two times. However, it is not clear if such extreme engineering is practical. Thus, a zero loss-plasmonic material is elusive, and a practical alternative is to minimize the losses to an extent it is tolerable.

The dielectric function described by the Drude model (Eq. (1)) suggests that reducing γ can directly scale down the losses. Many possibilities have been explored in reducing the carrier-damping losses in metals including cooling metals to cryogenic temperatures (Bouillard *et al.*, 2012) and using monolithic metal crystals (Wang *et al.*, 2015). However, the reduction in losses was either not adequate or its practical implementation was a challenge (as in cooling).

Another possibility for reducing loss in metals is to reduce ω_p, which can reduce the magnitudes of both of the real permittivity and the imaginary permittivity. However, ω_p should not be reduced too much to retain a metallic response in the desired wavelength range. This lower limit on ω_p can be deduced from the Drude-model description of $\varepsilon'(\omega)$. If we require metallic behavior ($\varepsilon'(\omega) < 0$) for frequencies less than the cross-over frequency ($\omega < \omega_C$), then $\omega_p > \sqrt{\varepsilon_b(\omega_C^2 + \gamma^2)}$. This also sets the lower limit on the imaginary part of permittivity: $\varepsilon''(\omega) > \frac{\varepsilon_b \gamma}{\omega_C}$ for $\omega < \omega_C$. Both of these limits are useful in assessing and engineering different materials and their optical properties to produce alternative plasmonic materials.

There are two possibilities in producing alternative plasmonic materials based on the Drude description. One of them is to dope semiconductors heavily and create enough free carriers that the material's optical properties become metallic in the desired wavelength range. The other option is to remove excess free carriers from metals so as to reduce the carrier concentration to the desired value. Both of these techniques have their own benefits from the perspective of material design and are discussed in the following sections. In addition to small losses, an ideal plasmonic material would possess other useful properties such as adjustable real permittivity, thermal and chemical stability, easy to fabricate nanostructures and integrate them into devices, and earth abundance in availability. Certain applications may demand additional properties such as CMOS compatibility and bio-compatibility. In the following sections, promising material candidates are discussed after a brief review of conventional plasmonic materials.

Conventional Plasmonic Materials

Metals are candidates for plasmonic applications because of their high conductivity. Among metals, silver and gold are the two most often used for plasmonic applications due to their relatively low loss in the visible and NIR ranges. In fact, almost all of the significant experimental work on plasmonics has used either silver or gold as the plasmonic material. While metals other than silver and gold have been used in plasmonics, their use is quite limited, as their losses are higher than those of silver and gold. Blaber *et al.* (2010) review plasmonic properties properties of many metals and intermetallics, and present a table listing the highest figure-of-merit of metals in the periodic table. The figure-of-merit defined for localized surface plasmon applications (Q_{LSP}) is defined as the negative ratio of the real to the imaginary parts of the dielectric function. It is clear from reference Blaber *et al.* (2010) that silver and gold perform the best followed by copper, aluminum and alkali metals. Given the high chemical reactivity of alkali metals, silver, gold, copper and aluminum have been the conventional plasmonic materials.

Among all metals, silver has the lowest optical loss (see Drude damping rate in **Table 1**) in the visible and NIR ranges. However, in terms of fabrication, silver degrades relatively quickly and the thickness threshold for uniform continuous films is around 12–23 nm, making it difficult to integrate silver into nanophotonic devices. Additionally, silver losses are strongly dependent on the surface roughness. Gold is the next-best material in terms of loss in the visible and NIR ranges. Compared with silver, gold is chemically stable and can form a continuous film even at thicknesses around 1.5–7 nm, however losses are at least three times higher than in silver. Silver and gold films can be fabricated by various physical vapor deposition (PVD) techniques such as electron-beam/thermal evaporation and sputtering. Nanoparticles and metal-coated nanoparticles can be synthesized by colloidal and electrochemical techniques.

Copper has the second-best conductivity among metals (next to silver), and is expected to exhibit promising plasmonic properties. Indeed, ε'' of Cu is comparable to that of gold from 600 to 750 nm. Considering the cost of silver and gold, copper would be a good alternative to silver and gold if its chemical stability were better. Copper easily oxidizes and forms Cu_2O and

Table 1 Drude model parameters for metals. ω_{int} is the frequency of onset for interband transitions. Drude parameters tabulated are not valid beyond this frequency

	ε_b	ω_p (eV)	γ (eV)	ω_{int} (eV)
Silver (41,62,63)	3.7	9.2	0.02	3.9
Gold (41,63)	6.9	8.9	0.07	2.3
Copper (41,62,63)	6.7	8.7	0.07	2.1
Aluminum (64,65)	0.7	12.7	0.13	1.41

CuO which degrades its optical properties as well. However, copper is CMOS compatible and if oxidation is prevented by protective layers, copper can be a good plasmonic material.

Aluminum has not been an attractive plasmonic material due to the existence of an interband transition around 800 nm (1.5 eV), resulting in large ε'' values in the visible wavelength range (see **Table 1**). However, in the UV range, ε' is negative even below 200 nm where ε'' is still relatively low. Thus, aluminum is a better plasmonic material than either gold or silver in the blue and UV range. Similar to copper, aluminum is easily oxidized. Aluminum very rapidly forms a thin self-limiting layer of aluminum oxide (Al_2O_3) under atmospheric conditions, making device fabrication challenging. Despite these challenges, aluminum is an earth-abundant, inexpensive and CMOS compatible metal attracting significant attention in the recent past (Knight *et al.*, 2014).

Conventional plasmonic materials – gold, silver, aluminum, copper and other – offer a broad set of optical and material properties useful for many nanophotonic applications. However, the emerging nanophotonic devices such as TO metamaterials need even broader set of properties. As outlined in the previous section, plasmonic materials with adjustable real permittivity, smaller imaginary permittivity, chemically stable, and easy to fabricate and integrate into devices are necessary. Thus, finding alternatives to conventional plasmonic materials is necessary.

Finding Alternatives

Designing materials with a given set of properties is not easy, and often researchers employ a variety of strategies to tackle this problem. For example, in the section on ideal plasmonic materials, we noticed how smaller carrier concentration in plasmonic materials can lead to lower imaginary permittivity and smaller magnitude of real permittivity. Two strategies to achieve the goal of lowering or adjusting the carrier concentration would be: doping semiconductors to create just enough free carriers, and removing excess free carriers from conventional plasmonic materials by alloying and compounding. The following subsections elaborate on these two approaches to design plasmonic materials. In addition, a new class of materials – 2D or ultrathin materials – will be discussed as candidates for alternative plasmonic materials.

Semiconductors to Metals

Increasing the carrier concentration in semiconductors enough to cause them to behave like metals is accomplished through doping. Achieving metal-like optical properties ($\varepsilon' < 0$) in the spectrum of interest puts a lower limit on the carrier concentration that must be achieved by doping. The optical response of free carriers as described by the Drude model (Eq. (1)) can be used to estimate this minimum carrier concentration. To obtain metal-like properties ($\varepsilon' < 0$) for $\omega < \omega_C$, the lower limit on the plasma frequency (ω_p) and hence the carrier concentration (n) is given by Eq. (4)

$$\omega_p^2 > \varepsilon_b\left(\omega_C^2 + \gamma^2\right),\ n > \frac{\varepsilon_0 m^*}{e^2}\varepsilon_b\left(\omega_C^2 + \gamma^2\right), \tag{4}$$

where ε_0 is the vacuum permittivity, e is the electron charge, and m^* is the effective mass of the carrier. **Table 2** shows the estimates of carrier concentration that would be required for many common semiconductors in order to obtain $\varepsilon' = -1$ at the technologically important telecommunication wavelength ($\lambda = 1.55$ μm). In this scenario, ω_C (cross-over frequency) is slightly higher than the telecommunication frequency. For many of the common semiconductors such as silicon, the minimum carrier concentration to obtain metal-like optical properties in the NIR is about 10^{21} cm^{-3}. Smaller ε_b and m^* values would slightly reduce the minimum carrier concentration, but a high carrier density is inevitable for achieving metal-like properties in the NIR. Such high

Table 2 Comparison of different heavily-doped semiconductors as potential alternative plasmonic materials. The table evaluates the carrier concentration required to reduce the real permittivity of semiconductors to $\varepsilon' = -1$ at telecommunication wavelength ($\lambda = 1.55$ μm). The electronic parameters of the semiconductors reported in the literature are used in Drude model to evaluate the optical properties in the NIR.

Material	*Background permittivity (ε_b)*	*Carrier mobility when heavily doped (cm^2/V-s)*	*Effective mass (m*)*	*Relaxation rate (eV)*	*Carrier concentration required to achieve Re{ε} = −1 at λ = 1.55 μm (x 10^{20} cm^{-3})*	*Im{ε} or losses at λ = 1.55 μm*
n-Si	11.70	80	0.270	0.0536	16.0	0.8508
p-Si	11.70	60	0.390	0.0495	23.1	0.7853
n-SiGe	15.10	50	0.24	0.0965	18.2	1.9414
n-Gaas	10.91	1000[a]	0.068	0.017	3.76	0.2534
p-GaAs	10.91	60	0.44	0.0438	24.4	0.6528
n-InP	9.55	700	0.078	0.0212	3.82	0.2796
n-GaN	5.04	50	0.24	0.0965	6.83	0.7283
p-GaN	5.24	5	1.4	0.1654	42.3	1.290
Al: ZnO	3.80	47.6	0.38	0.064	8.52	0.384
Ga: ZnO	3.80	30.96	0.38	0.0984	8.59	0.5904
ITO	3.80	36	0.38	0.0846	8.56	0.5077

[a]Mobility corresponding to the free electron concentration of 1×10^{19} $cm.^{-3}$

carrier densities require ultra-high doping densities, which poses major limitations. The challenges in ultra-high doping will be discussed in the latter part of this section.

Another issue to be considered when choosing a semiconductor for creating metal-like behavior is the mobility of the carriers. The Drude relaxation rate (γ) can be related to the mobility (μ) of the charge carrier at a given optical frequency by Eq. (5).

$$\gamma = e/\mu m^* \tag{5}$$

Reducing the damping loss requires that the product of mobility and effective mass must be as large as possible. Mobility degrades significantly with higher doping levels due to increased impurity scattering. Hence, it is important to consider appropriate mobility numbers when assessing various semiconductors. **Table 2** shows the approximate values of γ evaluated using Eq. (5) where low-frequency m^* and high-doping mobility (about 10^{19} cm^{-3}) values are used. The damping losses in semiconductors are comparable to those of bulk gold and silver as reported by Johnson and Christy (1972).

Another consideration in choosing semiconductors is their optical bandgap. The optical bandgap corresponds to the onset of interband transitions, which cause additional optical losses. Hence, the optical bandgap needs to be larger than the frequency spectrum of interest. At telecommunication wavelength of 1.55 μm, the corresponding photon energy is 0.8 eV. Clearly there are many semiconductors with bandgaps greater than 0.8 eV that can be useful for applications in the NIR and longer wavelengths only if they can be doped heavily. From **Table 2**, we note that the only major bottleneck in turning semiconductors into low-loss plasmonic elements at optical frequencies is accomplishing the required ultra-high doping. A deeper understanding of the doping mechanism in semiconductors can provide insights for accomplishing this ultra-high doping, and hence we turn our attention to this process next.

Doping is a process of incorporating foreign atoms or impurities into the lattice of a semiconductor to controllably change the properties of the semiconductor. When certain atomic species are incorporated into the lattice sites of the semiconductor, the free-carrier concentration in the material can be changed proportionally. For our purposes, we will only consider doping that acts to increase the free-carrier concentration.

When dopant atoms replace (substitute) the semiconductor atoms in the lattice, this form of doping is called substitutional doping, which effectively contributes to an increase in the charge-carrier density in the material. When the dopant atom occupies an interstitial site, it is called interstitial doping. Interstitial doping is an ineffective doping method because it does not contribute any free carriers and, therefore, does not producing any electrical doping. There is another mechanism called doping compensation that can also result in ineffective doping. In this mechanism, dopants such as silicon in GaAs can behave as both an p-type and n-type dopant and self-compensate any net doping effect.

Doping is limited on the higher side by the solid solubility of the dopant in the semiconductor. Introducing dopants more than the solid-solubility limit may result in phase separation of dopants or compounds of dopants. This phase separation of excesses dopants leads to ineffective doping. Often, ultra-high doping results in more crystal defects, which could act as traps for free carriers. These trap states can counter the effect of doping and thus reduce the carrier concentration. In general, the doping mechanism is complicated and, most often, not all dopant atoms succeed in contributing free charge carriers due to a number of reasons, including those mentioned above.

Doping efficiency is an important quantity and can be defined as the fraction of dopants that contribute to the charge carrier density. Doping efficiency decreases sharply for high doping concentrations in many semiconductors, making ultra-high doping a tough challenge. Although there are many different material engineering approaches that can provide elegant solution to this problem (e.g., delta doping (Weir *et al.*, 1994)), a straightforward approach to the problem of achieving ultra-high doping would be to choose a dopant that has very high solid-solubility limit in the selected semiconductor material. This approach has been utilized in demonstrating many semiconductor-based plasmonic materials. Heavily doped silicon, III-V systems based on GaAs, and transparent conducting oxides (TCOs) such as indium-tin-oxide (ITO), and aluminum zinc oxide (AZO) are widely explored as alternative plasmonic materials.

III-V semiconductors

Compounds of III-V semiconductors, such as those based on GaAs and InP, are semiconductors with an optical bandgap in the NIR. This is useful to know because a bandgap comparable to that of silicon results in a value of ε_b that is comparable to that of silicon for these materials. In addition, the electron mobility is very high in these materials due to a small effective mass. This relaxes the carrier concentration requirement for observing metallic properties in the NIR (see **Table 1**). In these materials, a carrier concentration in excess of 10^{20} cm^{-3} is required to observe plasmonic properties at the telecommunication wavelength. However, such high doping is very challenging in these materials owing to lower solid solubilities of the dopants and poor doping efficiency (Walukiewicz, 2001). Doping higher than 10^{19} cm^{-3} is known to produce effects such as doping compensation. For instance, silicon is a common n-type dopant in GaAs, but at high doping levels, silicon can behave not only as an n-type substitutional dopant, but also as a p-type dopant. Thus, silicon can compensate itself at high doping densities, resulting in low doping efficiency. N-type doping beyond 10^{19} cm^{-3} is clearly difficult in GaAs films. Many other studies report similar trends, and it was pointed out that the carrier concentration in GaAs shows a saturating trend for donor densities higher than 5×10^{18} cm^{-3}. On the other hand, p-doping with Be or C can be used to achieve carrier concentrations in excess of 10^{20} cm^{-3} (Yamada *et al.*, 1989). However, holes have a higher effective mass and poor carrier mobility, which raises the minimum bar on carrier concentration to turn p-GaAs plasmonic at the telecommunication frequency (see **Table 1**). Other III-V semiconductor families such as InGaAs, AlGaAs, GaInSb, InP, and GaInAsP have properties similar to GaAs and doping them high is challenging. However, they all have high electron mobility and make low loss plasmonic materials in the infrared.

Transparent conducting oxides

Oxide semiconductors such as zinc oxide, cadmium oxide and indium oxide can be highly doped to make them conducting films (Minami, 2005). Since these semiconductors have a large bandgap, they are transparent in the visible range. Hence, these materials are known as transparent conducting oxides (TCOs). Because TCOs can be doped very heavily, TCOs exhibit high DC conductivity. It is exactly this property that gives them metal-like optical properties in the NIR range. Like any other semiconductor, the optical properties of TCOs can be tuned by changing the carrier concentration/doping. They can be grown into thin films and many different nanostructures, polycrystalline and crystalline structures, patterned by standard fabrication procedures and integrated with many other standard technologies. Thus, TCOs form an obvious choice as alternative plasmonic materials in the NIR. Among the many TCOs, previous studies have shown that heavily-doped ZnO and ITO are good candidates for NIR applications (Naik and Boltasseva, 2010). Hence, in the subsequent discussion on TCOs, we focus on doped ZnO and ITO.

Thin films of TCO materials may be grown by many different techniques such as sputtering, laser ablation, evaporation, solution processing and chemical vapor deposition. Since TCOs are non-stoichiometric oxides, their properties depend on the deposition technique used. Deposition schemes such as laser ablation and sputtering are more suitable in cases where the stoichiometry needs to be controlled to produce desired film properties. In an example, Naik *et al.* deposit aluminum-doped ZnO (AZO), gallium-doped ZnO (GZO) and ITO thin films using the pulsed laser deposition (PLD) technique. **Fig. 2** shows the carrier concentration in these TCO films as a function of doping concentration. The carrier concentration was extracted by fitting the Drude model to the measured data. Optical characterization was carried out using a spectroscopic ellispometer (J.A. Woollam Co.). The films were deposited under an oxygen partial pressure of 1 mTorr. Their results indicate that AZO shows lower losses with higher doping because of improved crystallinity in the highly-doped films. However, the highest carrier concentration achieved is smaller in AZO than in ITO or GZO. The losses are higher with GZO and ITO, but much higher carrier concentrations are possible in these films. When the oxygen partial pressure was decreased below 1 mTorr, the resultant AZO films exhibited higher carrier concentrations and performed the best compared to the other two materials. **Fig. 3** shows the optical properties of TCO films extracted from ellipsometry measurements. The retrieval of the dielectric functions of TCOs was based on Drude-Lorentz model. Drude model was added to account for the free carriers and the Lorentz oscillator was added in UV to account for the interband transitions at the band edge. **Table 3** shows the parameters of Drude-Lorentz model for AZO, GZO and ITO retrieved from ellipsometry measurements. Clearly, AZO is the TCO with the lowest loss, followed by ITO and then GZO.

Other than semiconductors and TCOs, there are non-stoichiometric compounds such as In_2O_3, TiO_2, IrO_2 and metallic VO_2 that are intrinsically doped. All of these materials can support large carrier concentrations ($> 10^{19}$ cm^{-3}) to provide Drude metal-like behavior, though their losses are typically higher. However, there are niche applications in which these materials can be very useful as plasmonic components. For example, perovskite oxides show magnetoresistance properties useful for data storage applications (Rao, 1998). Vanadium dioxide shows a metal-semiconductor transition at temperatures smaller than 100 °C, which can be useful in switching applications (Dicken *et al.*, 2009).

Metal Alloys

Metallic alloys, intermetallics and metallic compounds are potential candidates for alternative plasmonic materials owing to their large free electron densities. Because of the strong plasmonic performance of noble metals, one approach to improve these materials is by shifting their interband transitions to another (unimportant) part of the spectrum. This can be achieved by

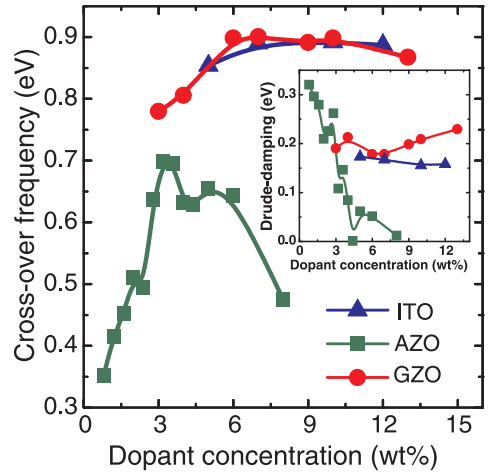

Fig. 2 Cross-over frequency (frequency at which Re(ε) = 0) or screened plasma frequency) as a function of dopant concentration in three TCOs: indium-tin-oxide (ITO), Ga:ZnO (GZO) and Al:ZnO (AZO) (adapted from reference Naik, G.V., Kim, J., Boltasseva, A., 2011. Opt. Mater. Express 1, 1090.) The inset shows the Drude damping coefficient as a function of dopant concentration. All the films were deposited using pulsed laser deposition and the optical parameters are extracted using a spectroscopic ellipsometer.

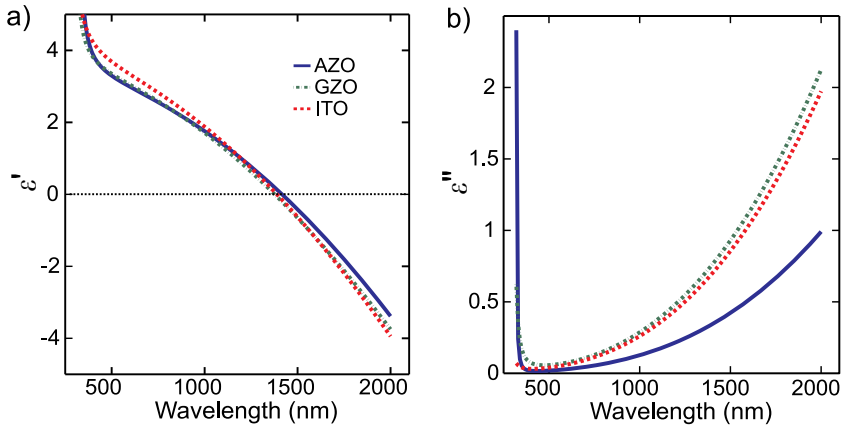

Fig. 3 Optical constants of TCO thin films deposited using pulsed laser deposition. The optical constants are measured using spectroscopic ellipsometry. (a) Real (b) Imaginary parts of dielectric function of TCO films: Al:ZnO (2 wt%), Ga:ZnO (4 wt%) and ITO (10 wt%). The deposition conditions were optimized to produce lowest losses and highest plasma frequency. Figure reproduced with permission from reference Naik, G.V., Kim, J., Boltasseva, A., 2011. Opt. Mater. Express 1, 1090.

Table 3 Drude-Lorentz parameters of five alternative plasmonic materials retrieved from ellipsometry measurements. The dielectric function of the materials may be approximated in the wavelength range of 350–2000 nm by the equation: $\varepsilon_m(\omega) = \varepsilon_b - \frac{\omega_p^2}{\omega(\omega+i\gamma)} + \frac{f_1\omega_1^2}{\omega_1^2-\omega^2-i\omega\Gamma_1}$, where the values of the parameters are as listed in the table

	AZO (2 wt%)	*GZO (4 wt%)*	*ITO (10 wt%)*	*TiN*	*ZrN*
ε_b	3.5402	3.2257	3.528	4.855	3.4656
ω_p (eV)	1.7473	1.9895	1.78	7.06	8.018
γ (eV)	0.04486	0.1229	0.155	0.291	0.5192
f_1	0.5095	0.3859	0.3884	2.125	2.4509
ω_1 (eV)	4.2942	4.050	4.210	3.862	5.48
Γ_1 (eV)	0.1017	0.0924	0.0919	1.45	1.7369

alloying/reacting two or more elements to create unique band structures that can be fine-tuned by adjusting the proportion of each alloyed material/reactant.

Alloys of noble metals have received significant attention given that noble metals are the conventional plasmonic materials. Studies on alloys of conventional plasmonic materials such as Ag–Cu, Au–Ag, Ag–Al, Au–Cd, Ag–In, and Ag–Mg show that the optical properties generally become worse than the best of the metals being alloyed. However, alloying helps in shifting the interband transitions where a simple mixing rule has been observed to be a good approximation. Similar studies and observations have been made on various intermetallics as summarized in reference Blaber *et al.* (2010). It has been concluded that intermetallics with more atoms in a unit cell usually suffer from higher optical losses. Thus, gold and silver continue to dominate as metallic plasmonic materials.

Metals to "Less-Metals"

Another approach to decrease the excessive optical losses in metals is to reduce the carrier concentration in the material. A smaller carrier concentration may be achieved in intermetallic or metallic compounds, which we will term "dilute metals" or "less-metals" in this work. In general, introducing non-metallic elements into a metal lattice reduces the carrier concentration. This also alters the electronic band structure significantly and might result in undesirable consequences, such as large interband absorption losses and higher Drude-damping losses. However, there is a lot of room for optimization, and this direction of research has yet to receive much attention from researchers. There are many advantages of plasmonic intermetallic or metallic compounds such as tunability and ease of fabrication and integration, which could out-weigh the disadvantages of these materials. Thus, it is worth pursuing a search for materials that are dilute metals.

A survey of the literature shows that the optical properties of many different classes of dilute-metals have been studied previously. Among them are metal silicides and germanides, ceramics such as oxides, carbides, borides and nitrides, which were all found to exhibit negative real permittivities in different parts of the optical spectrum. None of these materials possess all the desirable properties, and in particular none of them exhibit low optical losses. However, there has been a growing interest in materials that are important from the view point of technological applications. Specifically, silicides, and metal nitrides which are already used in silicon CMOS technology have received significant attention.

The dielectric functions of these silicides show metallic behavior in the mid-infrared (MIR), NIR and visible ranges. The losses are quite high in these materials due to interband transitions, especially in the NIR where their real permittivities are small in magnitude. Comparing to noble metals, silicides exhibit very high losses. While they are useful at longer wavelengths ($\lambda > 3$ μm) where there are no interband transitions, their real permittivity values are large in magnitude, similar to noble metals. Nevertheless, the manufacturing advantages of silicides make them promising materials for infrared plasmonics. Recently, Soref *et al.* proposed using $PdSi_2$ as plasmonic material for MIR plasmonics (Soref *et al.*, 2008) and Cleary *et al.* demonstrated an application of silicides for an SPP-based infrared biosensor where $PdSi_2$ was used as a plasmonic material (Cleary *et al.*, 2008). Blaber *et al.* (2010) provided a brief review of silicides for plasmonic applications which shows that the calculated quality factors for LSPR applications can be high for some silicides such as $TiSi_2$.

Metal nitrides such as titanium nitride, zirconium nitride, tantalum nitride and hafnium nitride exhibit metallic properties in the visible and longer wavelengths (Naik *et al.*, 2011). These are non-stoichiometric, interstitial compounds with large free-carrier concentrations. These materials are also refractory, stable and hard and allow for tuning of their optical properties by varying their composition. Technologically, they are important because many of them are currently used in silicon CMOS technology as barrier layers in copper-Damascene processes and as gate metals to n-type and p-type transistors. Thus, metal nitrides offer fabrication and integration advantages which could be useful in integrating plasmonics with nanoelectronics.

Refractory metal nitrides can be deposited by many techniques such as chemical vapor deposition (CVD), atomic layer deposition (ALD), physical vapor deposition (PVD) such as reactive sputtering pulsed laser deposition and ion-assisted reactive evaporation, and wet-chemical techniques. Since metal nitrides have nearly three orders of magnitude smaller surface energy, they can be grown into ultra-thin (< 10 nm) and epitaxial or single crystal films. Many of these nitrides are grown epitaxially on various substrates such as c-sapphire, MgO and Si. Epitaxial growth leading to ultra-smooth, ultra-thin films is very important for plasmonic applications, and noble metals usually fail in this aspect.

Optical properties of metal nitrides are receiving significant attention only recently. Naik *et al.* recently studied the optical properties of titanium nitride, tantalum nitride, hafnium nitride and zirconium nitride (Naik *et al.*, 2011). **Fig. 4** shows the dielectric functions of thin films of these metal nitrides deposited by DC reactive sputtering. The dielectric functions were retrieved from the spectroscopic ellipsometer measurements on a thin film sample. The Drude-Lorentz model parameters of these metal nitrides extracted from ellipsometry measurements are listed in **Table 3**. Note that the parameters listed for TiN in the table are for films with improved optical properties. Small improvements in the deposition conditions such as chamber seasoning allowed better quality TiN films. For all nitride films excepting titanium nitride, the deposition was carried out in nitrogen-deficient ambient, which resulted in metal-rich films. When the composition of sputtering ambient was varied from 100% nitrogen to 20% nitrogen and 80% argon, all the metal nitrides except titanium nitride exhibited optical properties that changed from dielectric behavior to metallic behavior. Titanium nitride shows a lower carrier concentration under nitrogen-rich deposition conditions. However, the metallic property in TiN is not lost because the change in the carrier concentration is small. It is also important to note that most of these nitrides have larger carrier relaxation rates than those of the noble metals, which increases the optical losses in these nitrides. Additionally, interstitial nitrides exhibit interband transition losses in the visible and ultraviolet ranges. Thus, their optical properties are undesirable in these ranges unless significant growth-parameter optimization is employed. The deposition conditions can change the Drude relaxation rate, the plasma frequency and the strength of the interband transitions. While nitrogen-rich films showed a monotonic decrease in the plasma frequency with increasing nitrogen content, the dependence was not monotonic for the other parameters. Thus, careful optimization of the deposition conditions is necessary to achieve a low-loss TiN film.

Surface plasmon-polaritons are shown to be supported by titanium nitride films. Steinmüller-Nethl *et al.* demonstrated SPP excitation on an approximately 30-nm-thick TiN film deposited by sputtering (Steinmüller-Nethl *et al.*, 1994). In that study, SPP

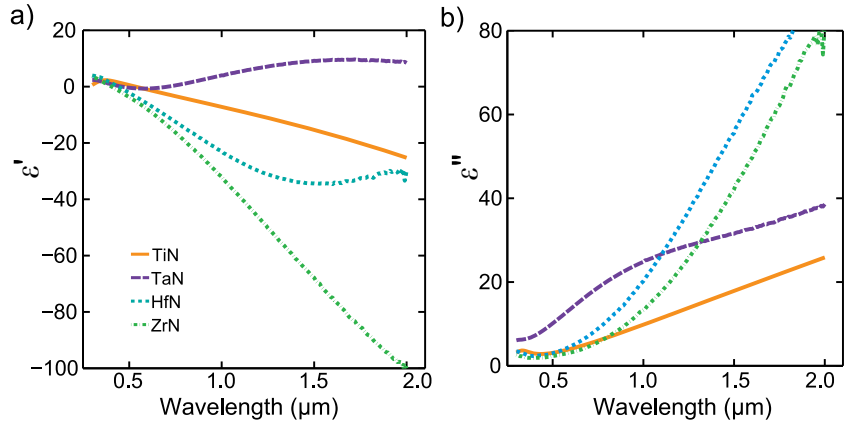

Fig. 4 Optical constants of metal nitride thin films deposited using DC reactive sputtering. The (a) real and (b) imaginary parts of dielectric functions extracted from spectroscopic ellipsometry measurements. Figure reproduced with permission from reference Naik, G.V., Kim, J., Boltasseva, A., 2011. Opt. Mater. Express 1, 1090.

excitation was demonstrated at wavelengths of 700, 750, 800 and 850 nm. Hibbins *et al.* extended this work to many wavelengths in the range from 500 to 900 nm (Hibbins *et al.*, 1998). Recently, Chen *et al.* demonstrated SPP excitation on TiN films in the visible wavelengths using the Kretschmann geometry (Chen *et al.*, 2011). Naik *et al.* also reported SPP coupling on epitaxial TiN thin films deposited by DC reactive sputtering on c-sapphire substrates (Naik *et al.*, 2012b). The SPP modes were coupled onto the TiN film using a dielectric grating coupler formed by patterning a thin layer of polymer (electron-beam resist, ZEP-520A) on top of the TiN film. The quality factor of surface plasmon polaritons on TiN thin films were predicted to be comparable to that on gold films. Given that TiN is a much harder material than gold, stable at temperatures in excess of 1200 °C, and possess low surface energy allowing it to form epitaxial ultra-thin films, TiN has systematically started to replace gold in plasmonic devices.

Though 'less-metals' such as metal nitrides exhibit plasmonic behavior in the optical range, their optical loss is no smaller than that of silver. Nevertheless, materials such as TiN have optical properties similar to gold, the second best noble metal in terms of losses. Given the other advantages of metal nitrides of noble metals, metal nitrides hold great promise as alternative plasmonic materials in the visible and NIR ranges.

Two-Dimensional Plasmonic Materials

Recently, plasmonics with 2D materials, or flatland plasmonics has gained much attention (Maier, 2012). Two-dimensional materials such as graphene have many advantages over bulk three dimensional (3D) materials from both scientific and technological perspectives. The optical properties of 2D materials are quite different from those of bulk, 3D materials, which results in significantly different plasmon dispersion relationships. Two-dimensional materials are technologically important because their properties are useful in many other applications, including electronics. The properties of 2D materials can be dynamically tuned by electrical, chemical, electrochemical and other means. For example, the optical properties of graphene can be tuned by electrical field-effect methods (Novoselov *et al.*, 2012). In addition, 2D materials can be processed via conventional, planar fabrication techniques. Currently, the synthesis of large-area single, crystalline 2D materials is a challenge that limits the set of suitable applications for these materials. However, there is now significant effort invested in overcoming this bottleneck.

There are reports of the observation of plasmons in 2D electron gas (2DEG) systems such as semiconductor inversion layers, semiconductor heterostructures, and graphene. Except in the case of graphene, plasmons were observed in other cases only at low temperatures because of the high losses (low carrier mobility or high carrier scattering rates) occurring at room temperature. All of the observations of plasmons were made in the MIR or longer wavelength ranges. Plasmons in the visible or NIR ranges were not observed in any of these 2D materials because of insufficient carrier densities.

Applications

The search for alternative plasmonic materials has already led to the discovery of many new materials for plasmonics. Yet, the exploration continues because there is not one material that works for all applications. Depending on the dominant physical phenomenon underlying the application, the performance metrics differ and so do the required optical properties of the plasmonic material. Hence, different plasmonic materials work the best for different applications. Based on the applications, plasmonic devices can be broadly classified into integrated plasmonics, sub-diffraction imaging, chemical sensing, high-temperature plasmonics, and metamaterials & transformation optics. Plasmonic materials for each of these applications are reviewed in the following.

Integrated Plasmonics

Integrated plasmonics is an analogue of integrated electronics aimed at information processing applications. Plasmonics can confine light to nanoscale and offer huge bandwidth unlike electronics, making plasmonics a promising technology to substitute/complement nanoelectronics. Some of the on-chip plasmonic devices are interconnects or waveguides, resonators, de-mux/multiplexers, and modulators or switches. Except for switches, all other components are passive devices and rely upon propagation of SPPs.

SPP propagation at metal/dielectric interface is characterized by two important metrics: mode confinement and propagation length. A small mode size and long propagation length are desirable, though these characteristics trade-off with each other. In general, for SPP waveguiding applications, the real and imaginary parts of the metal permittivity influence the trade-off between confinement and propagation loss significantly. A larger negative real permittivity gives rise to smaller field penetration into the metal and larger mode size, and a smaller imaginary permittivity leads to lower losses. Thus, the figure of merit would take the form $[\mathrm{Re}(\varepsilon_m)]^2/\mathrm{Im}(\varepsilon_m)$, where ε_m is the permittivity of the plasmonic material. This quantity is plotted for different alternative plasmonic materials in **Fig. 5**. Noble metals outperform all other materials because of their large negative real permittivity. The large negative real permittivity is a consequence of large plasma frequency or large carrier concentration. Since the figure-of-merit, $[\mathrm{Re}(\varepsilon_m)]^2/\mathrm{Im}(\varepsilon_m)$ exhibits stronger dependence on real permittivity than imaginary permittivity, having a larger plasma frequency results in a more significant improvement in performance than reducing the losses. Because all the alternative plasmonic materials we considered have lower plasma frequencies than noble metals, their performance metrics as integrated plasmonic elements are worse than those offered by noble metals.

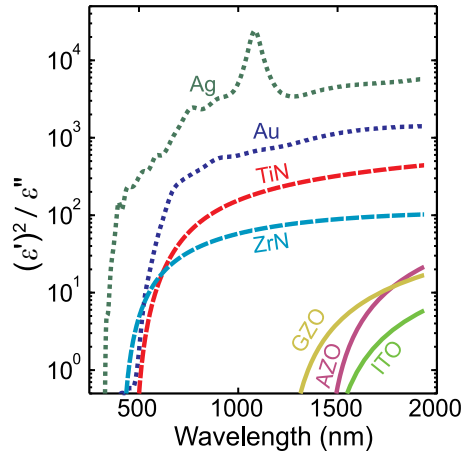

Fig. 5 The performance of many 2D plasmonic waveguides, non-spherical LSPR structures, resonant metamaterial devices such as negative-index-metamaterials depend on the quantity $\frac{Re(\varepsilon_m)^2}{Im(\varepsilon_m)}$ (ε_m is the permittivity of plasmonic material). This quantity is plotted in the figure for various plasmonic materials.

However, alternative plasmonic materials offer advantages like CMOS compatibility which could be a bigger factor than a slightly lower figure-of-merit. Aluminum, a CMOS compatible metal has been used in plasmonic systems for SPP waveguiding (MacDonald *et al.*, 2008), color filters and hot carrier photo-detectors (Zheng *et al.*, 2014). Copper waveguides are also shown as CMOS-compatible solutions to SPP waveguiding (Fedyanin *et al.*, 2016). Titanium nitride, another CMOS compatible material is also explored for integrated plasmonic applications. Kinsey *et al.* demonstrated TiN interconnects whose performance was comparable to SPP waveguides made from gold (Kinsey *et al.*, 2014). Though alternative plasmonic materials have not been able to perform as well as or better than silver in integrated plasmonic applications, they offer other advantages which makes them promising.

Optical modulator being the key element in an information processing application, plasmonic modulators offer superior performance in terms of modulation contrast. Semiconductor based plasmonic materials such as ITO and 2D material – graphene has been shown to work well as electro-optic materials. Often, these switchable materials together with other plasmonic metals have been incorporated into modulators. While silicon photonic modulators have extinction ratio in the order of 1 dB/mm, plasmonic modulators with extinction ratio greater than 3 dB/μm have been realized (Liu *et al.*, 2015). When a plasmonic modulator is incorporated into silicon photonics, high insertion loss has been observed due to the mismatch in plasmonic and photonic modes. However, in a completely plasmonic technology, insertion loos would not be a problem. Thus, alternative plasmonic materials are promising for nanoscale light switches.

Chemical Sensing

Chemical sensing applications rely upon strong field enhancement near plasmonic resonators. Metallic nanostructures exhibiting localized surface plasmon resonance (LSPR) concentrate light to nanoscale enhancing the local electric field by many times. Using a dipole approximation in the electrostatic limit, the near-field enhancement, absorption cross-section and extinction cross-section of spherical plasmonic nanoparticles can be calculated for conventional and alternative plasmonic materials. **Fig. 6** plots the maximum field enhancement ($|E_{max}/E_0|$) on the surface of spherical nanoparticles of gold, silver, titanium nitride, zirconium nitride and TCOs of diameter $\lambda/10$. The refractive index of the host medium surrounding the nanoparticles is assumed to be 1.33. Clearly, silver outperforms all other materials with resonances in the blue part of visible spectrum. Other materials show comparable performance in different parts of the NIR and visible ranges. Notably, TCOs could enable LSPR applications in the NIR without the need for complex geometries such as core-shell structures. Titanium nitride and zirconium nitride have similar performance and exhibit LSPR modes in the visible part of the spectrum. In particular, titanium nitride could be of significant interest for biological applications because its LSPR mode lies in the biological transparency window. Also, TiN may be useful in plasmonic heating applications because it is refractory.

When the particle geometry is more complex than a sphere, the LSPR wavelength and its strength change. In general, small spherical particles exhibit a resonance strength that is proportional to $Re\{-\varepsilon_m\}/Im\{\varepsilon_m\}$, where ε_m is the permittivity of the plasmonic material. When the geometry changes from a sphere to cigar-like elongated structures, for example, the resonance quality varies as $[Re\{\varepsilon_m\}]^2/Im\{\varepsilon_m\}$. Alternative plasmonic materials have smaller magnitudes of real permittivity due to their smaller plasma frequencies, and hence they perform poorly when compared to noble metals for such complicated geometries.

High Temperature Plasmonics

Application such as heat-assisted magnetic recording (HAMR), photothermal cancer treatment, and thermophotovoltaics (TPV) require plasmonic devices operating at elevated temperatures. Refractory metals are more preferred than low-melting noble metals

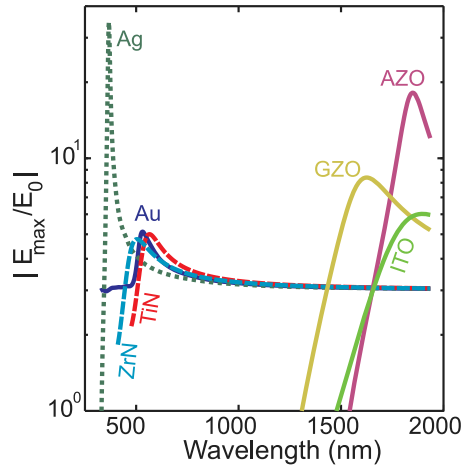

Fig. 6 Calculated maximum field enhancement on the surface of a spherical nanoparticle of the plasmonic material embedded in a dielectric host of refractive index 1.33. The calculations use quasistatic approximation. Reproduced from Naik, G.V., Shalaev, V.M., Boltasseva, A., 2013. Adv. Mater. 25 (24), 3264–3294. doi: 10.1002/adma.201205076.

in such applications. Though noble metals have high melting points ($\sim$1000 °C), their nanostructures with high ratio of surface area to volume deform at relatively low temperatures. One of the primary factors contributing to this deformation is high surface energy of these metals. On the other hand, refractory metals overcome such limitations and hence are promising for high temperature plasmonic applications.

Since high temperature plasmonic devices also rely upon LSPR phenomenon, the same trend as observed in **Fig. 6** holds. Given the problems of noble metals at elevated temperatures, alternative materials such as TiN and ZrN are very promising. Li *et al.* demonstrated the contrasting high-temperature stability between gold and TiN nanostructures (Li *et al.*, 2014). While 800 °C annealing destroyed the shape and optical properties of gold nanostructures, similar TiN nanostructures survived the annealing cycle and preserved their optical properties intact.

Recently, Guler *et al.* demonstrated LSPR in lithographically fabricated TiN nanodisks and showed better photothermal response than in similar gold nanostructures (Guler *et al.*, 2013). As an additional advantage, TiN nanostructures showed resonances in the biological transparency window of 800–1200 nm wavelength. Given the bio-compatibility of TiN, it is a promising material for photothermal cancer therapy. Ishii *et al.* demonstrated better solar steam generation in water suspended TiN nanoparticles than in similar suspensions of carbon-black or gold nanoparticles (Ishii *et al.*, 2016).

Recently, high temperature optical constants of gold and TiN are measured at high temperatures, and it is clear that TiN is stable at temperatures greater than 1200 °C and undergoes smaller degradation in its optical losses than gold (Briggs *et al.*, 2017). While gold doubles its Drude damping rate at 600 °C, TiN doubles at 1200 °C. Clearly, titanium nitride is a promising plasmonic material for high temperature applications and other refractory metal nitrides are yet to be explored.

Metamaterials and Transformation Optics

Metamaterials may be classified into two categories: LSPR-based and non-LSPR metamaterials. LSPR-based metamaterials such as negative index metamaterials rely upon LSPR phenomenon and hence the best performing plasmonic materials are noble metals followed by metal nitrides, further followed by TCOs (Tassin *et al.*, 2012). Non-LSPR metamaterials rely upon effective medium behavior of composite materials and examples are cloaks, hyperbolic metamaterials, and epsilon-near-zero (ENZ) materials. As a general note, most of the metamaterial implementations of transformation optics fall into the category of non-LSPR metamaterials. Such metamaterials benefit a lot from alternative plasmonic materials because of their smaller $\mathrm{Im}\{\varepsilon_m\}$.

Hoffman *et al.* showed negative refraction at an operating wavelength of 8 μm in a so-called hyperbolic metamaterial consisting of planar, alternating layers (a superlattice) of heavily doped ($1\text{–}4\times10^{18}$ cm^{-3}) $In_{0.53}Ga_{0.47}As$ and undoped $Al_{0.48}In_{0.52}As$ deposited by MBE (Hoffman *et al.*, 2007). This metamaterial device was shown to have a performance figure-of-merit of about 20, the highest reported so far. Yet another demonstration is the epsilon-near-zero (ENZ) properties of the InAsSb material when doped heavily. Adams *et al.* show that heavy doping ($1\text{–}2\times10^{19}$ cm^{-3}) of InAsSb causes the real permittivity to cross zero and turn metal-like in the MIR range (Adams *et al.*, 2011). At the zero cross-over of real permittivity (occurring at about 8 μm wavelength), the material behaves as an ENZ material and exhibits special properties such as photon funneling.

In the near-IR, Kim *et al.* showed ENZ property of TCO substrates pin the resonance of gold nanorods always to ENZ wavelength irrespective of their geometry (Kim *et al.*, 2016). Naik *et al.* demonstrated highest performing hyperbolic metamaterial in the near-infrared using Al-doped ZnO and undoped ZnO alternating layers (Naik *et al.*, 2012a). The figure-of-merit of this hyperbolic metamaterial was nearly 10 while similar noble-metal based structures have 3 orders of magnitude smaller

figures-of-merit. The implication of this demonstration is that TCO-based metamaterials are best suited for applications such as hyperlens – a sub-diffraction imaging metamaterial device.

In the visible spectral range, TiN based metamaterials have been demonstrated. In an example of a metamaterial for quantum optical applications, Naik et al. developed an epitaxial metal/dielectric superlattice based on TiN and AlN, and showed that the structure exhibits hyperbolic dispersion (Naik *et al.*, 2014). Layer thickness down to 5 nm – not possible with noble metals – was possible in TiN based metamaterial only because the surface energy of TiN is nearly 3 orders of magnitude smaller than that of noble metals. Such a metamaterial showed a huge enhancement in local photonic density of states, and a fluorophore probe near such metamaterial experienced about 80x enhancement in its radiative decay rate. Such huge enhancement in photon emission rate is useful in many light emitting devices including a single photon source for quantum optical applications. Shalaginov *et al.* extended this work to demonstrate fast single photon emission from NV centers of nano-diamonds (Shalaginov *et al.*, 2013).

Overall, alternative plasmonic materials are very promising for non-LSPR metamaterials owing to their smaller magnitude of real permittivity and small imaginary permittivity.

Summary and Outlook

With the rapid development of nanophotonics, it is clear that there will not be a single plasmonic material that is suitable for all applications at all frequencies. Rather, a variety of material combinations must be fine-tuned and optimized for individual situations or applications. While several approaches and materials have been discussed, many more have been and are being proposed. The problem of finding better plasmonic materials remains open-ended. An improved plasmonic material holds the key in transforming nanophotonics from a laboratory science to a ubiquitous technology for the coming generations.

References

Adams, D.C., Inampudi, S., Ribaudo, T., *et al.*, 2011. Phys. Rev. Lett. 107, 1090.
Blaber, M.G., Arnold, M.D., Ford, M.J., 2010. J. Phys. Condens. Matter 22, 143201.
Bouillard, J.S.G., Dickson, W., O'Connor, D.P., Wurtz, G.A., Zayats, A., 2012. Nano Lett. 12, 1561.
Briggs, J.A., Naik, G.V., Zhao, Y., *et al.*, 2017. Appl. Phys. Lett. 1.(in press).
Cai, W., Shalaev, V., 2009. Optical Metamaterials: Fundamentals and Applications. Springer Verlag.
Chen, N.C., Lien, W.C., Liu, C.R., *et al.*, 2011. J. Appl. Phys. 109, 43104.
Chen, W., Thoreson, M.D., Ishii, S., Kildishev, A.V., Shalaev, V.M., 2010. Opt. Express 18, 5124.
Cleary, J., Peale, R., Shelton, D., *et al.*, 2008. in (Cambridge Univ Press).
Dicken, M.J., Aydin, K., Pryce, I.M., *et al.*, 2009. Opt. Express 17, 18330.
Dressel, M., Grüner, G., Bertsch, G.F., 2002. Am. J. Phys. 70, 1269.
Fedyanin, D.Y., Yakubovsky, D.I., Kirtaev, R.V., Volkov, V.S., 2016. Nano Lett. 16, 362.
Guler, U., Ndukaife, J.C., Naik, G.V., *et al.*, 2013. Cleo 2013 QTu1A 2.
Hibbins, A.P., Sambles, J.R., Lawrence, C.R., 1998. J. Mod. Opt. 45, 2051.
Hoffman, A.J., Alekseyev, L., Howard, S.S., *et al.*, 2007. Nat. Mater. 6, 946.
Ishii, S., Sugavaneshwar, R.P., Nagao, T., 2016. J. Phys. Chem. C 120, 2343.
Johnson, P.B., Christy, R.W., 1972. Phys. Rev. B 6, 4370.
Khurgin, J.B., Sun, G., 2010. Appl. Phys. Lett. 96, 181102.
Kim, J., Dutta, A., Naik, G.V., *et al.*, 2016. Optica 3, 4.
Kinsey, N., Ferrera, M., Naik, G.V., *et al.*, 2014. Opt. Express 22, 12238.
Knight, M.W., King, N.S., Liu, L., *et al.*, 2014. ACS Nano 8, 834.
Kreibig, U., Vollmer, M., 1995. Springer Ser. Mater. Sci. 535.
Liu, K., Ye, C.R., Khan, S., Sorger, V.J., 2015. Laser Photonics Rev. 9, 172.
Li, W., Guler, U., Kinsey, N., *et al.*, 2014. Adv. Mater. 26, 7959.
MacDonald, K.F., Sámson, Z.L., Stockman, M.I., Zheludev, N.I., 2008. Nat. Photonics 3, 55.
Maier, S.A., 2012. Nat. Phys. 8, 581.
Minami, T., 2005. Semicond. Sci. Technol. 20, S35.
Naik, G.V., Boltasseva, A., 2010. Phys. Status Solidi – Rapid Res. Lett. 4, 295.
Naik, G.V., Kim, J., Boltasseva, A., 2011. Opt. Mater. Express 1, 1090.
Naik, G.V., Liu, J., Kildishev, A.V., Shalaev, V.M., 2012a. Proc. Natl. Acad. Sci. USA 109, 8834.
Naik, G.V., Saha, B., Liu, J., *et al.*, 2014. Proc. Natl. Acad. Sci. USA 111, 7546.
Naik, G.V., Schroeder, J.L., Ni, X., *et al.*, 2012b. Opt. Mater. Express 2, 478.
Novoselov, K.S., Fal, V.I., Colombo, L., *et al.*, 2012. Nature 490, 192.
Pashley, D.W., 1959. Philos. Mag. 4, 316.
Rao, C.N.R., 1998. Philos. Trans. R. Soc. London. Ser. A Math. Phys. Eng. Sci. 356, 23.
Shalaginov, M.Y., Ishii, S., Liu, J., *et al.*, 2013. Appl. Phys. Lett. 102, 2011.
Soref, R., Peale, R.E., Buchwald, W., 2008. Opt. Express 16, 6507.
Steinmüller-Nethl, D., Kovacs, R., Gornik, E., Rödhammer, P., 1994. Thin Solid Films 237, 277.
Tassin, P., Koschny, T., Kafesaki, M., Soukoulis, C.M., 2012. Nat. Photonics 6, 259.
Walukiewicz, W., 2001. Phys. B. Condens. Matter 302–303, 123.
Wang, C.-Y., Chen, H.-Y., Sun, L., *et al.*, 2015. Nat. Commun. 6, 7734.
Weir, B.E., Feldman, L.C., Monroe, D., *et al.*, 1994. Appl. Phys. Lett. 65, 737.
Yamada, T., Tokumitsu, E., Saito, K., *et al.*, 1989. J. Cryst. Growth 95, 145.
Zheng, B.Y., Wang, Y., Nordlander, P., Halas, N.J., 2014. Adv. Mater. 6318.

Raman Lasers

Marco Santagiustina, University of Padova, Padova, Italy

Introduction

Since their first appearance, Raman lasers have represented a simple and reliable approach to obtain laser sources at wavelengths not available through stimulated emission. Several experimental setups were developed (single pass, Fabry–Pérot cavities, ring cavities, intra- and extra cavity pumping etc.) each one with its own advantages, and disadvantages. As well, there were several materials that could be exploited, with a particular emphasis on solid-state lasers, which were mainly based on special crystals and optical fibers.

The main advantages of solid-state devices, with respect to those exploiting gases or liquids, are their reliability, long life, high power handling and high beam quality emission. In particular, crystals are known to be suitable materials for high power Raman lasers, due to their very good thermal properties, high damage threshold, high Raman gain. All these properties had finally made crystalline media Raman lasers a class of very efficient devices. Another class of solid-state devices was represented by fiber Raman lasers, realized in silica glass and showing two unique features, not possessed by Raman lasers based on crystalline media. The first was a broadband gain, due to the homogeneous broadening effect introduced by the amorphous glass. Such property enabled the achievement of wavelength tuning, a property lacking in crystal based devices. The second, unique feature was the seamless integration, through fusion splices, of the amplifying section (a fiber) with all the other devices required for the cavity, that are the mirrors (by means of fiber Bragg gratings) and the output ports (by means of fiber isolators, fiber directional couplers etc.). This design greatly simplified the assembly and the stability of the laser and enabled the realization of cascaded Raman lasers, in nested cavities, that maintained a high degree of compactness.

In the most recent years, the field of Raman lasers has taken advantage of the ongoing developments in photonics, extending such progresses into the specific content of Raman amplification and lasing. In particular, one should mention the extremely large advancements in the realization of photonic waveguides and circuits, which is leading to realize high quality on-chip devices that so far required bulk optics. As well, one should mention the new possibilities demonstrated by a new class of optical fibers based on photonic bandgap guidance. In the following sections we first briefly recall the fundamentals of Raman lasers and then we introduce the most relevant, recent advancements in Raman lasers.

Basic Principles

Raman lasers are based on the stimulated Raman scattering, an inelastic scattering process in which photons of a wave at one frequency, defined as the pump, are converted into photons at a smaller frequency, the Stokes wave. If the energy of the generated photon is larger than the pump photon energy, then the output wave is defined as anti-Stokes. In these processes part of the energy of the optical waves is exchanged with some internal form of energy of the Raman medium, typically molecular vibrations and rotations or lattice waves. So, Raman scattering can occur in several different media like gases, atomic vapors, liquids, crystals and glasses. In Fig. 1, the energy diagram of the stimulated generation of the Stokes and anti-Stokes waves is depicted.

The anti-Stokes interaction is less probable, because it requires the existence of a higher energy level $(v+1)$ much more populated than that of smaller energy (v); however, at thermal equilibrium, the populations are proportional to the Boltzmann distribution, $\exp(-E_v/K_BT)$ where E_v is the energy of level v, K_B is the Boltzmann constant and T the temperature. Therefore, the higher energy levels are less populated and generally Stokes waves (at wavelengths longer than the pump) are generated in a

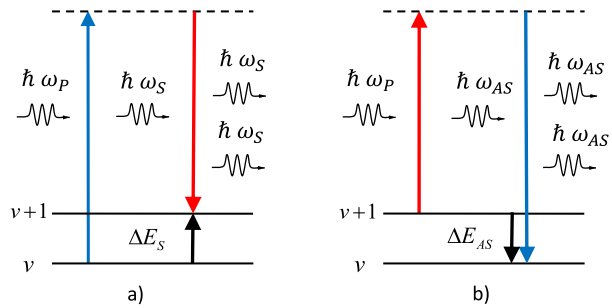

Fig. 1 Energy diagram of the stimulated Raman scattering; solid horizontal lines, v and $v+1$, indicate two medium energy levels, separated by an energy difference of $\Delta E_{S,AS}$; the dashed horizontal lines represent the unstable energy level reached by the medium during the process; $\hbar\omega_P$, $\hbar\omega_S$ and $\hbar\omega_{AS}$ are respectively the pump, Stokes and anti-Stokes photon energies. (a) refers to the Stokes interaction while (b) refers to the anti-Stokes interaction.

Raman laser. Though the pump to Stokes frequency shift is dictated by the medium energy levels, the lasing wavelength can be set to an arbitrary value, provided that a pump wave at a suitable wavelength is available. The efficiency of the processes depicted in Fig. 1 depends on several medium parameters and is summarized by the Raman gain coefficient at resonance g_R, measured in m W^{-1}. Energy levels are also characterized by a certain linewidth (with the typical Lorentzian distribution) that depends on which vibrational or rotational mode is excited, on the temperature, and on the coupling between the excited state and other states.

By imposing the conservation of the number of photons, that characterizes the interaction depicted in Fig. 1(a), a set of two nonlinear equations can be found for the evolution along the propagation direction (z) of the Stokes and pump wave intensities (I_S, I_P, measured in W m^{-2}):

$$\frac{dI_S}{dz} = (g_R I_P - \alpha_S) I_S, \quad \frac{dI_P}{dz} = \pm\left(\frac{\omega_P}{\omega_S} g_R I_S + \alpha_P\right) I_P$$

where α_S and α_P are the medium attenuation coefficients (measured in m^{-1}) respectively at the Stokes and pump frequencies, the $-(+)$ sign in the second equation applies to the case when the pump is co-propagating (counter-propagating) with respect to the Stokes wave. The equations are a valuable model to predict Raman laser properties and must be solved with boundary conditions that take into account the presence of a cavity. They are valid for the continuous and quasi-continuous wave cases; in other regimes (e.g. pulsed), propagation equations like the modified Schrödinger equations must be used instead.

Advancements in Crystalline Raman Lasers

The availability of new Raman crystalline materials, with large thermal damage thresholds and of suitable continuous wave pumping diode lasers have enabled to broaden the set of wavelengths covered by this type of laser sources and to achieve more easily the continuous wave emission.

The problem of thermal handling has been efficiently tackled by the growth of high quality Tungstate and Vanadate crystals (see Table 1). The heat generated during Raman scattering is an intrinsic problem of Raman lasers; in fact, it must be recalled that the energy difference between input and output photons is released to the medium in the form of vibrational or rotational energy so it amounts to

$$P_{\text{heat}} = P_S\left(\frac{\omega_P}{\omega_S} - 1\right)$$

where $P_S = \hbar\omega_S I_S$ is the Stokes wave power. In the near infrared region and with crystalline media the ratio between heat dissipated power and Stokes power (P_{heat}/P_S) can be as high as 10%. For this reason, the crystals used in Raman lasers should present a high thermal conductivity. Tungstates and Vanadates, as shown in Table 1, have a much higher thermal conductivity of Barium nitrate that was previously identified as the most performing crystalline media. In addition, Tungstates show a negative thermo-optic coefficient (refractive index decreases with temperature) and so thermal lensing, that can decrease the damage threshold, is avoided. The improved heating handling properties highly compensate the gain that is relatively smaller than that of Barium nitrate.

Crystalline Raman lasers pumped by an external pump can provide high output powers (up to about 2 W) with efficiencies that span from 5% to 17%. Pulse durations are in the range between 1 and 50 ns, with very few examples of sub-nanosecond operation, and repetition rates are in the range of 10–50 kHz.

A recent, major innovation is the possibility of achieving pump and Raman lasing in the same crystal (see Fig. 2), to finally realize the so called intra-cavity self-Raman lasers.

Table 1 Raman frequency shift, linewidth, gain coefficient, thermal conductivity and thermo-optic coefficient for various crystalline Raman media.

	Raman shift [cm^{-1}]	*Raman linewidth [cm^{-1}]*	*Raman gain coefficient @1064 nm [cm GW^{-1}]*	*Thermal conductivity @25°C [W m^{-1} K^{-1}]*	*Thermo-optic coefficient [10^{-6} K^{-1}]*
Barium nitrate $Ba(NO_3)_2$	1047	0.4	11	1.17	−20
Calcium tungstate $CaWO_4$	908	4.8	3	16	−7.1,[a] −10.2[a]
Potassium gadolinium tungstate $KGd(WO_4)_2$	768	6.4	4.4	2.5–3.4	−0.8,[a] −5.5[a]
Yttrium vanadate YVO_4	892	2.6	>4.5	5.2	3
Diamond C	1332	1.5	12	2000	9.6

[a]Value depending on the crystal orientation.

Source: All data except for diamond from Piper, J.A., Pask, H.M., 2007. Crystalline raman lasers. IEEE Journal Of Selected Topics In Quantum Electronics 13 (3), 692–704; data for diamond are from Lubeigt W., Bonner, G.M., Hastie,J.E., *et al.* Continuous-wave diamond Raman laser. Optics Letters 35 (17), 2994–2996.

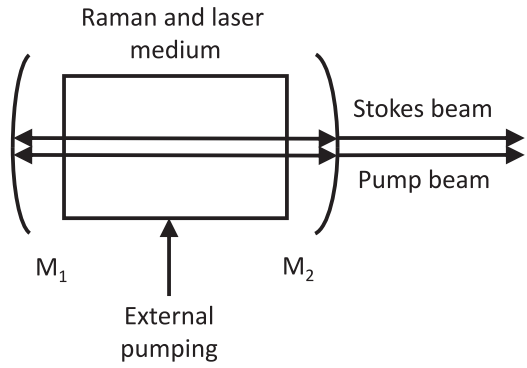

Fig. 2 Experimental setup of an intra-cavity, self-Raman oscillator. Mirrors M_1 and M_2 are highly reflective at the Stokes and at the pump wavelengths.

However, it must be remarked that, though simplifying the setup by eliminating the need for a separate Raman crystal, the thermal loading of the crystal in intra-cavity self-Raman lasers is extremely high, because of the coexistence of both amplification phenomena.

Another recent, remarkable advancement in the experimental setup of crystalline Raman lasers is the increased efficiency of intra-cavity frequency-doubling that has lead to high power output beams (1.8 W, with 9% efficiency) in the visible yellow-orange region.

The most significant recent development in crystalline Raman lasers is certainly represented by continuous lasing operation, also enabled by the exploitation of intra-cavity self-Raman lasing. Continuous wave laser output powers of about 700–800 mW have been achieved, with efficiencies spanning from 4% to 10%. Output power is still limited due to the high thermal loading that clearly affects the continuous wave regime.

From the thermal viewpoint, diamonds are the most promising crystalline media for realizing solid state Raman lasers because diamond thermal conductivity is the highest among solids (see **Table 1**). The quality of diamonds grown by means of chemical vapor deposition are by now comparable to the natural ones, with extremely broad transparency windows (0.225–2.6 μm, >6.2 μm). Therefore, diamond Raman lasers have reached, in a few years, a high degree of maturity. Though intra-cavity diamond Raman lasers may still suffer the thermal loading of lasing material, the external cavity setup enabled high conversion efficiencies (from 50% to 70%) and a very high output power: from about 10 W for continuous wave, to about 100 W in quasi-continuous wave and to a few kW in a pulsed regime (typically with kHz repetition rates and 20 ns pulses). The large diamond thermal conductivity also allows for the implementation of multiple pumping, with further enhancement of the output power (up to 6.7 kW was demonstrated). But the applications of diamond Raman laser can be extended also to the regime of very short pulses. In fact, though phonon dephasing time is in the order of 7 ps, short pulses (down to 25 fs) have been achieved exploiting the strong (soliton-like) compression effects due to the interplay between the cavity dispersion and the strong nonlinear chirp due to the Kerr nonlinearity. Finally, both monolithic and on-chip diamond Raman lasers can be realized and feature low thresholds (less than 100 mW) and Raman cascading up to the 4th order in the continuous wave regime.

On-Chip Raman Lasers

The evolution of the manufacturing technology of photonic waveguides and circuits has lead, in the recent past, to a process of miniaturization of several devices, including Raman lasers. The route for on-chip devices is particularly advanced with semiconductors, a field in which the high footprint technologies developed for electronics (epitaxy, e-beam lithography etc.) have been transferred to the photonic domain.

The most striking result in the field of Raman lasers is the demonstration and realization of silicon Raman lasers. Though silicon represents the by far most important material in electronics, its exploitation in photonics is hampered by the fact that the amplification through stimulated emission of radiation is hard to achieve, because it is an indirect bandgap semiconductor. To circumvent the lack of lasing action, Raman amplification has been proposed and successfully demonstrated.

Silicon presents a Raman shift of about 520 cm^{-1} with a linewidth of 4.6 cm^{-1} and a very large Raman gain coefficient (estimated about 76 cm GW^{-1} at 1.55 μm pumping). The physical process of Raman scattering in semiconductors is slightly different from that occurring in insulators, because it is mediated by electrons. In fact, as depicted in **Fig. 3**, the input pump photon first excites an electron–hole pair; the electron interacts with the medium and generates the lattice phonon.

The radiative recombination of electron–hole pairs finally results into the emission of the scattered Stokes photon. In silicon (and semiconductor media in general) other nonlinear phenomena can occur that can hamper Raman amplification. A particularly limiting effect is the two-photon absorption, which can cause pump depletion. Such phenomenon highly depends on the effective recombination lifetime of the free carriers, which can be reduced by a proper design of the waveguide. Several approaches can be used that include the interaction, recombination and diffusion with interface states at the boundary between the silicon and the buried oxide layer in SOI (silicon-on-insulator) waveguides or the realization of reverse-bias p-n junctions.

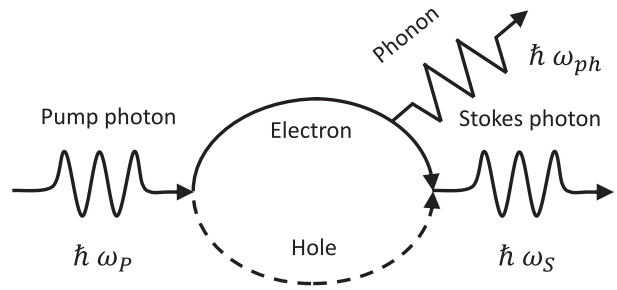

Fig. 3 Feynman diagram describing the Raman scattering process in semiconductors. A pump photon first generates an electron–hole pair. The electron loses energy by interaction with the lattice and a phonon is generated. The electron–hole recombination generates a Stokes (less energetic) photon. Based on Jalali, B., Raghunathan, V., Dimitropoulos, D., Boyraz, O. 2006. Raman-based silicon photonics. IEEE Journal of Selected Topics in Quantum Electronics 12 (3), 412–421. doi:10.1109/JSTQE.2006.872708.

Several types of silicon Raman lasers have been demonstrated by means of the previously mentioned approaches and with different type of cavities (that is ring, Fabry-Peròt, photonic crystals) getting continuous wave emission with a threshold as low as 20 mW, output power of about 50 mW and conversion efficiency of 12%. Slightly inferior performance (26 mW threshold, 10 mW output) can be achieved without any electrical bias. Remarkably, the Raman cascade lasing to the second order of continuous wave Stokes was also achieved with 5 mW of output power.

On chip Raman lasers has been manufactured also with other materials and techniques. Examples are the ultralow threshold (74 μW), high efficient (45%) monolithic silica micro-cavities and III-V Raman injection lasers that exploited triply resonant Raman scattering to enhance the low gain coefficient of such materials that is rather low (0.008 cm GW^{-1}) in comparison to other crystalline insulators or to silicon.

Advancements in Fiber Raman Lasers

Fiber Raman lasers based on step index single mode silica fibers proved to be one of the most reliable and effective types of Raman sources. Silica glass fibers present negligible losses over a rather large window ($<$0.5 dB/km between 1.2 and 1.7 μm). Moreover, the seamless integration of all the devices needed for the lasing make this type of source one of the most appealing in several practical applications. In fact, active lasing materials (in particular rare earths like neodymium, erbium, ytterbium, thulium, holmium etc.) can be easily embedded in the fiber glass matrix to realize an intra-cavity self-Raman pumping scheme, very similarly to what mentioned above for crystalline media. Highly reflective, narrow band mirrors can be permanently inscribed in the same fibers in the form of Bragg gratings that offer reliable and stable performance and, in principle, a certain degree of tuning, by means of the fiber temperature and fiber strain control. Other fundamental devices like power splitter, circulators and isolators are as well realized in optical fibers. Finally, all parts can be assembled by means of fusion splices, that present low losses (including low return losses) and that enable high power handling.

From the thermal handling viewpoint optical fibers are also very convenient. In silica fibers, the small frequency shift (about 490 cm^{-1}) reduces the total amount of heat power to 6.8% of the Stokes power. Moreover, the long length and the very small radius of the optical fibers by far compensate the value of the thermal conductivity (about 1.4 K m^{-1} W^{-1}) that is smaller than many crystalline media. The use of double-clad fibers enables cladding-pumped setups in which high total pump power can be launched and propagates in the internal cladding, while the Stokes signal is generated in the core. The outcome of all the previously mentioned combined properties and design characteristics is that extremely high power loading can be realized in fiber Raman lasers.

Continuous wave Raman fiber lasers have, in fact reached, the record kW regime (about 1.5 kW) at 1120 nm; in that case intra-cavity pumping was realized by doping a 26 m long fiber with ytterbium ions, made active by means of external semiconductor laser diodes emitting at 915 and 976 nm. Fiber Bragg gratings provided the cavity for both Stokes and pump waves and the overall optical efficiency was also very good (75%). Similar examples of this technology have been demonstrated with emission at other wavelengths, enabled by means of different co-doping, including some examples in the 2 μm range (thulium and holmium doped intra-cavity fiber Raman lasers). One of the main issues to achieve high spectral purity is that of suppressing high-order Stokes lasing. This is achieved by keeping the fiber short enough (typically 20–30 m) or by inserting intra-cavity filters that in fibers can be realized by designing fibers with specific cut-off wavelengths, so keeping, once again, the valuable seamless integration of the whole device.

Conversely, when cascaded Raman generation is a target this can be also achieved with very high power output and very large overall efficiency. A fifth order converter (1117–1480 nm) reported about 300 W output power at 1480 nm with a total conversion efficiency of about 64%.

Fiber Raman lasers, as well as the previously mentioned Raman laser types, also leveraged onto the most recent progresses in material science and technology to improve their performance and broadens the spectrum of emitted frequencies and therefore

applications. Though silica glass is still the most preferred medium for handling high power beams, other glasses have been tested. Fluoride glasses (Raman shift 572 cm^{-1}) enabled lasing beyond 2 μm by pumping at 1981 nm through an intra-cavity thulium doped fiber with an output power in the order of a few Watts. The chalcogenide glasses seem more promising for extending Raman lasing beyond 3 μm. In fact, losses of chalcogenide materials rapidly decay at wavelengths larger than 3 μm (for example As_2S_3 fiber has >1 dB/m at 3 μm but 0.1 dB/m at 3.345 μm). The Raman gain coefficient is much larger than that of silica (about 780 times) and that of fluoride glasses, so partly compensating the increase in linear losses. Thermal and mechanical properties of these glasses are, however, poorer with respect to silica glass and therefore it can be expected that their usage will be limited to achieve emission in spectral regions where silica fiber losses would be too high.

Another fundamental improvement in the field of optical fibers, namely the realization of photonic crystal fibers has also opened new perspectives for Raman fiber lasers. In photonic crystal fibers the guiding effect is no longer based on the total internal reflection, but on the photonic bandgap principle. The cladding is therefore substituted by a periodic structure, inhibiting light transmissions (bandgap) at the wavelengths that are to be guided so that they can only propagate in the photonic crystal defect, which is the fiber core. Among the various unusual properties of photonic crystal fibers, the one that is particularly attractive for Raman lasing is the property of guiding light in low refractive index materials, including vacuum or gases. Gases were among the first Raman media to be used in realizing single pass and cavity Raman lasers, because they presented rather large Raman gain coefficients (from tens o few hundred times that of silica glass), large transparency spectra and, moreover, Raman frequency shifts are the largest among Raman media. However, the vessels used to contain the gas were bulky and the nonlinearity was severely limited by diffraction. In fact, no light confinement was actually possible and therefore the Raman interaction can take place only in the region where the pump beam was focused, whose effective volume is diffraction limited. Hollow core photonic crystal fibers fully eliminated such problem and high intensity beams (fiber mode effective areas are in the order of tens of μm^2) can be maintained over much longer distances (from ten to a few hundred meters) before fiber loss becomes too large, so enabling Raman generation with a very high conversion efficiency. Experimental demonstrations were essentially made with hydrogen, that was the most widely studied in bulk vessels. A conversion efficiency of about 50% (line shifted by about 600 cm^{-1}) with 2.25 W of input power was achieved in a 30 m long hydrogen filled hollow core photonic crystal fiber. Conversion efficiency seems essentially limited by fiber loss, as demonstrated by the fact that 99.99% of output power was contained in the Stokes wave. Fully exploiting seamless fiber integration, experiments including fiber Bragg gratings mirrors (cavity) reduced the threshold to sub-Watt levels (600 mW). The very high efficiency derives also from the suppression of higher order Stokes waves, due to the large loss of the fiber at longer wavelengths.

Conversely, the availability of hollow core fiber (the so called Kagomé photonic crystal fibers) in which fiber loss coefficient is small (e.g. <2 dB m^{-1}) over a large band (a few hundreds of nm at wavelengths in the visible and infrared spectral ranges) enabled the realization of a Raman cascade and even the coherent excitation of multiple rotational or vibrational Raman lines. In the latter case, a completely novel feature for Raman lasers, which is the emission of a multi-octave frequency comb can be achieved. In fact, leveraging on the coherence of the Raman lines, all excited by the same pump wave, Raman combs have a potential to deliver sub-femtosecond pulses, as the spectrum can span over almost 1000 THz with a variable line spacing (from a few tens to more than one hundred THz).

Applications

Raman lasers continue to be applied in a large variety of fields. The availability of new wavelengths has broadened the range of applications in remote sensing for safety, like in atmospheric science but also in new areas like food quality evaluation. Biomedical applications are, by far, the area in which the recent progresses have had the largest impact. The availability of unconventional wavelengths (like the yellow-orange emission of crystalline lasers), the compactness and high-power delivery (like for fiber Raman lasers pumped by diode lasers) have made these devices a fundamental tool in diagnostics and therapies, in several medical areas like dermatology, ophthalmology, surgery and dentistry. Raman lasers remain fundamental tools for spectroscopic applications, now entailing also coherent combs and sub-femtosecond pulses, which also opened a completely new line of application in metrology. In optical communications, fiber Raman lasers are the most exploited solution for pumping Raman amplification along the fiber link, because of the reliability, efficiency and seamless integration with the rest of the network.

See also: Raman Spectroscopy

Further Reading

Abdolvand, A., Walser, A.M., Ziemienczuk, M., Nguyen, T., Russell, P. St. J., 2012. Generation of a phase-locked Raman frequency comb in gas-filled hollow-core photonic crystal fiber. Optics Letters 37, 4362–4364. doi:10.1364/OL.37.004362.

Bernier, M., Fortin, V., El-Amraoui, M., Messaddeq, Y., Vallée, R., 2014. 3.77 μm fiber laser based on cascaded Raman gain in a chalcogenide glass fiber. Opt. Lett. 39, 2052–2055. doi:10.1364/OL.39.002052.

Bernier, M., Fortin, V., Caron, N., *et al*., 2013. Mid-infrared chalcogenide glass Raman fiber laser. Optics Letters 38, 127–129. doi:10.1364/OL.38.000127.

Boyraz, O., Jalali, B., 2004. Demonstration of a silicon Raman laser. Optics Express 12, 5269–5273. doi:10.1364/OPEX.12.005269.

Cerný, P., Jelínková, H., Zverev, P.G., Basiev, T.T., 2004. Solid state lasers with Raman frequency conversion. Progress in Quantum Electronics 28, 113–143. doi:10.1016/j.pquantelec.2003.09.003.

Chen, Y.F., 2004. Efficient 1521-nm Nd: $GdVO_4$ Raman laser. Optics Letters 29, 2632–2634. doi:10.1364/OL.29.002632.

Chen, Y.F., 2004. Compact efficient all-solid-state eye-safe laser with self-frequency Raman conversion in a Nd: YVO_4 crystal. Optics Letters 29, 2172–2174. doi:10.1364/OL.29.002172.

Couny, F., Benabid, F., Light, P.S., 2007. Subwatt threshold cw Raman fiber-gas laser based on H_2-Filled hollow-core photonic crystal fiber. Physical Review Letters 99, 143903. doi:10.1103/PhysRevLett.99.143903.

Feng, Y., Taylor, L.R., Bonaccini Calia, D., 2009. 150 W highly-efficient Raman fiber laser. Optics Express 17, 23678–23683. doi:10.1364/OE.17.023678.

Feve, J.-P.M., Shortoff, K.E., Bohn, M.J., Brasseur, J.K., 2011. High average power diamond Raman laser. Optics Express 19, 913–922. doi:10.1364/OE.19.000913.

Georgiev, D., Gapontsev, V.P., Dronov, A.G., *et al.*, 2005. Watts-level frequency doubling of a narrow line linearly polarized Raman fiber laser to 589 nm. Optics Express 13, 6772–6776. doi:10.1364/OPEX.13.006772.

Grabitchikov, A.S., Linsinetskii, V.A., Orlovich, V.A., *et al.*, 2004. Multimode pumped continuous-wave solid-sate Raman laser. Optics Letters 29, 2524–2526. doi:10.1364/OL.29.002524.

Granados, E., Spence, D.J., Mildren, R.P., 2011. Deep ultraviolet diamond Raman laser. Optics Express 19, 10857–10863. doi:10.1364/OE.19.010857.

Hausmann, B.J.M., Bulu, I., Venkataraman, V., Deotare, P., Lončar, M., 2014. Diamond nonlinear photonics. Nature Photonics 8, 369–374. doi:10.1038/nphoton.2014.72.

Jackson, S.D., Anzueto-Sánchez, G., 2006. Chalcogenide glass Raman fiber laser. Applied Physics Letters 88, 221106. doi:10.1063/1.2208369.

Jalali, B., Raghunathan, V., Dimitropoulos, D., Boyraz, O., 2006. Raman-based silicon photonics. IEEE Journal of Selected Topics in Quantum Electronics 12 (3), 412–421. doi:10.1109/JSTQE.2006.872708.

Jiang, H., Zhang, L., Feng, Y., 2015. Silica-based fiber Raman laser at >2.4 μm. Optics Letters 40, 3249–3252. doi:10.1364/OL.40.003249.

Kippenberg, T.J., Spillane, S.M., Armani, D.K., Vahala, K.J., 2004. Ultralow-threshold microcavity Raman laser on a microelectronic chip. Opt. Lett. 29, 1224–1226. doi:10.1364/OL.29.001224.

Kitzler, O., McKay, A., Mildren, R.P., 2012. Continuous-wave wavelength conversion for high-power applications using an external cavity diamond Raman laser. Optics Letters 37, 2790–2792. doi:10.1364/OL.37.002790.

Kivistö, S., Hakulinen, T., Guina, M., Okhotnikov, O.G., 2007. Tunable Raman soliton source using mode-locked Tm-Ho fiber laser. IEEE Photonics Technology Letters 19, 934–936. doi:10.1109/LPT.2007.898877.

Latawiec, P., Venkataraman, V., Burek, M.J., *et al.*, 2015. On-chip diamond Raman laser. Optica 2, 924–928. doi:10.1364/OPTICA.2.000924.

Lee, A.J., Spence, D.J., Piper, J.A., Pask, H.M., 2010. A wavelength-versatile, continuous-wave, self-Raman solid-state laser operating in the visible. Optics Express 18, 20013–20018. doi:10.1364/OE.18.020013.

Li, B.-B., Xiao, Y.-F., Yan, M.-Y., Clements, W.R., Gong, Q., 2013. Low-threshold Raman laser from an on-chip, high-Q, polymer-coated microcavity. Optics Letters 38, 1802–1804. doi:10.1364/OL.38.001802.

Lin, J., Spence, D.J., 2016. 25.5 fs dissipative soliton diamond Raman laser. Optics Letters 41, 1861–1864. doi:10.1364/OL.41.001861.

Lubeigt, W., Bonner, G.M., Hastie, J.E., *et al.*, 2010. Continuous-wave diamond Raman laser. Optics Letters 35, 2994–2996. doi:10.1364/OL.35.002994.

McKay, A., Spence, D.J., Coutts, D.W., Mildren, R.P., 2017. Diamond-based concept for combining beams at very high average powers. Laser & Photonics Reviews 11, 1600130. doi:10.1002/lpor.201600130.

Mildren, R.P., Convery, M., Pask, H.M., Piper, J.A., Mckay, T., 2004. Efficient, all-solid-state, Raman laser in the yellow, orange and red. Optics Express 12, 785–790. doi:10.1364/OPEX.12.000785.

Mildren, R.P., Pask, H.M., Ogilvy, H., Piper, J.A., 2005. Discretely tunable, all-solid-state laser in the green, yellow, and red. Optics Letters 30, 1500–1502. doi:10.1364/OL.30.001500.

Mildren, R.P., Sabella, A., 2009. Highly efficient diamond Raman laser. Optics Letters 34, 2811–2813. doi:10.1364/OL.34.002811.

Parrotta, D.C., Kemp, A.J., Dawson, M.D., Hastie, J.E., 2013. Multiwatt, continuous-wave, tunable diamond raman laser with intracavity frequency-doubling to the visible region. IEEE Journal on Selected Topics in Quantum Electronics 19, 6471175. doi:10.1109/JSTQE.2013.2249046.

Pask, H.M., Dekker, P., Mildren, R.P., Spence, D.J., Piper, J.A., 2008. Wavelength-versatile visible and UV sources based on crystalline Raman lasers. Progress in Quantum Electronics 32, 121–158. doi:10.1016/j.pquantelec.2008.09.001.

Piper, J.A., Pask, H.M., 2007. Crystalline Raman Lasers. IEEE Journal of Selected Topics in Quantum Electronics 13 (3), 692–704. doi:10.1109/JSTQE.2007.897175.

Qin, G., Liao, M., Suzuki, T., Mori, A., Ohishi, Y., 2008. Widely tunable ring-cavity tellurite fiber Raman laser. Optics Letters 33, 2014–2016. doi:10.1364/OL.33.002014.

Reilly, S., Savitski, V.G., Liu, H., *et al.*, 2015. Monolithic diamond Raman laser. Opt. Lett. 40, 930–933. doi:10.1364/OL.40.000930.

Rong, H., Jones, R., Liu, A., *et al.*, 2005. A continuous-wave Raman silicon laser. Nature 433, 725–728. doi:10.1038/nature03346.

Rong, H., Kuo, Y.-H., Xu, S., *et al.*, 2006. Monolithic integrated Raman silicon laser. Optics Express 14, 6705–6712. doi:10.1364/OE.14.006705.

Rong, H., Xu, S., Cohen, O., *et al.*, 2008. A cascaded silicon Raman laser. Nature Photonics 2, 170–174. doi:10.1038/nphoton.2008.4.

Rong, H., Xu, S., Kuo, Y.-H., *et al.*, 2007. Low-threshold continuous-wave Raman silicon laser. Nature Photonics 1, 232–237. doi:10.1038/nphoton.2007.29.

Russell, P. St. J., Hölzer, P., Chang, W., Abdolvand, A., Travers, J.C., 2014. Hollow-core photonic crystal fibres for gas-based nonlinear optics. Nature Photonics 8, 278–286. doi:10.1038/nphoton.2013.312.

Sabella, A., Piper, J.A., Mildren, R.P., 2010. 1240 nm diamond Raman laser operating near the quantum limit. Optics Letters 35, 3874–3876. doi:10.1364/OL.35.003874.

Sabella, A., Piper, J.A., Mildren, R.P., 2014. Diamond Raman laser with continuously tunable output from 3.38 to 3.80 μm. Optics Letters 39, 4037–4040. doi:10.1364/OL.39.004037.

Supradeepa, V.R., Feng, Y., Nicholson, J.W., 2017. Raman fiber lasers. Journal of Optics 19 (2), 023001. doi:10.1088/2040–8986/19/2/023001.

Takahashi, Y., Inui, Y., Chihara, M., *et al.*, 2013. A micrometre-scale Raman silicon laser with a microwatt threshold. Nature 498, 470–474. doi:10.1038/nature12237.

Taylor, L.R., Feng, Y., Calia, D.B., 2010. 50W CW visible laser source at 589 nm obtained via frequency doubling of three coherently combined narrow-band Raman fibre amplifiers. Optics Express 18, 8540–8555. doi:10.1364/OE.18.008540.

Troccoli, M., Belyanin, A., Capasso, F., *et al.*, 2005. Raman injection laser. Nature 433, 845–848. doi:10.1038/nature03330.

GaN Lasers

Harumasa Yoshida, Mie University, Tsu, Mie, Japan

Introduction

There are three main types of wide bandgap semiconductors, which are group III-nitrides as GaN, II-chalcogenides as ZnSe and II-oxides as ZnO. Diamond and silicon carbide are also important wide bandgap semiconductors. Among those materials, the III-nitrides are direct bandgap semiconductor and one of the most promising candidates for the semiconductor light sources such as laser diodes and light emitting diodes (LEDs). The III-nitride-based semiconductors can provide a wide range of bandgap energies from 0.7 to 6.2 eV, compounding with InN, GaN, and AlN. In principle, it is possible to tune the emission spectral range of the GaN system down to infrared or toward deep ultraviolet region. At present, InGaN- and AlGaN-based LEDs cover the spectral range from red ($>$600 nm) to deep ultraviolet (210 nm) wavelengths. In contrast, the laser diodes require a more complex structure, thicker layers and higher crystalline quality than the LEDs to satisfy all the requirements of suitable optical and electrical confinements as well as high emission efficiency for lasing. As a result of the tight material requirements, the laser diodes are confined to a much narrower spectral range from green (530 nm) to near ultraviolet (320 nm) wavelengths.

The InGaN-based laser diodes are already in markets. In contrast, the AlGaN-based laser diodes set into the markets through a stage of research and development. Based on the 405-nm InGaN laser diodes, Blu-ray disk systems have been in widespread use for the high-density optical data storage application, instead of the conventional DVD storage system based on the 650-nm InGaP laser diodes. The blue InGaN laser diodes would be useful for metallic material processing and high brightness illuminations with scintillators because of the high power characteristic. For laser display application, the green InGaN laser diodes become alternative light source to bulky second-harmonic-generation YAG solid-state lasers. The ultraviolet AlGaN laser diodes could offer an alternative to the costly and bulky lasers such as excimer, HeCd and nitrogen gas lasers, and as third and fourth harmonic-generation YAG lasers. These conventional ultraviolet gas and solid-state lasers are widely used in the scientific and industrial fields. The merits and advantages of flexible wavelength, compactness, low-power consumption, and rapid startup of the laser diodes offer much convenience of the laser applied equipment, and provide a chance to create novel applications. For example, the ultraviolet laser diodes can provide excellent performance for the long range sensing applications such as light detection and ranging (LiDAR), and 3-dimensional mapping because the maximum permissible exposure for human eyes is much higher in 300-nm-band than that in 900-nm-band of the conventional GaAs-based laser diodes under the condition of pulsed mode operation.

In this section, after the brief overview on group III-nitrides, typical fabrication procedures and characteristics of both the AlGaN laser diodes for ultraviolet region and the InGaN laser diodes for visible region are described. Impacts on device characteristics caused by unique nature of the nitrides are also mentioned.

Group III-Nitrides

Group III-nitrides are compounds composed of group III-elements of boron (B), aluminum (Al), gallium (Ga) and indium (In), and group V-element of nitrogen (N). AlN, GaN, and InN are the binary nitrides suited to optical and electronic devices. Ternary and quaternary alloys such as AlGaN, InGaN, and AlInGaN are also synthesized for the complicated device fabrication. The GaN and the related nitrides crystallize in wurtzite or zincblende structure (Morkoç, 2008; Takahashi *et al.*, 2007; Nakamura *et al.*, 1997). The wurtzite structure is much stable at room temperature and pressure, whereas the zincblende structure is quasi-stable due to the difference of the degree of ionicity in the ionic bonding components. The crystalline structure of wurtzite is hexagonal, and the [0001] axis perpendicular to the hexagon structured plane is called as *c*-axis, as schematically illustrated in **Fig. 1**. The $\langle\bar{1}100\rangle$ axes perpendicular to the lateral plane of hexagonal column and the $\langle11\bar{2}0\rangle$ axes parallel to that plane are called as *m*-axes and *a*-axes, respectively. The III-nitrides exhibit remarkable polarization effects due to the asymmetry in atomic configuration along the [0001] direction. The strain-induced piezoelectric and spontaneous polarization fields have substantial impacts on the device characteristics. In contrast, the *m*-direction and *a*-direction are free from polarization due to symmetric structures along the directions. The *c*-plane is described as polar plane, and the *m*-plane and *a*-plane are described as nonpolar planes after the polarization properties. Other polarization-controlled planes can be obtained by slicing the crystal at some angles. Those are called as semi-polar planes. The typical one is $\{20\bar{2}1\}$ planes, as shown in **Fig. 2**. **Table 1** lists typical crystal lattice parameters for AlN, GaN, and InN. The wurtzite structures of AlN, GaN, and InN exhibit direct bandgap properties, whereas zincblende ones show indirect bandgap behavior. The typical values of bandgaps are also listed on the **Table 1**. Bandgaps of 6.2 (AlN), 3.4 (GaN), and 0.7 eV (InN) correspond to wavelengths of 200, 360, and 1800 nm, respectively. By controlling the composition of ternary compound AlGaN and InGaN, the optical emission is expected to be tuned extensively from deep ultraviolet to infrared spectral regions in principle.

The differences of lattice constants between AlN and GaN, GaN and InN, and AlN and InN are 2.5, 11, and 14%, respectively. It could be hard to grow quality crystals of AlInN alloy due to the large differences in optimal growth temperatures as well as lattice

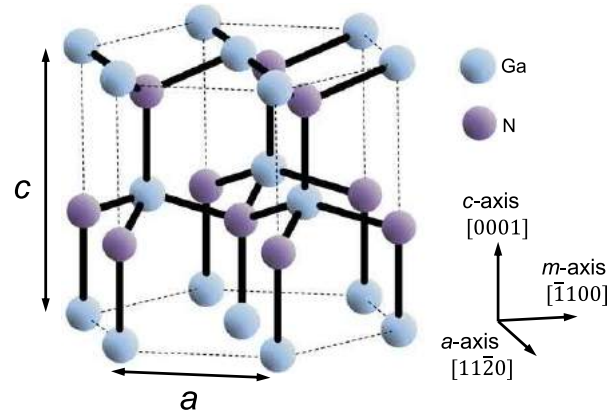

Fig. 1 Schematic illustration of wurtzite GaN crystalline structure. Reproduced from Takahashi, K., Yoshikawa, A., Sandhu, A. (Eds.), 2007. Wide Bandgap Semiconductors. Berlin Heidelberg: Springer-Verlag.

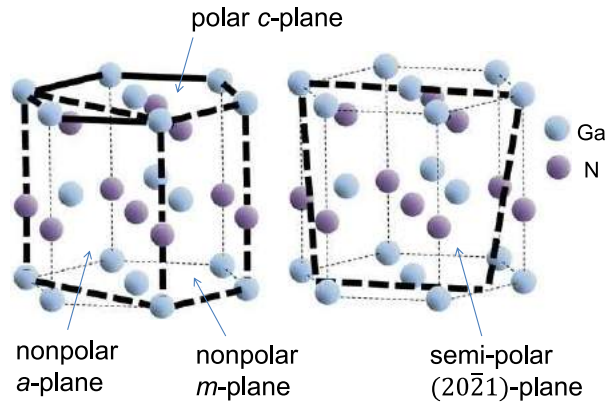

Fig. 2 Polar, nonpolar and semi-polar planes of GaN. Reproduced from Takahashi, K., Yoshikawa, A., Sandhu, A. (Eds.), 2007. Wide Bandgap Semiconductors. Berlin Heidelberg: Springer-Verlag.

Table 1 Lattice constants and bandgaps (wurtzite)

	a *(Å)*	c *(Å)*	*Bandgap (eV)*
AlN	3.111	4.980	6.2
GaN	3.189	5.186	3.4
InN	3.538	5.703	0.7

Source: Morkoç, H. (Ed.), 2008. Handbook of Nitride Semiconductors and Devices, vol. 1, Weinheim: WILEY-VCH Verlag GmbH & Co. KGaa and Takahashi, K., Yoshikawa, A., Sandhu, A. (Eds.), 2007. Wide Bandgap Semiconductors. Berlin Heidelberg: Springer-Verlag.

constants between AlN and InN. The calculated lattice parameters a and c of AlGaN and InGaN alloys can be approximately predicted by applying Vegard's law from several reports:

$$a_{\mathrm{AlGaN}} = 3.1986 - 0.0891x\ \text{Å} \text{ and } c_{\mathrm{AlGaN}} = 5.2262 - 0.2323x\ \text{Å} \tag{1}$$

$$a_{\mathrm{InGaN}} = 3.1986 + 0.3862y\ \text{Å} \text{ and } c_{\mathrm{InGaN}} = 5.2262 + 0.574y\ \text{Å} \tag{2}$$

where x and y are the AlN and InN molar fractions in the alloys, respectively. The bandgaps of these alloys can also be approximated, respectively, as:

$$E_{\mathrm{AlGaN}} = 6.1x + 3.4(1-x) - b_{\mathrm{AlGaN}} \cdot x(1-x)\mathrm{eV} \tag{3}$$

$$E_{\mathrm{InGaN}} = 0.7y + 3.4(1-y) - b_{\mathrm{InGaN}} \cdot y(1-y)\mathrm{eV} \tag{4}$$

Among a lot of reported parameters, the typical bowing parameters will be $b_{\mathrm{AlGaN}} = 1.0$ eV and $b_{\mathrm{InGaN}} = 1.43$ eV.

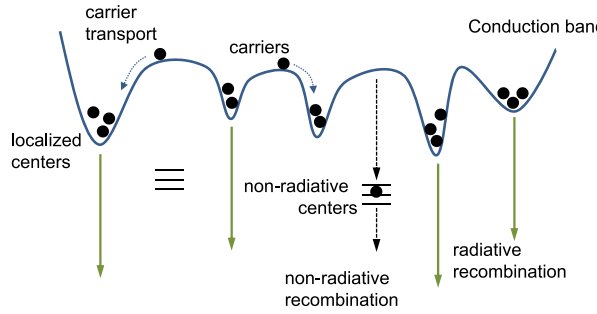

Fig. 3 Schematic representation on potential fluctuation. Reproduced from Takahashi, K., Yoshikawa, A., Sandhu, A. (Eds.), 2007. Wide Bandgap Semiconductors. Berlin Heidelberg: Springer-Verlag and Nakamura, S., Pearton, S., Fasol, G., 1997/2000. The Blue Laser Diode. Berlin Heidelberg: Springer-Verlag.

For the growth of nitride semiconductors, sapphire substrates are widely used because of the thermal and chemical stabilities. Although the large differences in lattice constants and thermal expansion coefficients between the sapphire ($a=4.763$ Å, $\Delta a/a=7.0\times10^{-6}\,K^{-1}$) and GaN ($a=3.189$ Å, $\Delta a/a=5.59\times10^{-6}\,K^{-1}$), a low-temperature AlN or GaN buffer layer enables the growth of quality GaN film on the sapphire substrates. Commercially available bulk GaN substrates are more suitable for the growth of high quality GaN film because of the lattice matching. However, it is difficult to grow high quality AlGaN film with high AlN molar fraction due to the lack of lattice matched substrates such as bulk AlN substrates.

There is a strong localization effect especially in the InGaN alloy. The localization effect fortunately leads to an excellent emission efficiency of the InGaN-based optical devices such as LEDs. The fluctuations of well width and alloy composition result in the potential fluctuation in active layers. The degree of potential fluctuation is caused by alloy disorder and correlates to the energy gap difference between the two binaries such as InN and GaN. InN and GaN provide an energy gap difference of 2.7 eV which is much larger than a typical value of a few hundred milli-electron volt in order for the II–VI and III–V semiconductor alloys. And the InGaN alloy is easy to generate a compositional modulation due to a large difference of 11% in the lattice constants between InN and GaN. The localized centers act as quantum dots and reduce the trap probability of excitons in nonradiative centers, as schematically illustrated in Fig. 3. However, the excess potential fluctuation reduces the gain for lasing of the laser diodes and relates to the difficulty in green InGaN laser diodes, as mentioned later.

AlGaN-Based Laser Diodes

The AlGaN laser diodes, in which the active layer was composed of the In-free AlGaN alloys, were reported for the first time in 2008 (Yoshida *et al.*, 2008). The AlGaN laser diodes were grown on the foreign substrates such as sapphire and silicon carbide because of lack of lattice matched substrates. The lattice mismatch results in the dislocations and the cracking of layers as well as the wafer bending. Epitaxial lateral overgrowth (ELO) method is one of solutions to relax the strain between the AlGaN layers and the lattice mismatched substrates, and to reduce the dislocations of the AlGaN layers grown on the substrates.

The quality AlGaN crystal films can be grown by the ELO method on the sapphire substrates. Fig. 4 shows a schematic illustration of typical layer structure. The material growth for crack-free and low-dislocation-density AlGaN layers is carried out by metaloganic vaporphase epitaxy (MOVPE). Trimethylgallium (TMG), trimethylaluminum (TMA), and ammonia (NH_3) are used as the source materials. A *c*-plane (0001) sapphire is used as the substrate. After the deposition of a buffer layer with a thickness of several tens of nanometers at a relatively low growth temperature of 500°C on the sapphire substrate, a 2.5-μm-thick GaN layer is grown at 1050°C. Following the deposition of SiO_2 striped masks along the *m*-axis on the GaN layer, inclined-facet GaN seed crystals can be selectively grown at 900°C through the masks by facet-controlled epitaxy. Both the width and the spacing of stripe masks are 2–3 μm. Then, a subsequent AlGaN layer is laterally overgrown at 1150°C on the inclined facets of the GaN seeds. A coalescence of each overgrown AlGaN layer from opposite facets of the GaN seeds is achieved. There is no crack over the entire area of the 2 in. diameter wafer. On the contrarily, a number of cracks are easily generated in the wafer grown by the conventional growth sequence without the inclined-facet GaN seed crystals. The cathode-luminescence observation reveals the low dislocation density of the AlGaN film with an average dark spot density in the order of $10^8\,cm^{-2}$. The dislocation density of AlGaN film grown by the ELO method is almost two orders of magnitude lower compared with that of the film directly grown on the sapphire substrate with the low-temperature buffer layer.

The AlGaN laser diodes are fabricated on the quality AlGaN film. The adequate AlN molar fractions of the AlGaN film depend on the emission wavelengths of laser diodes to satisfy the sufficient optical and electrical confinements in the active region. However, the layer configuration and the compositional allocation are constrained and compromised not to generate the clacks in film due to the lattice mismatched crystal growth despite the ELO method. As a typical example, the laser diodes with emission wavelengths of 320–340 nm are based on the AlGaN film with an AlN molar fraction of 0.3. The layer structure on the $Al_{0.3}Ga_{0.7}N$ film is composed of a 2.8-μm-thick n-$Al_{0.3}Ga_{0.7}N$ contacting layer, a 600-nm-thick n-$Al_{0.3}Ga_{0.7}N$ cladding layer, a 90-nm-thick AlGaN guiding layer, AlGaN multiple quantum wells (MQWs), a 120-nm-thick AlGaN guiding layer, a 20-nm-thick p-AlGaN

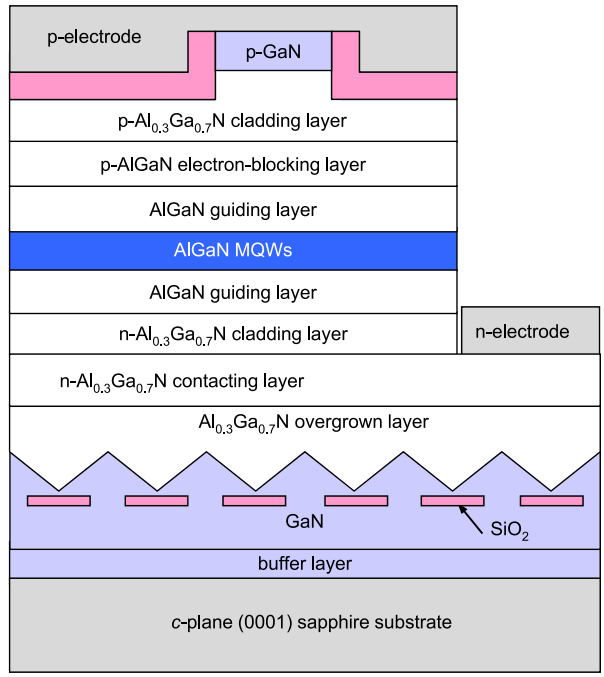

Fig. 4 Layer structure of AlGaN laser diodes. Reproduced from Yoshida, H., Yamashita, Y., Kuwabara, M., Kan, H., 2008. A 342-nm ultraviolet AlGaN multiple-quantum-well laser diode. Nature Photonics 2, 551–554. Available at: http://www.nature.com/nphoton/journal/v2/n9/abs/nphoton.2008.135.html.

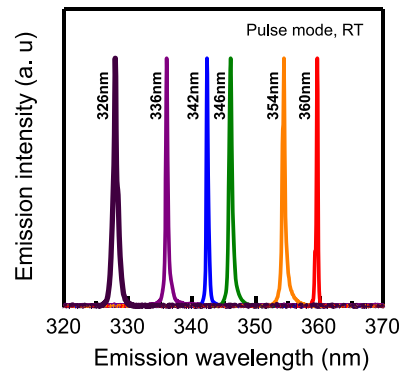

Fig. 5 A series of spectra for AlGaN laser diodes. Reproduced from Yoshida, H., Kuwabara, M., Yamashita, Y., *et al.*, 2009. AlGaN-based laser diodes for the short-wavelength ultraviolet region. New Journal of Physics 11, 125013. Available at: http://iopscience.iop.org/article/10.1088/1367-2630/11/12/125013/meta.

electron-blocking layer, a 500-nm-thick p-$Al_{0.3}Ga_{0.7}N$ cladding layer and a 25-nm-thick p-GaN contacting layer. The devices are processed as ridge waveguide lasers in accordance with the following procedure. After forming laser stripes and exposing the n-AlGaN contacting layer by conventional dry etching, an n-electrode is deposited on the exposed n-AlGaN contacting layer. A p-electrode is deposited on the p-GaN contacting layer. Facets of the laser cavities are also formed by dry etching. It is difficult to employ a highly reflective coating exactly for the cavity facets due to the obstruction of remaining unetched terraces in front of the facets.

Fig. 5 shows a series of lasing spectra for the AlGaN laser diodes with the AlN molar fraction being from 0 to 6% in the AlGaN wells. These spectra are measured under a pulsed-current mode. The emission wavelengths depend on both the MQW compositions and the well widths. An emission wavelength of 336 nm is nearly identical to 337 nm of the conventional nitrogen gas lasers. At present, the lasing wavelength has reached up to 326 nm corresponding to 325 nm of the HeCd gas lasers.

Fig. 6 shows a typical light output–current (L–I) characteristic of the AlGaN laser diodes with an emission wavelength of 342 nm. The device exhibits clear nonlinear behavior in the L–I characteristic. The threshold current of 390 mA corresponds to a threshold current density of 8.7 kA cm^{-2} for the 5-μm-ridge stripe and the 900-μm-long cavity. The peak output power from one

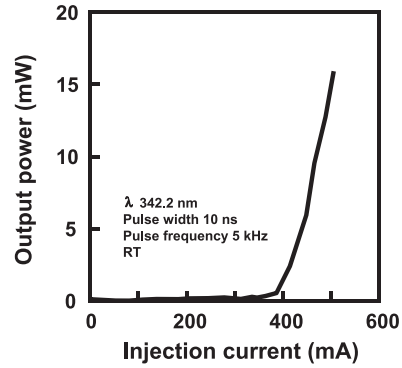

Fig. 6 Light output-current characteristic of AlGaN laser diode on sapphire substrate. Reproduced from Yoshida, H., Yamashita, Y., Kuwabara, M., Kan, H., 2008. A 342-nm ultraviolet AlGaN multiple-quantum-well laser diode. Nature Photonics 2, 551–554. Available at: http://www.nature.com/nphoton/journal/v2/n9/abs/nphoton.2008.135.html.

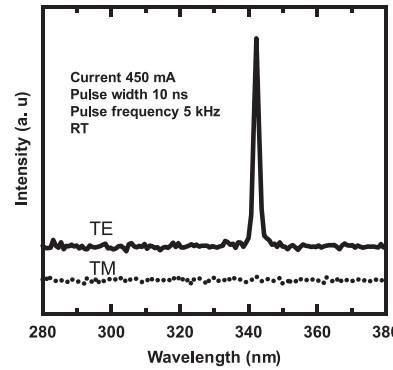

Fig. 7 Polarized emission spectra of AlGaN laser diode. Reproduced from Yoshida, H., Yamashita, Y., Kuwabara, M., Kan, H., 2008. A 342-nm ultraviolet AlGaN multiple-quantum-well laser diode. Nature Photonics 2, 551–554. Available at: http://www.nature.com/nphoton/journal/v2/n9/abs/nphoton.2008.135.html.

side of the facets approaches as high as 15 mW. The differential external quantum efficiency (DEQE) for the output from both facets is estimated to be 8.2%.

The typical transverse electric (TE) and transverse magnetic (TM) emission spectra from the AlGaN laser diode are shown in **Fig. 7**. Above the threshold, a very strong polarized emission can be observed. The emission is TE polarized due to the dominance of TE optical gain in the AlGaN MQWs with relatively low AlN molar fraction as well as the high reflectivity for the TE mode as a result of the cavity facets. Due to the large negative crystal field splitting in AlN compared with the positive value in GaN, an unusual valence band structure of AlN gives rise to unique optical polarization properties in the AlGaN alloys. The optical emission is dominantly TE polarized ($E \perp c$) for GaN, whereas the emission is predominantly TM polarized ($E \| c$) for AlN because a crystal-field split-off-hole state lies quite lower than heavy and light hole-states. The introduction of Al significantly changes the polarized optical gain property in the MQW. For the AlGaN MQWs with rather high AlN molar fraction, the TM emission is expected to be the dominant emission due to the polarization property of AlGaN alloys.

The peak output power of AlGaN laser diodes on the sapphire substrates are around several tens of milli-watts, and higher powers are expected to satisfy various application requirements. A broad-area striped structure is one of approaches for the higher output powers. The laser diodes fabricated on sapphire unavoidably accept lateral conduction due to the insulating property of sapphire. The laser diodes on sapphire substrates are unsuitable for the broad-area stripes because the lateral conduction leads to an un-uniform current injection into the wide emitter stripes. Moreover, the mirror facets of AlGaN laser diodes on the sapphire substrates are formed typically by dry etching due to the difficulty in cleaving the sapphire. Far-field patterns (FFPs) of the laser diodes are deteriorated by optical interference between the direct beam and the reflected beam from the lying down terrace in front of the etched mirror facet. The commercially available bulk GaN substrates are a potential candidate for both the vertical conducting structure and the cleaved mirror facets because of the excellent conductivity and cleavability of GaN.

The broad-area striped AlGaN laser diodes can be fabricated by the similar manner as mentioned above on the GaN substrates as an alternative to the sapphire substrates (Aoki *et al.*, 2015; Taketomi *et al.*, 2016). **Fig. 8** shows a typical *L–I* characteristic of the 50-μm-ridge stripe AlGaN laser diodes with an emission wavelength of 338 nm under the pulsed operation. The peak output power exhibits a clear nonlinear behavior around a threshold current of 7 A corresponding to a current density of 38.9 kA cm^{-2},

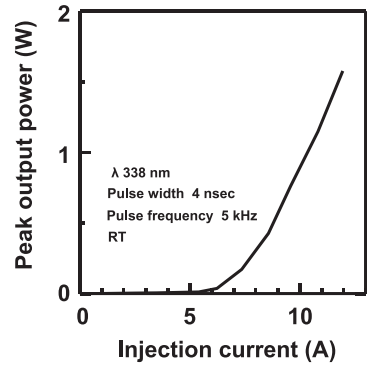

Fig. 8 Light output-current characteristic of broad-area AlGaN laser diode on GaN substrate. Reproduced from Taketomi, H., Aoki, Y., Takagi, Y., *et al.*, 2016. Over 1W record-peak-power operation of a 338 nm AlGaN multiple-quantum-well laser diode on a GaN substrate. Japanese Journal Applied Physics 55 (5S), 05FJ05. Available at: http://iopscience.iop.org/article/10.7567/JJAP.55.05FJ05/meta.

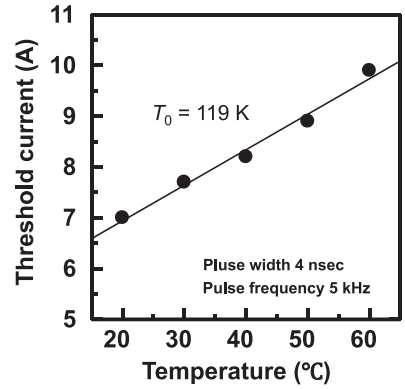

Fig. 9 Temperature dependence on threshold currents of AlGaN laser diode. Reproduced from Taketomi, H., Aoki, Y., Takagi, Y., *et al.*, 2016. Over 1W record-peak-power operation of a 338 nm AlGaN multiple-quantum-well laser diode on a GaN substrate. Japanese Journal Applied Physics 55 (5S), 05FJ05. Available at: http://iopscience.iop.org/article/10.7567/JJAP.55.05FJ05/meta.

and linearly increases above the threshold. A peak power of over 1 W is observed at operating currents of over 10 A. The slope efficiency and DEQE are estimated to be 0.31 W A^{-1} and 8.5%, respectively. The threshold current density and the slope efficiency are inferior to those of InGaN laser diodes (see next section of InGaN-based laser diodes). It could result from both the weak optical confinement due to the smaller difference in refractive index between the guiding layers and the cladding layers, and the electron overflow from the active region into the p-AlGaN cladding layer due to the low hole concentration as well as being free from the In effect.

Fig. 9 shows the temperature dependence on the threshold currents under the pulsed operation in a temperature range from 20 to 60°C. The threshold currents are increased from 7.0 to 9.9 A with increasing the temperature. The characteristic temperature T_0 of threshold current can be estimated to be 119K from the fitted line. This value is comparable to that of the AlGaN laser diodes on sapphire substrates. Fig. 10 shows the temperature dependence on emission wavelengths at a peak power of approximately 500 mW. The wavelengths move from 338.4 to 339.8 nm with a temperature rise of 40°C. The temperature coefficient of emission wavelengths is also estimated to be 0.033 nm K^{-1}. The small temperature coefficient is comparable to that of the infrared GaAs-based distributed feedback laser diodes with Bragg grating inside. The excellent spectral stability is favorable characteristic for various applications.

Fig. 11(a) shows the horizontal and vertical FFPs of the laser diode with GaN/AlGaN MQWs on the GaN substrate. The ridge width and cavity length are 1.5 and 300 μm, respectively. For the comparison, the FFPs of similar device on sapphire substrate are also shown in Fig. 11(b). The two dimensional images of FFPs are also shown on the graphs. Both the horizontal and vertical FFPs of the device on GaN substrate exhibit unimodal profiles, respectively. In contrast, the distorted FFPs can be seen for the device on sapphire substrate. The deteriorated FFP of the laser diode on sapphire substrate are considered to be caused by the interference between the direct beam from emitting area and the reflected beam from the residual terrace lying down in front of the etched mirror facet. The cleaved mirror facets provide the excellent FFP for flexible optical beam control owing to the cleavability of GaN substrates.

There are difficulties in the crystal growth of the AlGaN layer with higher AlN molar fraction due to the lack of lattice matched substrates, and in the conductivity control of Mg doped p-type AlGaN due to the low hole concentration originated from the high

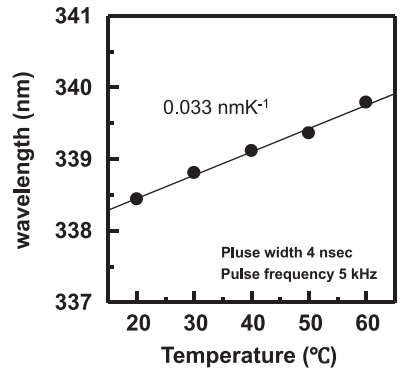

Fig. 10 Temperature dependence on emission wavelengths of AlGaN laser diode. Reproduced from Taketomi, H., Aoki, Y., Takagi, Y., *et al.*, 2016. Over 1W record-peak-power operation of a 338 nm AlGaN multiple-quantum-well laser diode on a GaN substrate. Japanese Journal Applied Physics 55 (5S), 05FJ05. Available at: http://iopscience.iop.org/article/10.7567/JJAP.55.05FJ05/meta.

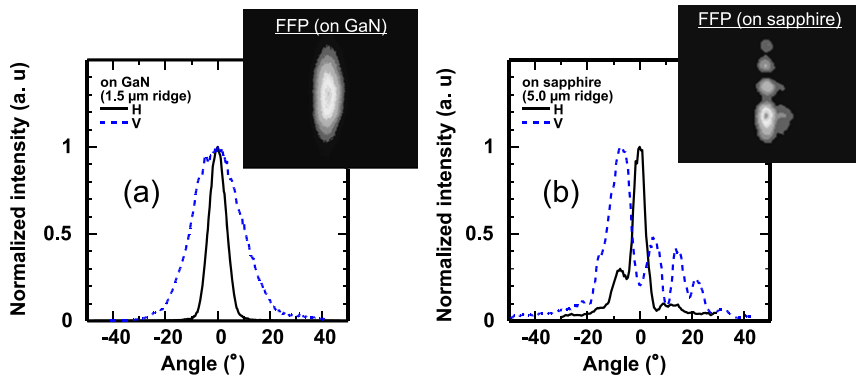

Fig. 11 FFPs of AlGaN laser diodes on (a) GaN and (b) sapphire substrates. Reproduced from Aoki, Y., Kuwabara, M., Yamashita, Y., *et al.*, 2015. A 350-nm-band GaN/AlGaN multiple-quantum-well laser diode on bulk GaN. Applied Physics Letters 107, 151103-1-151103-4. Available at: http://aip.scitation.org/doi/abs/10.1063/1.4933257.

binding energy of Mg. The Mg acceptor activation energy of 510 meV in AlN is much higher than 150 meV in GaN. These difficulties limit continuous wave (CW) operations and extending the wavelengths toward deep ultraviolet.

InGaN-Based Laser Diodes

The first InGaN laser diodes, in which the active layer was composed of the InGaN alloys, were demonstrated by Nakamura *et al.* in 1996 (Nakamura *et al.*, 1996, 1997/2000). The laser diodes were fabricated on the sapphire substrates because the bulk GaN substrates were unavailable commercially in the early stages. Nowadays, the bulk GaN substrates are commonly used for the InGaN laser diodes. According to the report on the first InGaN laser diodes, the laser diodes were grown by MOVPE on the *c*-plane (0001) sapphire substrate. **Fig. 12** shows a schematic illustration of the layer structure. The device is composed of a low temperature GaN buffer layer on the sapphire, a n-GaN contacting layer, a n-InGaN buffer layer, a n-AlGaN cladding layer, a n-GaN guiding layer, InGaN MQWs, a p-AlGaN layer, a p-GaN guiding layer, a p-AlGaN cladding layer, and a p-GaN contacting layer. The mirror cavity facets were formed by dry etching due to the difficulty in cleaving the epilayers grown on the sapphire. In order to reduce the threshold current, high reflection facet coating was employed. Evaporated Ni/Au and Ti/Al were used for the p-GaN and n-GaN contact layers, respectively. Reportedly, the threshold current was 1.7 A, corresponding to a current density of 4 kA cm^{-2}, under the pulsed operation of a pulse width of 2 μs and a period of 2 ms. A voltage of 34 V was applied at the threshold. The DEQE is estimated to be 13% from a pulsed output power of 215 mW per facet at an injection current of 2.3 A at room temperature. The laser exhibited a dominantly strong emission at a wavelength of 417 nm with a full width of half maximum (FWHM) of 1.6 nm. Because of the etched facets, it was difficult to obtain a unimodal FFP.

It was not easy to operate the InGaN laser diodes in CW mode. Under the direct current operation, the devices were easily failed and burned in a moment at room temperature due to heat generation caused by the high operating voltages. The room-temperature CW operation of InGaN laser diodes with a long lifetime of 40 min was achieved in 1997 by optimizing the growth conditions, the ohmic contacts and the doping profiles. Further long life CW laser diodes were demonstrated using the GaN

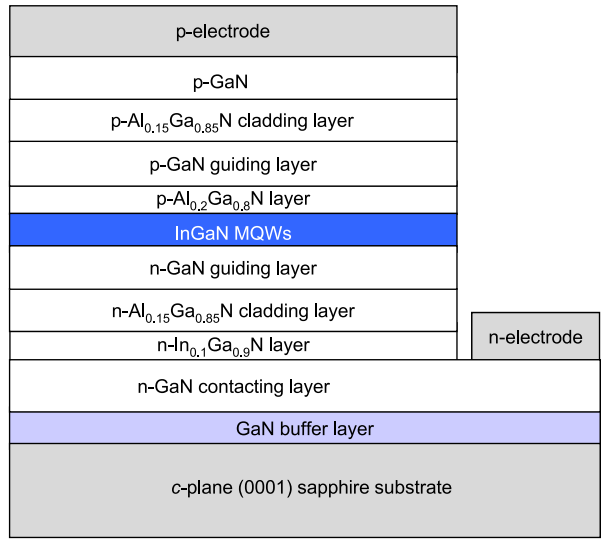

Fig. 12 Layer structure of InGaN laser diodes. Reproduced from Nakamura, S., Pearton, S., Fasol, G., 1997/2000. The Blue Laser Diode, Berlin Heidelberg: Springer-Verlag.

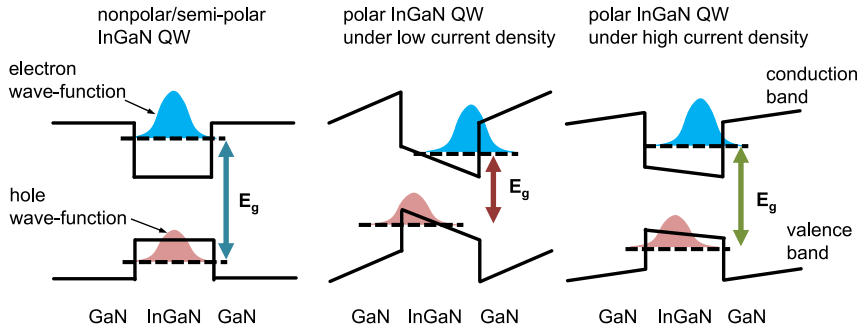

Fig. 13 Schematic band models of nonpolar and polar InGaN QWs. Reproduced from Kafafi, Z.H., Martin- Palma, R.J., Nogueira, A.F., *et al.*, 2015. The role of photonucs in energy. Journal of Photon Energy 5 (1), 050997. Available at: http://photonicsforenergy.spiedigitallibrary.org/article.aspx?articleid=2463108.

substrates as alternative to the sapphire substrates. Owing to an excellent GaN thermal conductivity of 1.3 W cm^{-1} K^{-1} higher than 0.5 W cm^{-1} K^{-1} of the sapphire, as well as the low dislocation density of the GaN substrates, the laser diodes were operated for 800 h. The 405-nm violet InGaN laser diodes led to Blu-ray disk technology in early 2000s.

For the violet InGaN laser diodes, the InN molar fractions of InGaN well layer are around 0.15. With increasing the InN molar fraction in the InGaN well layer of MQWs, the emission can shift theoretically to longer wavelengths. The blue laser diodes with an emission wavelength of 450 nm were reported in 2000. An estimated lifetime of 200 h for the blue laser diodes was seriously inferior to that of 2000 h for the violet laser diodes. This relates to poor crystalline quality of the InGaN MQW structure caused by dissociation of the InGaN active layer with a high InN molar fraction of 0.3. In recent years, a high output power of over 4 W has been marked for the blue laser diodes by the optimization of growth condition and the improvement in thermal management. The high output-power blue laser diodes are applied to laser displays, high brightness lamps, and material processing.

The red AlInGaP laser diodes are already available. Green laser diodes are the last key semiconductor-light-source for compact full-color-display applications consisting with three primary colors of red, blue, and green. Ideally, wurzite InGaN alloys grown in the nonpolar (*m* and *a* directions) and the semi-polar (for instance,$\{20\bar{2}1\}$) directions are the most potential candidate because of less blue-shift of the emission and higher internal quantum efficiency (IQE) in QWs compared to the polar *c*-plane QWs. The QWs on the nonpolar and semi-polar planes are free from and/or less quantum confined Stark effect (QCSE). The internal electric field is induced along the *c*-axis of the QWs by spontaneous and piezoelectric polarizations due to the asymmetry of the crystal as well as the strain resulting from the lattice mismatch between the layers. The internal electric field of QWs can be screened with increasing the injection current. The counter effect for the QCSE by the screening results in the increase of bandgap, or the blue-shift of emission wavelength. Moreover, the internal electric field spatially separates an electron wave-function and a hole wave-function in the QWs. The separate wave-functions reduce the radiative recombination probability or IQE. These phenomena are observed not only in the InGaN system but also in the AlGaN system. Fig. 13 schematically illustrates the phenomena. The

blue-shift behavior as well as the less IQE inhibits the shift to longer wavelengths of the InGaN laser diodes. The less blue-shift and the sufficient wave-function overlap are the important properties of nonpolar and semi-polar QWs.

There are other issues on In compositional fluctuation with increasing In contents in the InGaN active layer caused by the immiscibility of InN with GaN. Although the moderate In fluctuation is much effective for the improved emission of blue InGaN LEDs because of the localized carriers, as already shown in **Fig. 3**, the excessive fluctuation generate much defects in the QWs. The defects act as non-radiative centers and result in less IQE of the green InGaN laser diodes. The In fluctuation also leads to broad emission spectrum, and reduces the gain for lasing of the laser diodes.

Several groups demonstrated the CW operations of green laser diodes taking advantages of the nonpolar and semi-polar InGaN layers (Okamoto *et al.*, 2009; Takagi *et al.*, 2012; Miyoshi *et al.*, 2009). However, it is difficult to grow or to fabricate the large size nonpolar and semi-polar GaN substrates. The poor availability of non- or semi-polar substrates constrains the developments of green laser diodes. Contrarily, the polar *c*-GaN substrates with a diameter of 2 in. are commercially and widely available. The diameter reaches up to 6 in. at present. By optimizing the growth conditions of InGaN even with high InN molar fraction, the watt-class 537-nm green laser diodes have been realized on the *c*-GaN substrates. Based on the technological progress of the blue and green InGaN laser diodes as well as the red AlInGaP laser diodes, the standard wavelengths of three primary colors have been determined to be 467, 532, and 630 nm for the super high vision displays (ITU-R BT.2020-1).

References

Aoki, Y., Kuwabara, M., Yamashita, Y., *et al.*, 2015. A 350-nm-band GaN/AlGaN multiple-quantum-well laser diode on bulk GaN. Applied Physics Letters 107, 151103-1–151103-4.

Miyoshi, T., Masui, S., Okada, T., *et al.*, 2009. 510–515nm InGaN-based green laser diodes on c-plane GaN substrate. Applied Physics Express 2, 062201-1–062201-3.

Morkoç, H. (Ed.), 2008. Handbook of Nitride Semiconductors and Devices, vol. 1. Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA.

Nakamura, S., Pearton, S., Fasol, G., 1997/2000. The Blue Laser Diode. Berlin Heidelberg: Springer-Verlag.

Nakamura, S., Senoh, M., Nagahama, S., *et al.*, 1996. InGaN-Based Multi-Quantum-Well-Structure Laser Diodes. Japanese Journal Applied Physics 35, L74–L76.

Okamoto, K., Kashiwagi, J., Tanaka, T., Kubota, M., 2009. Nonpolar m-plane InGaN multiple quantum well laser diodes with a lasing wavelength of 499.8 nm. Applied Physics Letters 94, 071105-1–071105-3.

Takagi, S., Enya, Y., Kyono, T., *et al.*, 2012. High-power (over 100mW) green laser diodes on semipolar {20$\bar{2}$1} GaN substrates operating at wavelengths beyond 530 nm. Applied Physics Express 5, 082102-1—082102-3.

Takahashi, K., Yoshikawa, A., Sandhu, A. (Eds.), 2007. Wide Bandgap Semiconductors. Berlin Heidelberg: Springer-Verlag.

Taketomi, H., Aoki, Y., Takagi, Y., *et al.*, 2016. Over 1W record-peak-power operation of a 338nm AlGaN multiple-quantum-well laser diode on a GaN substrate. Japanese Journal Applied Physics 55, 05FJ05.

Yoshida, H., Yamashita, Y., Kuwabara, M., Kan, H., 2008. A 342-nm ultraviolet AlGaN multiple-quantum-well laser diode. Nature Photonics 2, 551–554.

Optically Pumped Semiconductor Lasers

Jerome V Moloney and Alexandre Laurain, University of Arizona, Tucson, AZ, United States

Introduction

Optically pumped semiconductor lasers (OPSLs), also referred to as vertical external cavity surface emitting lasers (VECSELs), or semiconductor disk laser (SDL) exhibit tremendous performance in terms of power, spatial and temporal coherence, noise characteristic, short pulse generation, and offer unique features such as tunability, power scalability, or intracavity frequency conversion (second harmonic generation (SHG) or difference frequency generation (DFG)). They combine the advantage of a semiconductor laser technology: compact and highly efficient gain medium that can be engineered to emit at a targeted wavelength, covering the visible up to the mid-infrared (IR) range, with the advantage of diode-pumped solid-state lasers: power scalability with circular low divergence beam and high-Q optical cavity. Unlike most semiconductor lasers, the laser light is amplified perpendicular to the surface of the semiconductor rather than the edge of the semiconductor chip. In this sense, they can be viewed as close cousins of the vertical cavity surface emitting laser (VCSEL), a widely employed commercial source that is generally electrically pumped and exhibits a typical low power of a few mill-watts in single mode operation. Fundamentally different, however, is the physical layout of the OPSL cavity depicted in **Fig. 1**.

Compared to a VCSEL, the optical cavity is extended by using an external mirror instead of a top distributed Bragg reflector (DBR), leaving an air gap between the OPSL chip and the external output coupler. The external optical cavity can be independently designed to provide a single transverse mode operation, by matching the cavity fundamental mode size with the pumped area. The external cavity also facilitate the insertion of various optical elements like an etalon and birefringent filter for high power single frequency operation, a saturable absorber for passive modelocking of ultrashort pulses, or a non-linear crystal that can exploit the very high intra-cavity circulating field to efficiently convert the fundamental wavelength to visible and UV frequencies via harmonic generation down to terahertz frequencies via DFG. Thanks to the very broad absorption spectrum of semiconductor media and extremely thin active region, OPSL chips can be excited with a spatially and spectrally incoherent pump beam, permitting the use of commercially available high power fiber-coupled diode laser modules. The chip itself consists of a high reflectivity distributed Bragg mirror (DBR) consisting of multiple repeats of two semiconductor materials, followed by an extremely short gain region (typically $\sim 2\ \mu m$) consisting of pump absorbing barriers and multiple quantum wells (QWs) or quantum dots (QDs) placed strategically according to the standing wave of the optical field. For high power operation, the chip is typically bonded to a chemical vapor deposition (CVD) diamond heat spreader which is placed in close contact with a water cooled copper heat sink. Thermal management is critically important as most of the pump induced heat is generated in an extremely small volume (disk of diameter 500 μm and thickness 2 μm).

The wavelength versatility and tunability of these compact sources result from the many possible semiconductor gain material combinations that can be grown to provide direct lasing in the UV (InGaN, AlGaN, GaN), visible GaInP/AlGaIn/GaAs, near-IR (InGaAs, AlGaAs, GaAs), and mid-IR (InGaSb, AlGaSb, GaSb). Due to the relatively recent development and immaturity of some material systems, especially in the visible and UV, it is sometimes more advantageous to generate these wavelengths via SHG using a more mature material system like InGaAs/GaAs. This usually offers even greater power at the expense of added cavity complexity. It is thus not surprising that many companies such as Coherent (Santa Clara) commercialize high power semiconductor lasers with a technology based on InGaAs QWs, and offer visible and ultraviolet wavelengths via efficient intra-cavity SHG in OPSLs.

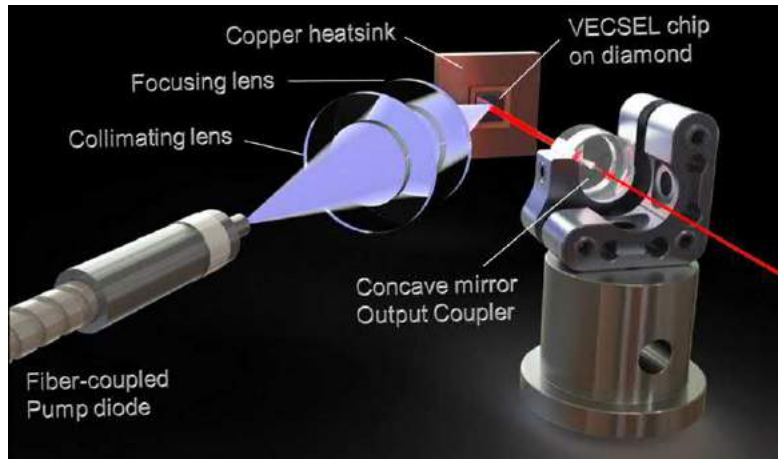

Fig. 1 Representation of an optically pumped semiconductor laser (OPSL) setup in a linear cavity configuration.

 doi:10.1016/B978-0-12-803581-8.09631-4

Another remarkable feature of OPSLs is their ability to generate ultrashort pulses at repetition rates unreachable with solid state or fiber laser technology due to Q-switching instabilities. The optical cavity can be arranged to integrate a semiconductor saturable absorber mirror (SESAM) to passively modelock the modes with repetition rates ranging from tens of megahertz up to 100 GHz. This is a major advantage for the development of compact femtosecond mode-locked source with output peak powers exceeding kilowatts and average output powers exceeding 5 W.

Background

The history of the development of OPSLs is relatively recent, dating back to the mid-1990s. The semiconductor active region of the OPSL chip typically consists of 8–12 QWs arranged to lie close to the anti-nodes of the signal standing-wave within the OPSL, as illustrated in Fig. 2. Due to the micro-cavity or micro-resonator formed by the DBR and the semiconductor/air interface, the field intensity within the structure can be significantly increased at the nominal wavelength. This resonance can be exploited to enhance the relatively low gain resulting from the fact that light is extracted to normal rather than along the QWs plane. A typical resonant periodic gain (RPG) design provides a gain of about 5 to 10%, depending on the number of QWs used and the pumping rate. The resulting device has marked advantages relative to other solid state lasers and semiconductor edge emitters in particular. Specifically, power scaling is simply achieved by increasing the diameter of the pump and mode waist on the chip. For single transverse mode operation, the fundamental mode is matched to the pump areas which serve as a soft aperture to discriminate the high order transverse modes. Indeed, since the transverse modes are not guided like in an edge emitter, they naturally diffract in the free space external cavity and the higher order modes spatially extend beyond than the fundamental mode and are absorbed more in the unpumped region of the OPSL chip. TEM_{00} operation with large diameter provide a very low divergence beam (<1 degree), often close to the diffraction limit ($M^2 < 1.1$) with a circular symmetry. These large beam diameters also strongly reduce the risk of a light induced surface damage or facet damage, the plague of diode edge emitters. In contrast to high power solid disk lasers, more than 80% of the pump power is absorbed in a single pass, simplifying the pumping scheme.

Most modern OPSLs are now grown via MBE or MOCVD as bottom emitters, namely, the RPG QW stack is grown first on a lattice matched substrate. This growth mode greatly facilitates heat extraction, as the poorly thermal conductive thick substrate (GaAs for InGaAs/AlGaAs/GaAs) can be entirely removed by chemical etching after bonding to a CVD diamond or other heat spreader. Bottom emitters require growth of an etch stop layer (typically AlGaAs or InGaP) just before a cap and the QW stack. The cap layer is needed to confine carriers within barrier regions so that they are efficiently captured in the QWs. The large refractive index of the semiconductor material relative to air ensures strong confinement of the signal field within the resulting semiconductor micro-cavity. In some instances, it is desirable to AR coat the surface of the chip to facilitate mode-locked pulse generation or internal signal conversion via SHG or DFG. The linear cavity depicted in Fig. 1 represents the simplest and optimal arrangement for raw power generation although a V-cavity arrangement provides a better control of the transverse modes on the OPSL chips and can include a SESAM for mode-locking.

Overview

OPSL Design Strategies

While other solid state disk laser sources are limited to wavelengths associated with specific transitions in dopant materials, the semiconductor gain medium supports wide tunability by simply changing the relative composition of the semiconductor

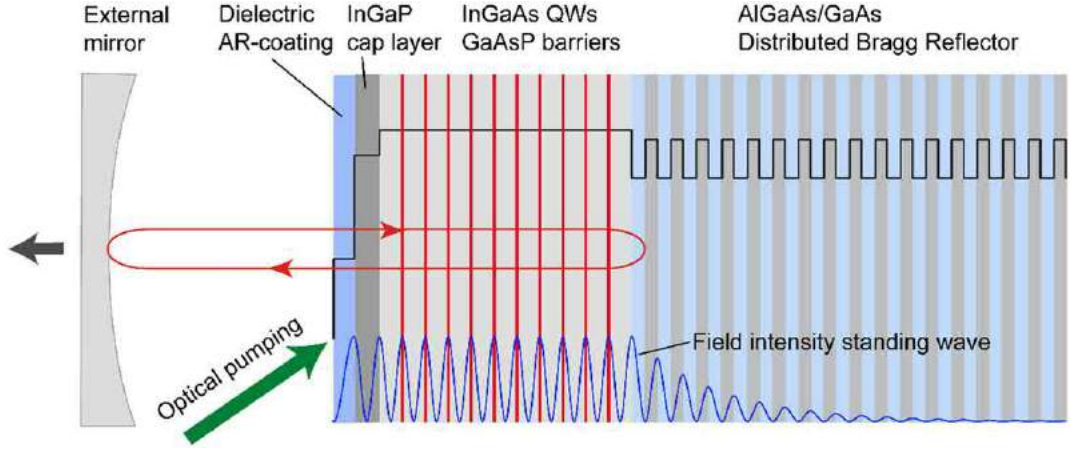

Fig. 2 Schematic representation of a typical GaAs-based vertical external cavity surface emitting laser (VECSEL) structure with a resonant periodic gain (quantum wells (QWs) placed on the field antinodes).

materials. For example, InGaAs QWs can be tuned to lase between 920 and 1200 nm by changing the composition of the Indium in the well. Increasing Indium increases strain and requires that strain compensating layers to be grown between individual QWs in an RPG stack, for example. The tuning range can be extended further toward 1.5 μm by slightly doping with atomic nitrogen. The nitrogen with concentration of a few percent is not incorporated into atom lattice sites but acts as an independent atom inclusion that shifts the valence upwards and reduces the bandgap in InGaAs. The 1.5 μm window can also be accessed via InGaAsP QWs grown on InP. DBRs are difficult to grow for this material system and the combination of InP/InGaAlAs has a low index contrast and consequently requires very thick structures. It is possible to wafer fuse the active region to a GaAs/AlAs Bragg mirror on a GaAs wafer, but the fabrication of such hybrid structures is difficult.

Performance of OPSLs for different targeted applications such as power scaling, high power single frequency operation or mode-locked short pulse generations is strongly dependent on many factors including semiconductor epitaxy design and thermal management. Typical diagnostics of growth accuracy prior to chip processing involve measuring surface and/or edge photo-luminescense (PL). The two PL diagnostics differ in that the surface PL, which can be done on a full semiconductor wafer, is convolved with the DBR reflectivity, whereas the edge PL gives a direct measure of the individual QW quality and growth accuracy.

However, the wavelength of the PL at room temperature is usually shorter than the wavelength of the gain peak in a running OPSL. In designing the OPSL chip for targeted performance, one must know accurately the temperature increase under varying pump power as the gain peak itself shifts strongly with temperature increase. Therefore the QWs need to be designed with PL at shorter wavelength anticipating the location of the peak gain at the final laser operating temperature (typically around 400K). One must also take into account the spectral shift of the micro-cavity resonance, as it will also shift with temperature but at a slower rate than the QWs gain. **Fig. 3** shows a typical evolution of the unsaturated gain spectrum and micro-cavity resonance under pumping and it's correlation with the output power characteristic. At low pump power the structure is close to room temperature and the QW gain peak is significantly detuned (− 30 nm) from the nominal wavelength of 1000 nm, whereas the resonance is only slightly detuned (− 7 nm). As the pump power increase and thus the temperature, the gain spectrum gets higher, broader and shifts at a rate of about 0.3 nm/K whereas the resonance shifts at a rate of about 0.07 nm/K. The gain peak and the resonance will be spectrally aligned when the internal temperature reaches ∼400K giving the maximum output power. If the pump power is further increased, the spectral shift and the gain amplitude decrease with temperature lead to a thermal rollover of the output power.

Obviously, the temperature increase for a given pump power depends on the thermal impedance of the device and is the main limiting factor of high power OPSLs. The thermal impedance depends critically on the bonding quality of the OPSL chip to the CVD diamond heat spreader, the thickness and thermal conductivity of the semiconductor membrane and the pump beam size. Increasing the pump beam size reduces the thermal impedance but also increases the lasing threshold. As long as the thermal impedance scales with the beam area, a linear power scaling is possible but, for very large pump beam (∼1 mm diameter), the lateral thermal heat flow becomes a limiting factor and the impedance no longer scales with the beam area and limits the achievable power output. This limitation together with the possible gain depletion from lateral lasing with large pump beam has set the limit of raw power from a single device to just above 100 W to date, which is remarkable coming from a semiconductor laser.

Current sophisticated software tools (SimuLase) utilize the microscopic physics of the semiconductor active structure to aid in targeting a specific wavelength and power performance. Such software tools reduce dramatically the usual required multi-parameter ad hoc phenomenological treatment needed in less sophisticated approaches. In addition to the optical and modal

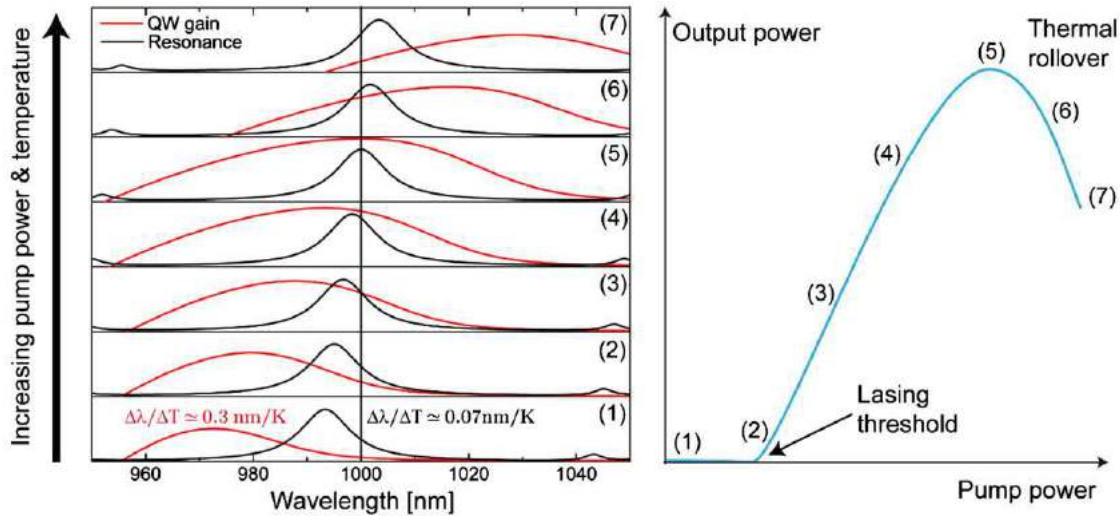

Fig. 3 Evolution of the intrinsic quantum wells (QWs) gain and micro-cavity resonance with increasing pump power for a nominal emission wavelength of 1000 nm at maximum power, and correlation with the output power characteristic.

gain, the OPSL performance also depends strongly on both radiative and nonradiative carrier losses (spontaneous emission and Auger recombination). Auger recombination tends to become increasingly important at high temperature and for longer wavelengths (i.e., lower bandgap).

OPSL Gain Structure

OPSL gain chips are typically designed as RPG structures. Here individual QWs are placed at the anti-nodes of the standing wave laser signal field that is confined within the semiconductor micro-cavity. This gain structure has a high reflectivity distributed Bragg Mirror (DBR) at one end such that the structure acts as a "gain mirror." The DBR is grown of similar materials to that of the active structure, for example, AlGaAs/AlAS alternating layers are employed in structures with InGaAs QWs. This simple optical component can be placed in different ways relative to other elements (mirrors, SESAMs, etc.) to assemble different types of cavities.

The design of the optical gain structure is crucial when optimizing performance for direct continuous wave (CW), intra-cavity SHG/DFG or ultrashort pulse generation. A global optimization scheme of the structure before wafer growth must take into account the "cold" and "hot" lasing properties of the device. When pumped either optically or electrically, the chip heats up and the optical properties can be strongly modified. The goal is to minimize the influence of heating in the active structure making an efficient means of heat extraction critical. The growth mode can strongly impact the thermal management scheme (Fig. 4).

For example, many of the original OPSL devices were grown as top emitters – the DBR stack is first grown on the substrate followed by the multiple QW stack and carrier confinement cap layers. The rather high thermal impedance of semiconductor materials limits this kind of devices to relatively low power (<10 W). For higher power they generally require that optically transparent heat spreaders (single crystal diamond or silicon carbide) are capillary or wafer bonded to the top of the structure. Ring-shaped copper heatsinks are then bonded to the heat spreader to allow for accessible space for the pump and laser light to reach the OPSL chip. The main disadvantage of this technique is that the heatspreader is inside the laser cavity, requiring very high optical quality and transparent materials. This added element might also induce parasitic reflection, additional losses and birefringence or require a specific antireflection coating, etc., which can alter the performance of the device. It is however, very useful and sometime the only alternative with substrate materials which cannot be easily removed from the structure like GaSb for mid-IR wavelength.

Bottom emitters are grown in reverse order (flip-chip design) – the cap layer is grown first followed by the QW stack and finally the DBR. In some instances it is necessary to first grow an etch stop layer on the substrate. The bottom emitter can be directly bonded with Ti/Au/In or Cr/Au/In to a lower optical quality CVD diamond heat spreader. The great advantage of the bottom emitter is that the entire substrate can be removed with chemical etchants leaving behind an extremely thin OPSL chip. Not surprisingly bottom emitters exhibit dramatically better performance due to very efficient cooling.

OPSL Loss Mechanisms

Individual optical components inserted in OPSL cavities and output coupled transmission are generic optical loss mechanisms that need to be balanced against the relatively low gain of an OPSL chip. Due to the high finesse of the cavity, scattering losses on the surface of the chip and to a lesser extent on the surface of the dielectric mirror(s) cannot always be neglected. Typical scattering losses of 0.25 to 0.75% are expected, which can become increasingly important with low modal gain structures (long wavelengths, anti-resonant design, etc.), and need to be accounted for the design of the OPSL chip and cavity.

For passive mode-locking operation, saturable and non-saturable losses from the absorber (SESAM, Graphene-SAM, etc.) contribute further to the overall cavity loss when trying to achieve lasing threshold and will impact the dynamics of the pulsed lasing regime.

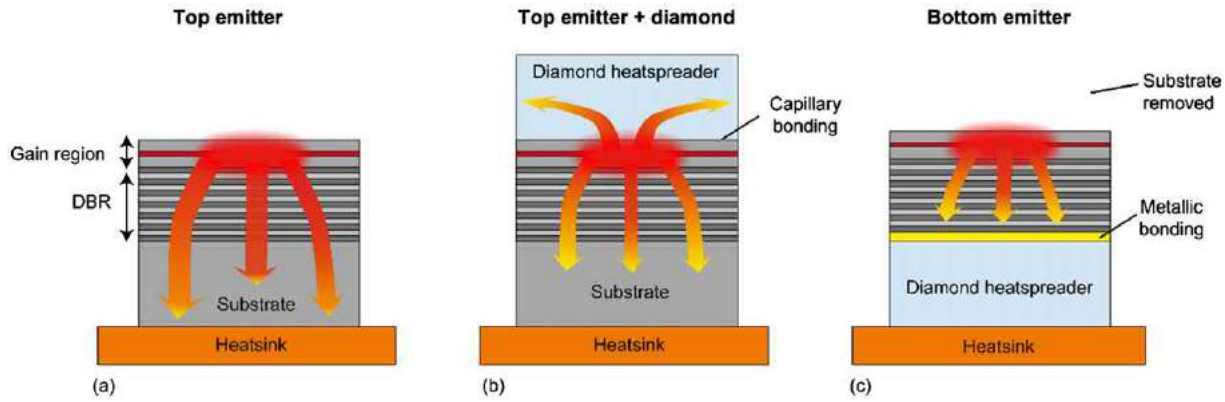

Fig. 4 Thermal management strategies for high power vertical external cavity surface emitting laser (VECSEL). (a) Top emitter: high thermal impedance from the substrate limits to low power regimes; (b) top emitter with intracavity diamond heatspreader: low thermal impedance but increased optical losses; and (c) bottom emitter bonded to a diamond: low thermal impedance and low losses.

On the other hand, intrinsic carrier losses in semiconductor active media directly affect the material gain and efficiency of the device and must be considered or even better microscopically calculated for a given material system, in order to fully optimize a structure. The carrier losses are usually dominated by radiative (spontaneous emission) and non-radiative (Auger losses) recombinations. Defect assisted recombinations which are assumed proportional to the carrier density are negligible for high quality grown structures. Auger losses, generally considered proportional to the carrier density cubed (often not the case!), become dominant at longer wavelengths but are also players at high operating temperatures at shorter wavelengths. The latter can suppress the gain to such an extent that the OPSL does not lase at all or, more typically, suddenly shuts off under strong pumping due to the added heat. Recently, novel Type-II W gain structures have been grown in an effort to suppress Auger losses. These designs displace the electron confining layer from the hole confining layers so as to minimize Auger losses. This is done at the expense of reducing the gain but has been successfully demonstrated for both electrical and optically pumped OPSLs.

OPSL Pumping Schemes

As mentioned previously, the best performing OPSLs are typically pumped by an external multimode fiber-coupled diode laser. The quality of the pump laser is not important nor is the pumping wavelength if the OPSL is designed for barrier pumping. In this scenario, the QWs are grown between pump absorbing barrier layers. For example, InGaAs QW can be surrounded by layers such as GaAs, AlGaAs or GaAsP to provide broad absorption features that are guaranteed to absorb over a broad spectral bandwidth. For GaAs-based materials, one can benefit from commercial pump modules emitting at 808 nm available at relatively low cost, which were initially developed for the pumping of solid state lasers. The disadvantage of barrier absorption is that the carrier confinement in the QWs imposes an energy difference between the pump and the lasing photons. This energy or wavelength difference between the shorter wavelength pump (808 nm) and the emitting wavelength (>1000 nm) can be quite large leading to a large Stoke shift or quantum defect (Fig. 5(a)). This energy difference is wasted in the form of heat when the excited carriers relax into the wells via phonon interactions with the crystal lattice. This is one of the major sources of pump induced heat in the structure and also limits the achievable optical-to-optical efficiency of the device. Another source of heat comes from the incomplete absorption of the pump in the barriers, typically between 10 and 20% of the incident pump power, which will be absorbed in the DBR or in the bonding materials such as Ti or Cr at the back end of the DBR if it is transparent at the pump wavelength.

A second option is to directly pump the QW in the RPG stack (Fig. 5(b)). This strongly reduces the quantum defect, but it is at the expense of the absorption. The reduced interaction length (<100 nm) between absorbing layers and the pump requires a recycling of the unabsorbed pump to increase the overall efficiency and to avoid unnecessary heating. This form of in-well pumping is analogous to a multi-pass pumping scheme employed with solid state thin disk lasers. While the QW absorption per pass is considerably stronger than the thin disk absorbing medium, it still require numerous passes for efficient absorption (>4) which is impractical to implement. The requirements on the pump laser wavelength and linewidth are also more stringent in this case. If the pump wavelength is too close to the gain peak, the absorption will decrease with the carrier density in the QW as the gain bandwidth gets broader and becomes more transparent at the pump wavelength. Nevertheless, in-well pump OPSL has been demonstrated with different material system with rather impressive power performance but much less than barrier pumping. It is also a good solution for OPSL emitting in the visible (red), where barrier pumping becomes impractical due to the lack of available pump sources.

Although not strictly characterized as OPSLs, it is important to mention that a structure can also be pumped electrically. Even though they are technologically more challenging, electrically pumped VECSELs (EP-VECSEL) have been demonstrated in the near

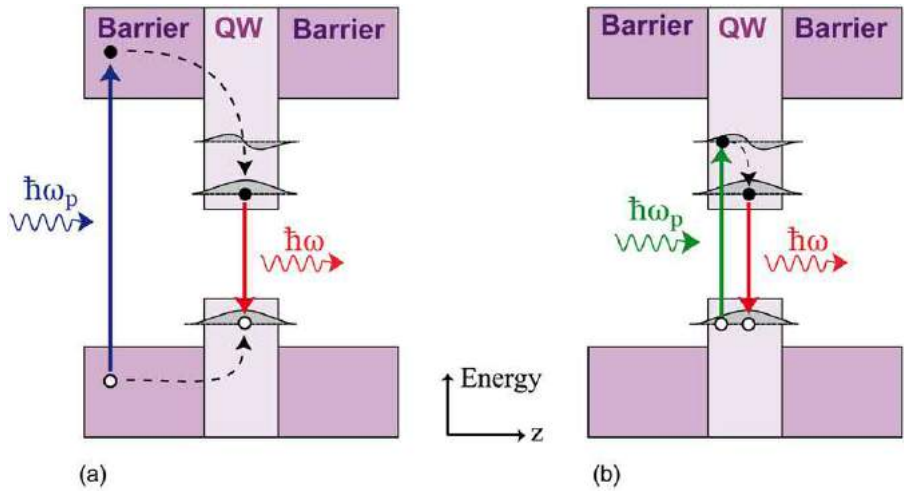

Fig. 5 Optical pumping scheme of OPSLs. (a) Barrier pumping: strong absorption, large quantum defect and (b) in-well pumping: weak absorption, low quantum defect.

IR and have exhibited impressive output powers in a TEM_{00} beam. This technology was used to generate both CW operation and mode-locked picosecond pulse trains. Just like the OPSL, thermal management plays a crucial role in the performance of these lasers – electrically pumped systems suffer additionally from Joule heating whereas optically pumped systems produce heat via the quantum defect and incomplete pump absorption in the active region. EP-VECSELs have a gain structure similar to that of a VCSEL, where the active area is sandwiched between a circular electrode on the bottom (mesa) and a ring electrode on the top to permit light extraction. In order to compensate the increased optical losses from free carrier absorption in the required doped layers, the structure usually contain a partial DBR reflector on the top of the active region to boost the micro cavity resonance and increase the modal gain. Unfortunately, this technique cannot be power scaled easily, as the contact geometry does not allow a uniform injection of carriers in the active region with a large area. The weak lateral carrier diffusion in the center of the active area is a major limitation for the output power in a single transverse mode. So far, the powers achievable with such devices appear to be limited to the order of 4 W. The compactness of this technology is, however, very attractive for numerous applications and have been successfully commercialized by NECSELs, primarily for visible displays and laser lighting markets.

OPSL Cavity Configurations

The simplest cavity configuration is the linear cavity depicted in **Fig. 1**. Here, the OPSL chip with backing DBR is mounted on a CVD diamond heat spreader and the latter soldered to a copper heat sink. An air gap on the order of a few millimeters to tens of centimeter exists between the active OPSL chip and an external output curved mirror with transmission varying between typically 0.5 and 10% for the higher gain chips. The external pump beam is imaged on the OPSL chip at a finite angle, typically between 20 and 45 degrees. This simple cavity setup is best for initially testing OPSL chip performance and for scaling of power beyond tens of Watts. This external free space cavity is a stable multimode optical resonator as long as the cavity length is shorter than the radius of curvature of the concave mirror ($L < R_c$). This cold cavity supports a very large number of transverse modes and the length of the cavity and curved mirror will dictate the fundamental and higher modes waists. By finely adjusting the fundamental mode waist to the pumped area, it is possible to spatially filter all the higher order mode with the help of the gain/absorption profile. However, to stabilize large TEM_{00} beams the cavity length required might be too long and the thermal lensing in the semiconductor can destabilize the output beam at high pump levels. In this case, it is preferable to use a more sophisticated cavity configuration, like a V-cavity where a large beam waist can be more easily obtained.

In a V-shaped cavity, the OPSL chip can be placed either at the vertex of the V (**Fig. 6(a)**) or as an end mirror (**Fig. 6(b)**), depending on the application targeted.

In a mode-locking setup, the SESAM absorber is typically placed at the end of the V-cavity and the other end mirror acts as an output coupler (**Fig. 6(a)**).

For SHG, the non-linear crystal might be placed between a flat and a curved mirror to ensure a double pass in the crystal before SHG light extraction through the curved mirror (**Fig. 6(b)**). There are many other variants of cavity geometries, a few are depicted in **Fig. 6**, including a Z-cavity, a ring cavity which offers further advantages over the V-cavity for modelocking. In some cases, it may be desirable to pump multiple OPSL chips for further power scaling within a single cavity, like in the W-shaped cavity of **Fig. 6(f)**. It has been demonstrated that power scales close to linearly with number of OPSL chips. There are much more possible geometry which is not shown here, that might serve best a particular application.

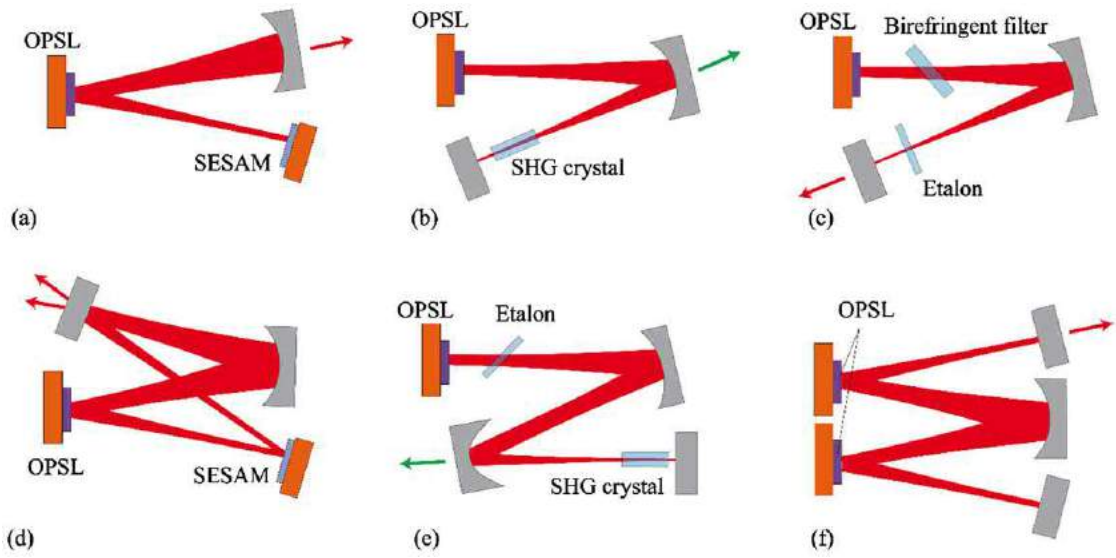

Fig. 6 Examples of VECSEL cavity geometries. (a) V-cavity for passive modelocking; (b) V-cavity for second harmonic generation; (c) V-cavity for single frequency operation; (d) folded ring cavity for colliding pulse modelocking; (e) Z-cavity for SHG; and (f) W-cavity for multi-chip power scaling.

High Power, Wavelength Tunable, High Brightness CW OPSLs

CW powers exceeding 100 W have already been demonstrated from a single OPSL chip. No attempt was made to enforce a single TEM_{00} mode at these power levels and thermal management will offer the greatest challenge. These deliver the final circular beam with $M^2 < 10$ which is much less than what was demonstrated by comparable power diode bars. TEM_{00} mode operation has been demonstrated up to powers of about 40 W and commercial OPSL intracavity SHG OPSL sources are available at powers up to 10 W. The broad gain bandwidth of the semiconductor QW allows for tuning over 30–40 nm albeit at lower multi-Watt powers.

The high brightness circular beam delivered by the OPSL chip offers many advantages over established high power broad area and diode bar emitters. Firstly, significantly improved brightness relative to these sources means that it can be imaged onto a smaller target spot or into a fiber – high power broad area edge emitters deliver a highly aberrated, highly divergent beam with an elliptical beam shape due to the rectangular shape and small vertical dimension of the emitting diode. Moreover, edge emitters are limited by the high power density at the emitted aperture leading to potential facet damage. The OPSL chip can maintain a much lower power density (power per unit area) as the emitter surface is much larger and power scaling is achieved by increasing the area of the pump and signal spot on the chip while maintaining a relatively constant power density.

High Power Narrow Linewidth CW OPSL Sources

Commercial single frequency edge emitters typically consist of a lower power single frequency DFB seed integrated with a flared planar amplifier (MOPA). These have been shown to perform up to 10 W but are extremely sensitive to any stray feedback or residual reflectivity at the end of the flare amplifier. The latter is present even with the best anti-reflection coatings. Moreover, they suffer from the same beam aberrations as their broad area high power counterparts.

Single mode OPSL sources with linewidths in the tens of kilohertz regime have been demonstrated with powers up to 23 W at a wavelength of around 1 μm. At this power level, single longitudinal mode operation needs to be enforced by inserting additional optical elements such as birefringent filters and etalon into a V-cavity. At lower power, it has been shown that single frequency operation is achievable with a very high spectral purity without any need for intra-cavity filter. This is accomplished thanks to the nearly ideal homogenous gain broadening of the OPSL combined to the high finesse cavity which strongly enforces the effect of the spectral gain curvature, and lead to high temporal coherence.

Fundamentally, the intensity and frequency noise of a single frequency laser is limited by the amount of spontaneous emission contained in the lasing mode. Because of the random nature of its phase, the spontaneous emission creates a fluctuating electrical field in time, which converts to a phase and frequency fluctuation of the output. In the high-Q cavity of a single frequency OPSL, the spontaneous emission is strongly filtered by the cold-cavity mode which gives rise to the emitted laser mode. For this reason, the fundamental limits in term of intensity and frequency noise are extremely low for an OPSL (sub-hertz linewidth). On the contrary, in most laser diodes the lasing mode propagates in an optical waveguide where the spontaneous emission is directly coupled and confined with the lasing mode and can be largely amplified along the cavity (amplified spontaneous emission (ASE)), which increases the fundamental limits by several orders of magnitude.

The current limitations in term of linewidth are however, not from the spontaneous emission but from technical contributions, such as mechanical vibrations and intensity noise from the pump source. These contributions have however, a spectral signature in a relatively low frequency range (< 500 kHz), giving the possibility of a noise stabilization for ultralow noise operation. It also means that OPSL can reach shot-noise limited operation over a very wide frequency range in the RF-microwave domain, in spite of poor quality multimode pumping, since the high-Q cavity will filter out the noise of the pump above the cold-cavity cutoff frequency. These OPSLs properties provide a frequency noise and linewidth much lower than other laser technologies, surpassing conventional diode-laser technologies by orders of magnitude.

Quantum Dot CW OPSL Sources

OPSLs incorporating QDs as a gain medium have also been demonstrated with CW output power exceeding 8 W. In most cases the QDs are grown using the Stranski–Krastanov growth method, in place of the QWs at the anti-nodes of the standing wave in the micro-cavity. These have the advantage of being wavelength adjustable by controlling the average dot size and the inevitable inhomogeneous size distribution of the dots provides a broad but lower gain relative to the QWs. InP QDs facilitate lasing in the visible around 600–700 nm whereas InAs QD VECSELs operate at longer wavelengths around 1 μm. Since QDs inherently exhibit strong inhomogeneous gain broadening, they may not be the medium of choice for single frequency operation, since the weaker mode coupling in the gain will likely lead to multimode operation. However, the increased gain bandwidth and expected spectral hole burning could provide an excellent platform for dual or multiple frequencies operation, to generate terahertz via DFG, for example. The increased gain bandwidth should also be able to support short pulse durations in a modelocked state. Mode-locking of QD OPSLs has been reported with pulses of picosecond duration in the visible using InP and near-IR with gain layers based on self-assembled InAs QDs. Even though QDs have proven to be a suitable candidate for the gain medium of a CW or modelocked OPSL gain medium, it is not clear at this point if QDs are superior to QWs. The intrinsic low saturation fluence of QDs, which is an advantage for a SESAM, could be a major drawback for a modelocked gain medium where a strong saturation fluence is usually desired.

Low Noise Multi-Gigahertz Repetition Rate Femtosecond Pulsed OPSLs

OPSLs are emerging as new sources of low noise, multi-gigahertz repetition rate femtosecond duration mode-locked sources. The compact V-cavity, Z-cavity, or ring cavity configuration facilitates replacement of one of the cavity HR mirrors by a SESAM. The SESAM typically consists of a single QW whose absorption spectrum is matched to the central wavelength of the desired pulse. It is grown on a DBR to provide for high reflectivity when the QW absorption is saturated at high intra-cavity field intensities. SESAMs need to be designed to have as fast a carrier recombination rate as possible to ensure short pulses. This can be achieved by low temperature growth of the QW creating defects to cause rapid defect recombination or by growing the single quantum well in close vicinity to the semiconductor surface (few nanometer) in order to facilitate fast recombination with surface states.

The low gain and compact cavity of a mode-locked VECSEL creates challenges in regards to the implementation of dispersion management to minimize group delay dispersion for femtosecond pulse generation. Unlike solid state mode-locked lasers such as Ti-Sapphire, one cannot afford to add many internal dispersion compensating optical components due to their intrinsic loss even when AR-coated. Typically, dispersion control is implemented at the semiconductor epitaxial design phase and further complemented with AR-coatings on the VECSEL and SESAM chips. The modelocking of OPSLs has been demonstrated at repetition rates ranging from 85 MHz to 100 GHz, and at operating wavelength ranging from 665 to 1960 nm. To date, the best performance in terms of power and pulse duration has been demonstrated between 960 and 1080 nm, where the semiconductor media benefit from the maturity and advantages of GaAs-based materials. To date, the record peak power obtained in a modelocked state exceeds 6.3 kW with 410 fs pulses at a repetition rate of 390 MHz. The shortest pulse duration was demonstrated at around 1030 nm with pulses of 107 fs, externally compressed to 96 fs.

A compact variant of a mode-locked OPSL is the MIXSEL which involves integrating the saturable absorber into the same epitaxial structure as the gain element. The saturable absorber elements can be either a QW or a single quantum dot layer. This approach is ideally suited very high repetition rate as the cavity can be extremely compact. It appears however, that this approach is less flexible and leaves less room for growth error, which has limited the performance to lower power and longer pulses (sub-300 fs) than the external SESAM approach.

Intra-Cavity SHG and DFG OPSLs

OPSLs offer maximum flexibility in implementing various modes of intra-cavity SHG and DFG generation. The potential of achieving high kW-level internal circulating fields within the cavity guarantee high power signal generation at wavelengths spanning UV to terahertz. Frequency conversion via intra-cavity SHG has been extensively employed to generate a host of wavelengths in the visible and UV from an IR fundamental wavelength. Short UV wavelengths around 200 nm typically requires two SHG stages starting with IR wavelength originating from an InGaAs QW OPSLs frequency doubled in the OPSL cavity, the visible second harmonic output can be further converted to UV in an external resonator containing a BaB_2O_4 crystal, for example. This technique has been employed to generate stabilized single frequency output with a couple hundreds of mill watts in the UV.

OPSL are also promising for the development of higher-power terahertz sources via DFG. By designing the semiconductor structure to provide a flat broadband gain, one can exploit the unique carrier dynamics of QWs or QDs gain medium, to favor the emission of multiple wavelengths simultaneously. The spectral spacing of these multiple wavelengths can be enforced by a thin etalon in the cavity to allow DFG in the terahertz regime. The broad gain bandwidth and the etalon thickness selection and tilt can provide a large and fine tunability of the terahertz emission, from 100 GHz to several terahertz. The stability and gain competition dynamics of multiple wavelength OPSLs have been investigated and have shown that with high intracavity powers in high-Q cavities, a relatively stable simultaneous two-color operation is feasible, probably due to strong kinetic hole burning effects and/or spatial hole burning. Moreover, the circular nearly diffraction-limited beam profile is ideal for driving nonlinear processes with high efficiency. However, one has yet to investigate the influence of a nonlinear crystal, such as a periodically poled lithium niobate (PPLN) crystal, on the dynamics of an OPSL operating in a multi-color regime. A DFG in an external resonant cavity would be an alternative to prevent against instabilities caused by the frequency conversion in the crystal.

Applications

Current commercial applications of OPSLs are primarily based on intra-cavity SHG based on the InGaAs/AlGaAs/GaAs material system. These lasers emit in the near-IR and can generate impressive power in the visible and UV wavelength range. Coherent's commercial OPSL based products include OBIS Lasers for life sciences applications, from the UV (375 nm) to the near IR (785 nm); OBIS LG Lasers offer unparalleled OPSL laser performance in the smallest form factor for plug-and-play capability; Sapphire Lasers: CW Lasers available at multiple wavelengths from 458 to 594 nm and ultra-narrow linewidth versions; Verdi G Series family of OPSL lasers with powers up to 20 W at 532 nm and TracER a portable and lightweight green forensic laser system. The commercial EP-VECSEL implemented by NEXSEL provides a host of single and multiple wavelengths SHG generated sources covering the RGB spectrum mainly for commercial lighting applications. In recent years, M Squared has been commercializing modelocked OPSLs under the name of Dragonfly to replace the more conventional titanium sapphire laser. The Dragonfly series is a simple turnkey laser system providing source of picosecond mode-locked pulses with excellent beam quality near the 1 μm region. Innoptics is another emerging company developing a compact commercial VECSEL module capable to generate a low noise continuously tunable single frequency output in the near IR around 1 μm (OPSilent series) and in the mid-infrared around

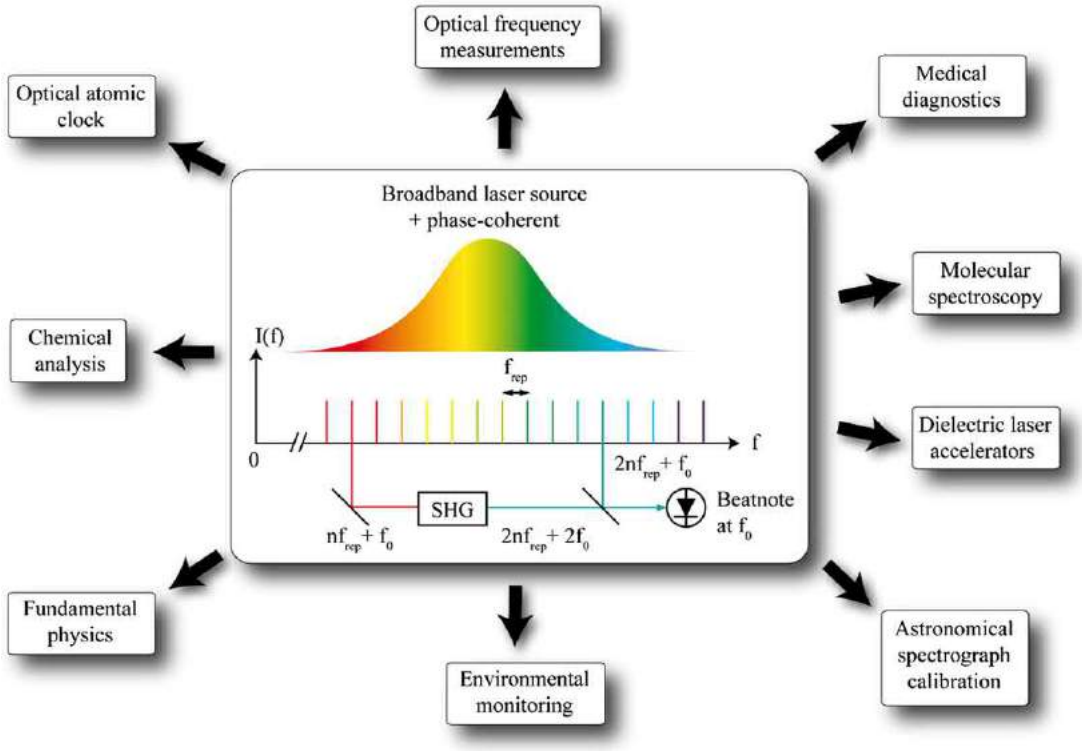

Fig. 7 Principle of a self-referenced frequency comb and potential applications.

2.3 μm (OPScan series). These wavelengths are interesting for gas analysis as they benefit from an atmospheric transparency window and from the presence of numerous gas signatures (methane, carbon monoxide, etc.).

However, OPSLs remain a nascent technology with a plethora of potential applications. Laboratory demonstrations in the CW single frequency regime have shown tremendous performance in term of coherence, tunability and power, which are extremely interesting for applications such as remote guidestar (at a wavelength of 1141 or 1178 nm doubled to 589 nm), high resolution spectroscopy, remote sensing, radar/lidar, optical metrology (frequency reference, gyroscope), or atom trapping to cite a few.

The high finesse external cavity of OPSLs give access to an intense optical field and prolonged interaction with an intracavity medium, like a nonlinear crystal for efficient SHG conversion or DFG for coherent terahertz emission, or like a dilute gas for intracavity laser-absorption spectroscopy (ICLAS) to improve the sensitivity of in-situ trace-gas analyzers.

When modelocked, OPSLs can provide a high peak power at high repetition rate, which is advantageous for nonlinear applications such as multiphoton imaging, tomography, surgery or material processing. Ultrashort pulse generation from an OPSL could also find applications in high-resolution time-domain terahertz spectroscopy with asynchronous optical sampling, ultrafast communication systems or to generate a low noise self-referenced gigahertz frequency combs. The realization of OPSLs-based frequency comb would open up a lot more applications that are summarized in Fig. 7.

Finally, OPSL cavities can be tailored to generate more exotic transverse mode and wavefront such as a vortex mode carrying an orbital angular momentum, that find applications in optical handling of microscopic particles, atoms manipulation, sub-diffraction limit microscopy, material processing, and quantum telecommunication.

Disclaimer

This material is based upon work supported under Air Force Contract No. FA9550-14-1-0062. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Air Force.

Further Reading

Hader, J., Scheller, M., Laurain, A., *et al.*, 2017. Ultrafast non-equilibrium carrier dynamics in semiconductor laser mode-locking. Semiconductor Science and Technology 32, 013002.

Moloney, J.V., Hader, J., Koch, S.W., 2007. Quantum design of semiconductor active materials: Laser and amplifier applications. Laser & Photonics Reviews 1, 24–43.

Okhotnikov, O.G., 2010. Semiconductor Disk Lasers. New York, NY: John Wiley & Sons.

Relevant Website

www.nlcstr.com
Nonlinear Control Strategies Inc. (SimuLaseTM).

Optical Parametric Amplifiers

Giulio Cerullo, Sandro De Silvestri, and Cristian Manzoni, Institute of Photonics and Nanotechnology, Milano, Italy

Introduction

Thanks to spectacular technological progress over the last three decades, nowadays sources of ultrashort (femtosecond and picosecond) laser pulses have become reliable, rugged and power scalable, and are used in many fields, ranging from fundamental studies in physics, chemistry and biology to industrial and biomedical applications. Such sources are typically based on solid state gain media, such as Ti:sapphire or Nd:Yb-doped crystals, which typically emit at specific wavelengths (800 nm for Ti:sapphire and $\approx$ 1 μm for Yb and Nd) with very limited tunability.

On the other hand, many applications require tunability of the ultrashort pulses, over a broad range from the ultraviolet (UV) to the mid-infrared (mid-IR). As there are no classical laser active media, based on population inversion, capable of providing gain over such a broad frequency range, frequency tunability must be achieved by a nonlinear optical effect, exploiting the very high peak powers produced by ultrashort pulse laser sources. In most cases, frequency tunability is achieved by the second order nonlinear optical effect known as optical parametric amplification (OPA) (Giordmaine and Miller, 1965; Baumgartner and Byer, 1979). The principle of OPA is quite simple (**Fig. 1(a)**): in a suitable nonlinear crystal, energy is transferred from an high frequency and high intensity beam (the pump beam, at frequency ω_3) to a lower frequency, lower intensity beam (the signal beam, at frequency ω_1) which is thus amplified; in addition a third beam (the idler beam, at frequency ω_2) is generated, to fulfill energy conservation. The OPA process can have a simple corpuscular interpretation (see **Fig. 1(b)**): a photon at frequency ω_3 is absorbed by a virtual level of the material and a photon at frequency ω_1 stimulates the emission of two photons at frequencies ω_1 and ω_2. In this interaction, energy conservation:

$$\hbar\omega_3 = \hbar\omega_1 + \hbar\omega_2 \tag{1}$$

is fulfilled. The signal frequency to be amplified can vary in principle from 0 to ω_3, and correspondingly the idler varies from ω_3 to 0 (as a matter of fact, the lowest frequency is limited by absorption of the nonlinear crystal). The so-called degeneracy condition is achieved when $\omega_1 = \omega_2 = \omega_3/2$. As will be discussed in the next section, the OPA process is efficient when, in addition to energy conservation, the phase matching condition

$$\Delta\mathbf{k} = \mathbf{k}_3 - \mathbf{k}_1 - \mathbf{k}_2 = 0 \tag{2}$$

is satisfied, where $\mathbf{k}_i$ are the wave vectors of the interacting beams.

In summary, the OPA process transfers energy from a high-power, fixed frequency pump beam to a low-power, variable frequency signal beam, generating an idler beam to satisfy energy conservation. The OPA can thus be seen as a "photon cutter," cutting a high energy photon $\hbar\omega_3$ into the sum of two lower energy photons $\hbar\omega_1$ and $\hbar\omega_2$. The OPA process thus provides an optical amplifier with continuously variable center frequency (determined by the phase-matching condition) and represents an easy way of tuning over a broad range the carrier frequency of an otherwise fixed wavelength laser system. On the other hand, if suitably designed, an OPA can simultaneously fulfill the phase matching condition for a very broad range of signal frequencies. The OPA thus acts as a broadband amplifier, efficiently transferring energy from a narrowband pump pulse to a broadband signal (idler) pulse; it can therefore be used to dramatically shorten, by more than an order of magnitude, the duration of the pump pulse. The concept of broadband OPA is very flexible and can be applied to produce few-optical-cycle pulses over a very wide range of frequencies, provided that a broadband yet weak seed is available and the proper phase-matching conditions are identified. This article is a self-consistent review of OPAs and is organized as follows: Section Theory of Optical Parametric Amplification introduces the theory of OPA; Section Phase Matching discusses the concept of phase matching; Section Architecture of an OPA presents the general architecture of an OPA and separately analyzes the different components; Section Ultra-Broadband OPAs describes ultra-broadband OPAs; finally, Section Optical Parametric Chirped pulse Amplification introduces the concept of the optical parametric chirped pulse amplifier (OPCPA).

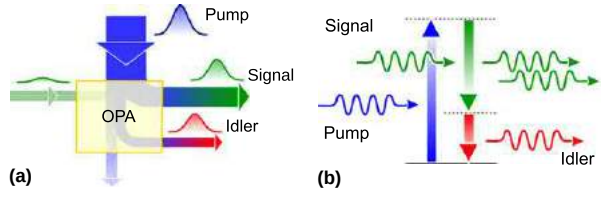

Fig. 1 (a) General scheme of the optical parametric amplifier (OPA) process and (b) corpuscular interpretation of the OPA.

 doi:10.1016/B978-0-12-803581-8.09632-6

Theory of OPA

Let us consider an optical field, consisting of the superposition of three monochromatic waves, at frequencies ω_1, ω_2, and ω_3:

$$E(z,t) = \frac{1}{2}\{A_1(z)\exp[i(\omega_1 t - k_1 z)] + A_2(z)\exp[i(\omega_2 t - k_2 z)] + A_3(z)\exp[i(\omega_3 t - k_3 z)] + c.c.\} \tag{3}$$

satisfying the condition $\omega_1 + \omega_2 = \omega_3$, impinging on a non-centrosymmetric medium generating a second order nonlinear polarization:

$$P_{\mathrm{NL}}(z,t) = \varepsilon_0 \chi^{(2)} E^2(z,t) \tag{4}$$

Such a situation is known as "nonlinear second-order parametric interaction" and corresponds to an exchange of energy between the three fields mediated by the second order nonlinearity. The nonlinear polarization contains three components at frequencies ω_1, ω_2 and ω_3, given by

$$P_{1\mathrm{NL}}(z,t) = \frac{\varepsilon_0 \chi^{(2)}}{2} A_2^* A_3 \exp\{i[\omega_1 t - (k_3 - k_2)z]\} + c.c. \tag{5a}$$

$$P_{2\mathrm{NL}}(z,t) = \frac{\varepsilon_0 \chi^{(2)}}{2} A_1^* A_3 \exp\{i[\omega_2 t - (k_3 - k_1)z]\} + c.c. \tag{5b}$$

$$P_{3\mathrm{NL}}(z,t) = \frac{\varepsilon_0 \chi^{(2)}}{2} A_1 A_2 \exp\{i[\omega_3 t - (k_1 + k_2)z]\} + c.c. \tag{5c}$$

Obviously there are other terms on P_{NL} at different frequencies, such as, for example, $2\omega_1$, $2\omega_2$, ω_1–ω_2… Here we consider only the terms at ω_1, ω_2 and ω_3 because we assume that only the interaction between these three fields is efficient, due to the phase-matching condition.

The nonlinear polarization at frequency ω_j can thus be written as:

$$P_{\mathrm{NL}j}(z,t) = \frac{1}{2} p_{\mathrm{NL}j}(z) \exp\{i[\omega_j t - k_{pj} z]\} + c.c. \tag{6}$$

where we emphasize that the wave vector of the polarization k_{pj} is in general different from that of the corresponding field. The propagation equation for each field in the nonlinear medium can be written, in the monochromatic limit, as:

$$\frac{\partial A_j}{\partial z} = -i\frac{\mu_0 \omega_j c}{2 n_j} p_{\mathrm{NL}j} \exp[-i\Delta k_j z] \tag{7}$$

where $\Delta k_j = k_{pj} - k_j$ is the so-called "wave-vector mismatch" between the nonlinear polarization and the field. By inserting Eq. (5) into Eq. (7), we can derive the following three equations for the fields at ω_1, ω_2 and ω_3 (Boyd, 2003):

$$\frac{\partial A_1}{\partial z} = -i\kappa_1 A_2^* A_3 \exp[-i\Delta k z] \tag{8a}$$

$$\frac{\partial A_2}{\partial z} = -i\kappa_2 A_1^* A_3 \exp[-i\Delta k z] \tag{8b}$$

$$\frac{\partial A_3}{\partial z} = -i\kappa_3 A_1 A_2 \exp[i\Delta k z] \tag{8c}$$

where we have defined the nonlinear coupling coefficients: $\kappa_i = \frac{\omega_i \chi^{(2)}}{4 c n_i}$ and the "wave-vector mismatch" as: $\Delta k = k_3 - k_1 - k_2$. Note that the first two equations are formally identical, indicating that the fields at ω_1 and ω_2 play the same role. According to the boundary initial conditions $A_i(0)$ ($i = 1, 2, 3$), two main nonlinear processes can arise: sum frequency generation (SFG) and difference frequency generation (DFG). In SFG the input fields are $A_1(0)$ and $A_2(0)$ at ω_1 and ω_2; their interaction produces the field A_3 at frequency $\omega_3 = \omega_1 + \omega_2$. Second harmonic generation (SHG) is just a particular case of SFG in which $\omega_1 = \omega_2$. In DFG, the input fields are $A_3(0)$ at ω_3 and $A_1(0)$ at ω_1; the field A_3 loses energy in favor of A_1 and of the newly generated field A_2 at $\omega_2 = \omega_3 - \omega_1$. The OPA is a mechanism similar to DFG, except for the strength of the interacting fields: DFG arises when the fields at ω_3 and ω_1 have comparable intensities, while for an OPA the field at ω_1 is much weaker. In the OPA process, therefore, an intense beam at ω_3 (the pump) transfers energy to the beam at ω_1 (the signal), thereby amplifying it, and generates light at ω_2 (the idler). Note that the DFG/OPA processes also occur starting from the fields $A_3(0)$ at ω_3 and $A_2(0)$ at ω_2; in this case the newly generated field is at ω_1. This means that the signal and idler play an interchangeable role in the OPA process.

Eq. (8) can be solved by making the assumption that the pump field is not depleted during the interaction ($A_3 \approx$ const.), and that there is no initial idler beam ($A_{20} = 0$). After proper manipulation, the signal evolution along the nonlinear medium can be written as:

$$\frac{\mathrm{d}^2 A_1}{\mathrm{d}z^2} = -i\Delta k \frac{\mathrm{d}A_1}{\mathrm{d}z} + \Gamma^2 A_1 \tag{9}$$

where $\Gamma^2 = \frac{2 d_{\mathrm{eff}}{}^2 \omega_1 \omega_2}{c_0^3 \varepsilon_0 n_1 n_2 n_3} \mathrm{I}_3$ and $I_3 = \frac{1}{2} n_3 c \varepsilon_0 |A_3|^2$ is the pump beam intensity. Γ is known as the small signal gain of the OPA and $d_{\mathrm{eff}} = \frac{\chi^{(2)}}{2}$ is the element of the $\chi^{(2)}$ tensor that keeps trace of the polarization of the beams and their propagation directions within the nonlinear crystal lattice.

In the hypothesis of an initial signal field amplitude $A_1(0)=A_{10}$ (the "seed beam") and no initial idler ($A_2(0)=0$), the solutions of Eq. (9) are:

$$I_1(z) = I_{10}\left\{1 + \left[\frac{\Gamma}{g}\mathrm{senh}(gz)\right]^2\right\} \tag{10a}$$

$$I_2(z) = I_{10}\frac{\omega_2}{\omega_1}\left[\frac{\Gamma}{g}\mathrm{senh}(gz)\right]^2 \tag{10b}$$

where $g = \sqrt{\Gamma^2 - \left(\frac{\Delta k}{2}\right)^2}$. For the case of large gain ($gz \gg 1$) Eq. (10) further simplify to:

$$I_1(z) = \frac{I_{10}}{4}\left(\frac{\Gamma}{g}\right)^2 \exp(2gz) \tag{11a}$$

$$I_2(z) = \frac{\omega_2}{\omega_1} I_1(z) \tag{11b}$$

giving an exponential growth of both signal and idler intensities with crystal length, characteristic of an optical amplifier. It should also be noted, that, in the large gain limit, Eq. (11b) can be rewritten as:

$$\frac{I_2(z)}{\hbar\omega_2} = \frac{I_1(z)}{\hbar\omega_1} \tag{12}$$

which means that, in the large gain limit, signal and idler intensities are related by energy conservation, since for each annihilated pump photon a signal and an idler photon are simultaneously generated.

The parametric gain for the signal beam, in the large gain limit, can be written as:

$$G = \frac{I_1(z)}{I_{s0}} = \frac{1}{4}\left(\frac{\Gamma}{g}\right)^2 \exp(2gz) \tag{13}$$

which, for perfect phase matching ($\Delta k=0$) becomes:

$$G = \frac{1}{4}\exp(2\Gamma z) \tag{14}$$

The exponential growth of the gain with propagation length is qualitatively different from the quadratic growth that occurs in other second-order processes like SFG and SHG and shows that the OPA behaves like a real amplifier. With respect to a classical optical amplifier based on population inversion in an atomic or molecular transition, however, an OPA has three important differences:

1. it does not have any energy storage capability, i.e., the gain is present only during the temporal overlap of the interacting pulses;
2. the gain center frequency is not fixed, but can be continuously adjusted by varying the phase-matching condition;
3. as we will see in Section Phase Matching Bandwidth of an OPA, the gain bandwidth is not limited by the linewidth of an electronic/vibrational transition, but rather by the possibility of satisfying the phase-matching condition over a broad range of frequencies.

Let us now consider the factors influencing the parametric gain G:

1. $G \propto \exp(g)$ exponentially depends on the parameter g, which is maximum when $\Delta k=0$ (phase-matching condition). G rapidly decreases for nonzero values of Δk, suggesting that phase-matching is a key condition to be fulfilled in order to get significant amplification from the OPA process.
2. $G \propto \exp(d_{\mathrm{eff}})$ depends exponentially on the second order nonlinear optical coefficient of the crystal d_{eff}; one should therefore select the crystal with the largest nonlinear response. There are however other considerations leading to the choice of the crystal, such as phase matching range, dispersive properties, and optical damage threshold.
3. $G \propto \exp\left(\sqrt{I_3}\right)$ scales as the exponential of the square root of the pump intensity. This indicates the suitability of ultrashort pulses for OPAs, due to their high peak powers. One should try to use the highest possible pump intensity before the onset of other nonlinear optical phenomena such as self-focusing, self-phase modulation and beam breakup. In order to be able to use high pump intensities, it is however important to have a spatially clean beam profile, without hot spots.
4. $G \propto \exp(L)$ scales as the exponential of the crystal length L, as in an optical amplifier. With ultrashort light pulses, however, the optimum crystal length has to be chosen considering the durations and group velocities of the interacting pulses.
5. $G \propto \exp\left(\sqrt{\omega_1\omega_2}\right)$ scales as the exponential of the square root of the product of signal and idler frequencies. This seems to indicate an advantage to use higher pump frequencies. However this advantage is often offset by the larger difference in group velocities of the interacting pulses.

Phase Matching

Interaction Types in an OPA

Eq. (8) describes the nonlinear interaction between pump, signal and idler waves, assuming the same propagation direction (collinear interaction). In general, the interaction involves fields with different polarizations and different propagation directions; a complete description of the most general case goes beyond the purpose of this article. To take polarization of the fields into account, it is often sufficient to reduce the nonlinear susceptibility tensor to the scalar coefficient d_{eff}, which accounts for the average interaction strength among the fields. For the sake of clarity, we will limit ourselves here to the simpler, and more commonly adopted, uniaxial birefringent crystals; the reader is referred to (Dmitriev *et al.*, 1999) for a complete treatment of the case of biaxial crystals. In uniaxial crystals, the best reference frame to describe the electric field of an electromagnetic wave is defined by the optical axis of the medium and the wave-vector **k** of the propagating beam, which form a plane σ and two normal polarization directions, called ordinary (o) and extraordinary (e). According to the polarization configurations of the three interacting fields, we group them in the following interaction types:

Type 0: the three interacting waves have equal polarization states, either ordinary (ooo) or extraordinary (eee).

Type I: the low-frequency fields (ω_1, ω_2) propagate with the same polarization, either ordinary or extraordinary. The resulting configurations are therefore ooe or eeo, respectively.

Type II: the low-frequency fields are cross-polarized, and the possible configurations are eoe, oee, eoo, and oeo.

From the polarization configuration of the fields it is possible to evaluate the two important parameters governing the nonlinear interaction: the nonlinear effective coefficient d_{eff}, which accounts for the second-order nonlinear response of the medium, and the frequency-dependent phase-mismatch Δk, which determines the efficiency of the energy flow between the interacting fields.

Phase Matching Methods

Eq. (10) shows that the nonlinear parametric interaction is optimized when $\Delta k = 0$, at perfect phase matching. In a collinear configuration, recalling that $k = \omega n / c_0$, the phase-matching condition is equivalent to

$$n_3\omega_3 - n_2\omega_2 - n_1\omega_1 = 0 \tag{15}$$

which links the refractive indexes n_i seen by the interacting waves. In general, it can be shown that it is not possible to fulfill this condition in an isotropic medium, due to material dispersion. To achieve phase matching, two approaches can be followed:

1. exploit birefringence of the non-centrosymmetric nonlinear crystal by propagating the three interacting beams with different polarizations, leading to Type-I and Type-II interaction schemes (see **Fig. 2(a)**);
2. apply a periodic modulation to the sign of the nonlinear coefficient d_{eff}, leading to the so-called quasi-phase-matching (QPM) regime (Hum and Fejer, 2007) (see **Fig. 2(c)**); in this case the fields can propagate with parallel polarization (Type 0). Phase matching is not locally satisfied, but the modulation of the sign of d_{eff} leads to an average macroscopic net exchange of energy between the fields.

To study phase matching by birefringence, for simplicity we will again limit ourselves to uniaxial crystals. These crystals have two frequency-dependent refractive indexes: the ordinary index n_o and the extraordinary index n_e; if $n_o < n_e$, the crystal is called positive uniaxial crystal, whereas if $n_o > n_e$ it is called negative uniaxial crystal. The ordinary polarization propagates with index of refraction n_o, while the extraordinary polarization experiences an index of refraction which varies from n_o to n_e as a function of the

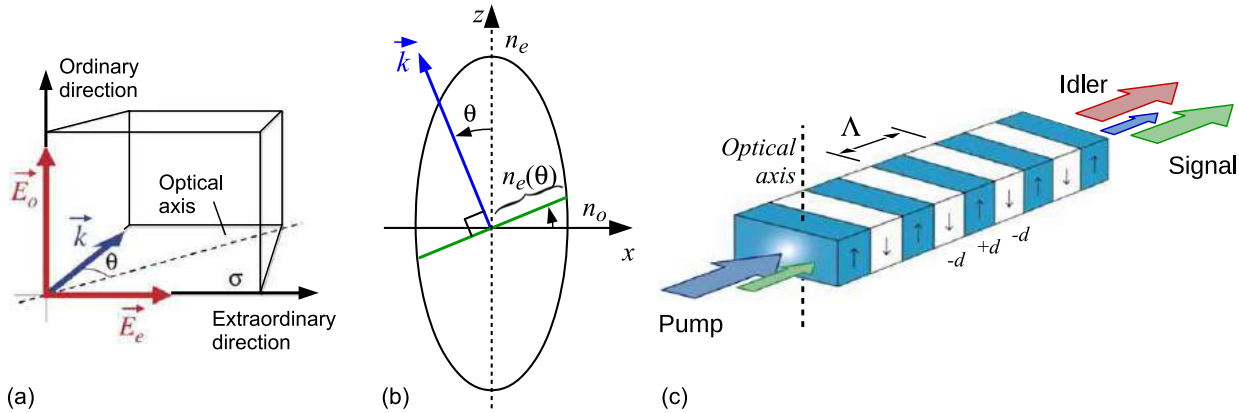

Fig. 2 (a) Ordinary and extraordinary directions in uniaxial birefringent crystals. (b) Section of the index ellipsoid and graphical meaning of ordinary and extraordinary refractive indices, for a positive uniaxial crystal. The optical axis is in the *z* direction. (c) Sketch of a periodically-poled crystal and the propagating fields of an optical parametric amplifier (OPA).

angle θ between $\mathbf{k}$ and the optical axis, according to the equation:

$$\frac{1}{n_e^2(\theta)} = \frac{\cos^2(\theta)}{n_o^2} + \frac{\sin^2(\theta)}{n_e^2} \tag{16}$$

By changing the propagation direction inside the crystal, $n_e(\theta)$ can thus be tuned from n_o to n_e (see **Fig. 2(b)**): in this way, there often exists one value of θ giving the right n_e which fulfills the phase-matching condition of Eq. (15). This angle is indicated with θ_m and is called phase-matching angle. If only one field has extraordinary polarization, θ_m can be exactly calculated from Eqs. (15) and (16), while the calculation is more involved if two fields have extraordinary polarization.

The QPM approach to phase matching is typically used with those nonlinear crystals whose $\chi^{(2)}$ tensor is such that the highest nonlinear coefficient manifests when the three interacting beams have the same polarizations, as for stoichiometric lithium tantalate (SLT) or lithium niobate ($LiNbO_3$, LN). In these crystals the strongest element of the nonlinear tensor is d_{33}, which contributes to the nonlinear process when the three interacting beams share the same extraordinary polarization (Type 0 configuration). In this case birefringence phase-matching can never be achieved, as discussed above. In QPM the nonlinear properties of the medium are spatially modulated by periodically changing the sign of the $\chi^{(2)}$ coefficient, reversing the alignment of the ferroelectric domains in the birefringent crystal upon application of a suitable voltage. When the modulation of the nonlinear coefficient is periodic with grating period Λ, the phase-matching condition now includes the grating wave-vector $K_g = 2\pi/\Lambda$ and reads:

$$\Delta k = k_3 - k_1 - k_2 - K_g = 0 \tag{17}$$

Finding the phase-matching condition, for the case of QPM, corresponds to calculating the poling period Λ_m.

Phase Matching Bandwidth of an OPA

Let us now discuss which conditions determine the gain bandwidth of an OPA. Ideally one would like to have a broadband amplifier, i.e., an amplifier which, for a fixed pump frequency $\overline{\omega}_3$, provides a more or less constant gain over an as broad as possible range of signal frequencies. To this end, one needs to keep the phase mismatch Δk as small as possible over a large bandwidth. Practically, however, the phase-matching condition can be satisfied only for a given set of frequencies $(\overline{\omega}_1, \overline{\omega}_2, \overline{\omega}_3)$, such that

$$\Delta k = k(\overline{\omega}_3) - k(\overline{\omega}_1) - k(\overline{\omega}_2) = 0 \tag{18}$$

If the pump frequency is fixed at $\overline{\omega}_3$ and the signal frequency changes to $\overline{\omega}_1 + \Delta\omega$, then by energy conservation the idler frequency changes to $\overline{\omega}_2 - \Delta\omega$. The ensuing wave vector mismatch becomes

$$\Delta k = k(\overline{\omega}_3) - [k(\overline{\omega}_1) + \Delta k_1] - [k(\overline{\omega}_2) + \Delta k_2] = -\Delta k_1 - \Delta k_2 \tag{19}$$

which can be approximated to the first order as

$$\Delta k = -\frac{\partial k_1}{\partial \omega_1}\Delta\omega + \frac{\partial k_2}{\partial \omega_2}\Delta\omega = \delta_{12}\Delta\omega \tag{20}$$

where $\delta_{12} = \frac{1}{v_{g2}} - \frac{1}{v_{g1}}$ is the group velocity mismatch (GVM) between signal and idler pulses. The full width at half maximum parametric gain bandwidth for a crystal of length L can then be calculated from Eq. (13), within the large gain and low pump-depletion approximations, as:

$$\Delta\omega = \frac{4\log(2)^{1/2}}{\pi}\left(\frac{\Gamma}{L}\right)^{1/2}\frac{1}{|\delta_{12}|} \tag{21}$$

Eq. (21) shows that the gain bandwidth is inversely proportional to the GVM between signal and idler and has only a square root dependence on small-signal gain Γ and crystal length L. For the case when $v_{g1} = v_{g2}$ (group-velocity matched OPA), Eq. (21) loses validity and Eq. (20) must be expanded to the second order in $\Delta\omega$, giving

$$\Delta\omega = \frac{4\log(2)^{1/4}}{\pi}\left(\frac{\Gamma}{L}\right)^{1/4}\frac{1}{\left|\frac{\partial^2 k_1}{\partial\omega_1^2} + \frac{\partial^2 k_2}{\partial\omega 2}\right|^{1/2}} \tag{22}$$

where $\frac{\partial^2 k}{\partial\omega^2}$ is the group velocity dispersion (GVD) of the pulses. In this case the gain bandwidth is thus inversely proportional to the square root of the sum of the GVDs of signal and idler pulses. The conditions for group-velocity matched OPAs will be discussed in a later section.

Architecture of an OPA

General OPA Scheme

The conceptual scheme of an OPA pumped by ultrashort pulses is summarized in **Fig. 3**. It fundamentally consists of three subsystems (Cerullo and De Silvestri, 2003):

1. A seed pulse generation stage, which exploits a nonlinear optical process to generate, starting from the pump pulse, from its second harmonic (SH) or from its third harmonic (TH), a weak pulse at the signal frequency that initiates the OPA process.

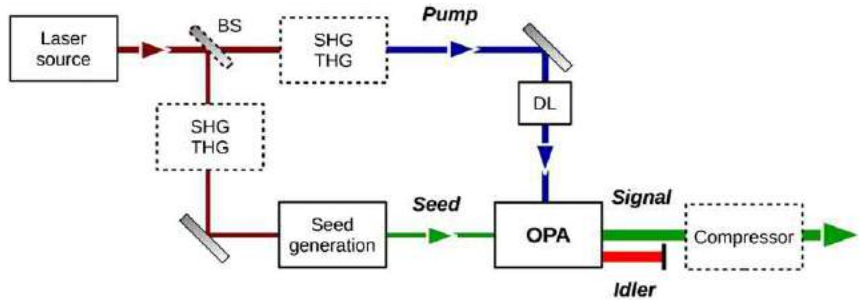

Fig. 3 General scheme of an optical parametric amplifier (OPA). BS: beam splitter; DL: delay line; SHG, THG: second- and third-harmonic generation stages.

2. Parametric amplification in one or more gain stages, pumped either by the fundamental wavelength (FW) of the driving laser or by its SH/TH.
3. An optional pulse compression stage by a dispersive delay line, to obtain the minimum pulse duration compatible with its bandwidth, the so-called transform-limited (TL) pulse duration.

Analyzing the operation of these three blocks provides the key for understanding the properties of the OPA and in particular the design required to achieve specific working parameters.

Seed Generation

In standard OPA systems, the two most widely adopted schemes for producing the seed pulse are optical parametric generation (OPG) and white-light continuum (WLC) generation. OPG, also known as parametric superfluorescence (PSF), is parametric amplification of the vacuum or quantum noise, which can also be thought of as two-photon spontaneous emission from a virtual level excited by the pump field (Harris *et al.*, 1967). In practice it is simply achieved by pumping a suitable second order nonlinear crystal, which is often of the same Type as the ones used in the subsequent OPA stages; OPG occurs at those frequencies for which the parametric interaction is phase-matched. Since PSF is a stochastic process that starts from noise, it has the disadvantages of an inherently large shot-to-shot energy fluctuations of the seed and a poor spatial beam quality. In addition, the longitudinal position inside the nonlinear crystal at which amplification takes place, and thus the absolute timing of the seed pulse with respect to the pump, varies stochastically in time. The main advantage of OPG is the relatively modest pulse energy requirement, especially using crystals with very high effective nonlinearity such as periodically poled LN and SLT. This feature is particularly useful in OPAs which are driven by high-repetition-rate, low-pulse-energy oscillators. One final remark about OPG is that it often occurs as a parasitic effect, as it can be generated in the initial stage(s) of the OPA, in the absence of a proper seed, and then amplified in the subsequent stages, thus spoiling the properties (energy stability, spectral purity) of the amplified pulses.

WLC generation (Gaeta, 2000) is achieved by focusing the ultrashort pulse in a suitable transparent dielectric medium (typically a sapphire plate) displaying a third-order ($\chi^{(3)}$) nonlinearity. WLC is a complicated nonlinear optical process, involving the interplay and coupling of temporal (self-phase modulation, dispersion, self-steepening) and spatial (self-focusing, diffraction) effects, as well as ionization and plasma generation, leading to the formation of a filament and to dramatic spectral broadening. WLC generation is performed by focusing a fraction of the driving pulse (typically the FF off the laser, but sometimes also the SH or the TH) into the plate (with thickness from 1 to 10 mm); by adjusting the pulse energy (using a variable attenuator), the focusing conditions (by moving the plate in the focus of the beam) and the numerical aperture of the focused beam (often by a diaphragm placed in front of the focusing lens) a single filament WLC is formed. Typical spectral energy density of the WLC is 10–30 pJ/nm in the visible range. The WLC displays excellent shot-to-shot stability and diffraction-limited spatial beam quality, so that it is the ideal seed for an OPA, although with limited spectral energy density. Using 800 nm Ti:sapphire driving lasers, with pulse duration of 100 fs or shorter, the best medium for WLC generation is undoped sapphire which, due to its very high thermal conductivity and damage threshold, guarantees excellent stability and damage-free operation. Using 1030 nm Yb-based driving lasers, which generate longer pulses, sapphire is not a suitable material for WLC generation. Other media with lower bandgap, such as undoped YAG, allow the generation of stable WLC under these conditions Bradler *et al.* (2009). The WLC provides a very broadband seed; the duration of the OPA pulses will depend on the fraction of this broad bandwidth which can be amplified, i.e., on the gain bandwidth of the OPA crystal(s), which has been discussed in Section Phase Matching Bandwidth of an OPA.

Parametric Amplification

In the parametric amplification stage(s) energy is transferred from the pump to the seed pulses. In order to optimize this process, one should carefully adjust the pump-pulse (wavelength, energy, and duration) and the nonlinear interaction (crystal type, crystal length, and phase-matching Type) parameters.

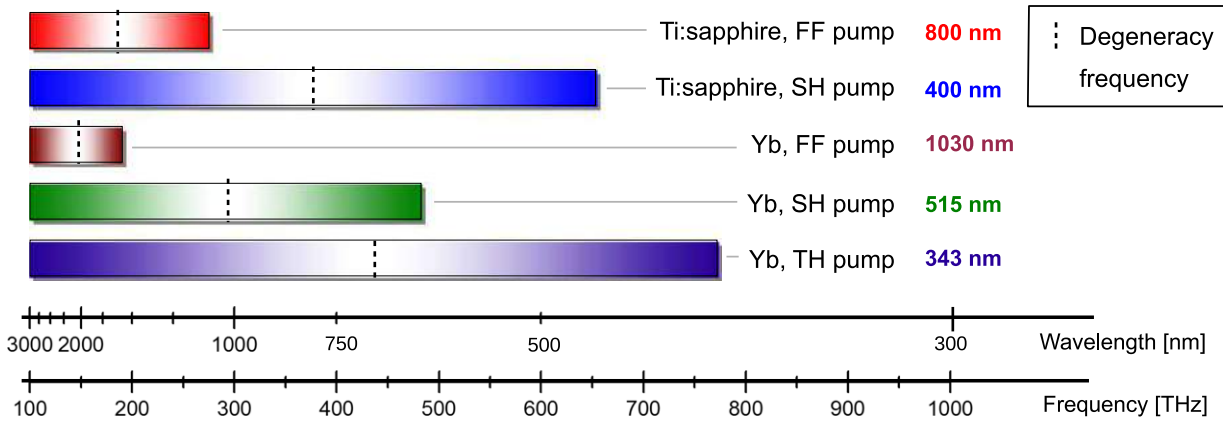

Fig. 4 Map of tuning ranges of an optical parametric amplifier (OPA) based on β-barium borate (BBO), whose infrared transparency is limited to 3 μm. The degeneracy point is at half the pump frequency.

Typically gain in an OPA is achieved in multiple stages, with the number dictated by the final pulse energy: a preamplifier (or signal amplifier, 1st stage) starting from the low-energy seed beam and displaying high gain (from 10^3 to 10^5) and one (2nd stage) or more power amplifiers, displaying relatively low gain (from 10 to 10^2). The multistage approach has several advantages:

1. it allows to compensate for the GVM arising between pump and signal pulses within each stage;
2. it enables to adjust the pump intensity, and thus the parametric gain, separately in the different stages. Most of the available pump energy should be used for the final stage, which is driven into saturation in order to efficiently convert the available pump energy and minimize shot-to-shot energy fluctuations of the amplified signal. The spot size of the beams progressively increases for the subsequent stages, in order to accommodate higher pulse energies while keeping comparable peak intensities.

Several criteria influence the choice of the pump wavelength for an OPA. The most important is the required tuning range, which depends on the pump wavelength, on the possibility to satisfy the phase matching condition and on the transparency range of the nonlinear crystal. Absorption is one of the most important factors limiting the tuning of an OPA; care must be taken in order to have both signal and idler frequencies in the transparency range of the nonlinear crystal(s).

Fig. 4 summarizes the OPA tuning ranges for the popular β-barium borate (BBO) crystal, when pumped by the FF and the SH of Ti:sapphire and by the FF/SH/TH of Yb. For this crystal and these pumping conditions wavelength tunability is between 390 nm and $\approx$3 μm, limited by the pump pulse wavelength and by the onset of IR absorption in BBO at wavelengths longer than 3 μm. Of course other frequency ranges can be covered by nonlinear frequency conversion of the OPA output, for example, by SHG of or by SFG with the FF or the SH of the driving pulses. In particular, the mid-IR region (3–20 μm), which is very important for vibrational spectroscopy, can be covered by DFG between the signal and the idler of a 800 nm pumped OPA. By inspecting Fig. 4, one notices that for FF-pumped OPAs signal/idler tunability is limited to the IR. By frequency doubling an FF-pumped OPA, one can only generate signal wavelengths down to 550 nm, thus leaving a significant portion of the visible range uncovered. The visible region (450–700 nm), which is very important for spectroscopic applications, can be well covered by a SH-pumped Ti:sapphire OPA; on the other hand, for Yb lasers a visible OPA requires pumping with the TH, with an additional frequency conversion process which reduces the overall conversion efficiency.

The tuning range of OPAs in the visible is limited to $\approx$400 nm (when pumped by the TH of an Yb laser) by the energy of the pump photon; pumping with more energetic photons, such as, for example, the TH of Ti:sapphire at 266 nm, is difficult because of the onset of strong two-photon absorption in the nonlinear crystals at the intensities required for the OPA process; generation of tunable UV pulses is nevertheless possible by SHG of the visible OPA or SFG with the FF pulses. Long-wavelength tuning is limited in BBO to 3 μm by the onset of mid-IR absorption. Other crystals with higher refractive index display a more extended mid-IR transparency, allowing the direct generation of mid-IR idler pulses out to a wavelength of $\approx$5 μm. Examples of these crystals are $LiIO_3$, $KNbO_3$, $KTiOPO_4$ (KTP) or its isomorphs, magnesium-oxide (MgO):$LiNbO_3$, LN and SLT. These crystals are however not transparent in the 5–10 μm region (the fingerprint region), where many molecular vibrational transitions occur. There are crystals with extended mid-IR transparency, such as $ZnGeP_2$ (ZGP), GaSe, $AgGaSe_2$ and $AgGaS_2$ (AGS), but they all display absorption onsets in the visible (0.74 μm for ZGP, 0.65 μm for GaSe, 0.73 μm for $AgGaSe_2$ and 0.53 μm for AGS) so that they all suffer from two-photon absorption when pumped at 800 nm, preventing the direct generation of mid-IR pulses as the idler of an 800 nm pumped OPA for wavelengths longer than 5 μm. The situation improves slightly at 1030 nm wavelength, which allows pumping of the AGS crystal and direct generation of idler pulses tunable out to 7 μm. However, due to the large GVM between signal and idler, the amplified bandwidth is rather narrow, resulting in pulsewidths of 160 fs or longer. There is a current ongoing research effort in developing powerful sources of ultrashort pulses at 2 μm, based on Ho: or Tm:doped gain media, which would allow direct pumping of mid-IR OPAs based on the above mentioned crystals.

According to the previous discussion, the choice of the nonlinear OPA crystal(s) is hence determined by the balance of several parameters: the transparency range; the nonlinear coefficients; the nonlinear interaction parameters; the optical damage threshold;

the availability and price. Due to the combination of several of the above listed parameters (Nikogosyan, 1991), BBO is the crystal of choice for many OPA applications. It has a broad enough transparency range to allow amplification from 400 nm to 3 μm. Its nonlinear coefficient is moderately high, with $d_{eff} \approx 2$ pm/V for Type I OPA pumped at 400 nm. Bismuth triborate (BiB_3O_6, or BIBO) is a biaxial crystal which possesses extremely large parametric gain bandwidth for degenerate collinear interaction when pumped at 800 nm. Its main advantage is the high second order nonlinear susceptibility ($d_{eff} = 3.2$ pm/V for Type-I SHG). Lithium triborate (LiB_3O_5, LBO), on the other hand, despite its lower nonlinear coefficient, has high damage threshold, which makes it suited for the final power amplification stages. LN and SLT have IR transparency extended to 5.2 μm and 5.5 μm, respectively, which make them suited for amplification in the infrared. For both crystals, the highest element of the nonlinear tensor is d_{33} (27 pm/V for LN, and 21 pm/V for SLT), an extremely high nonlinearity which can be only exploited with Type-0 interaction. For this reason, they are typically periodically poled to obtain QPM. LN has low optical and photorefractive damage threshold, which can be increased by doping the crystal with MgO without affecting its high nonlinear coefficient. In power stages or booster amplifiers, PPSLT is preferred thanks to its higher damage threshold.

Pulse Compression

The third optional subsystem of an OPA is pulse compression, to compensate for the spectral phase accumulated by the amplified pulses, and achieve TL pulse duration. Typically, for visible/near-IR wavelengths, the OPA pulses are positively chirped due to material dispersion introduced by the seed generation process (e.g., by linear propagation of the WLC in the generation plate), by optical elements in the signal path (lenses, beam splitters, spectral filters ...) and by the OPA crystals. For narrow amplified pulse bandwidths (i.e., corresponding to TL pulsewidths of ≈50 fs or longer) the effects of such dispersion are negligible, so that no compression is necessary and the OPA pulses can be used as generated. For broader bandwidths, pulse compression is necessary to achieve the TL pulse duration. For moderately broad bandwidths, it is sufficient to correct the second-order dispersion, or group delay dispersion (GDD), while for broader bandwidths, and in particular for sub-10-fs TL pulses, it is also necessary to simultaneously correct for the third order dispersion (TOD).

Fig. 5 shows some of the several available optical systems which are able to provide negative dispersion in the visible/near-IR including grating pairs, prism pairs, chirped mirrors (CMs) and adaptive optical systems.

Grating pairs (see panel (a)) are not commonly used with OPAs due to their high losses and the unnecessarily large GDD values that they introduce. The simplest compressor consists of a Brewster-cut prism pair in a folded configuration (panel (b));

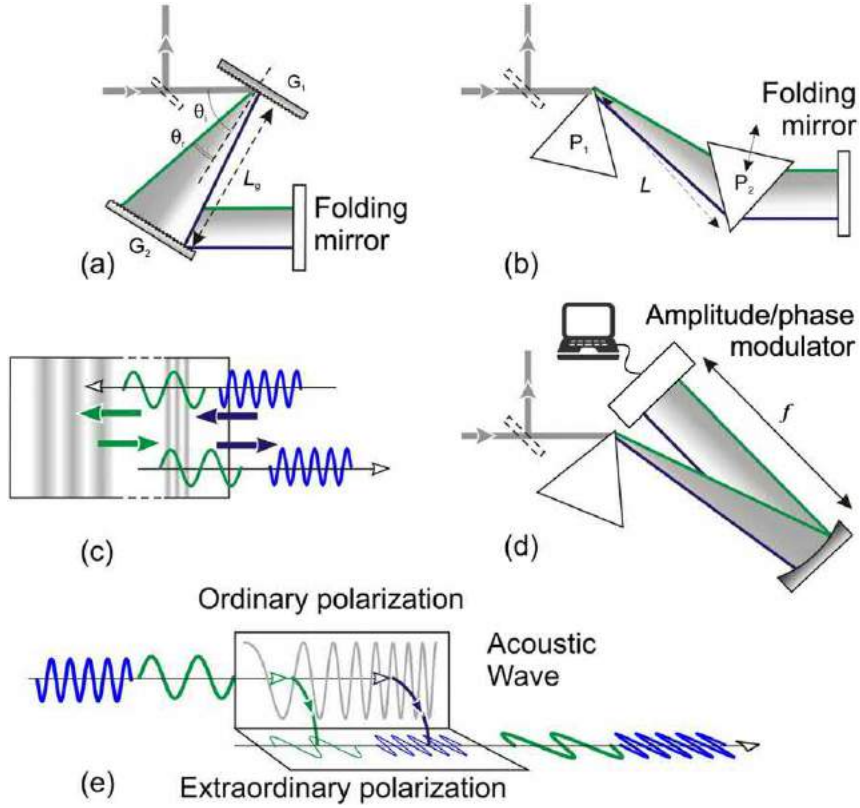

Fig. 5 Pulse-compression schemes. (a) Grating pair; (b) prism pair; (c) chirped dielectric mirror; (d) amplitude and phase-modulator in the Fourier plane of a 4*f* pulse shaper; and (e) acousto-optic programmable dispersive filter (AOPDF).

for a given prisms distance and glass insertion, it can be adjusted to introduce a negative GDD which compensates for the GDD of the OPA pulses, however it simultaneously introduces a large negative TOD, which cannot be independently controlled. Since both GDD and TOD are proportional to the prism tips separation, the ratio TOD/GDD is a characteristic of the prisms material and the wavelength, and can be minimized by choosing materials with low dispersion (such as fused silica, MgF_2, or CaF_2). In this case, however, the prism distance required to achieve a given value of GDD increases. Since it can only correct for GDD but not also for TOD, a simple prism compressor is suitable for TL pulse durations down to ≈ 20 fs, but not for broader bandwidths. An additional disadvantage of the prism pair is the fact that the prism distance needs to be changed as the OPA wavelength is tuned. A better solution consists in multiple bounces on a pair of CMs (panel (c)), which can be designed to correct simultaneously for both GDD and TOD and ensure either full compression of the ultra-broadband spectrum generated by an OPA (see Section Ultra-Broadband OPAs) or pulse compression throughout the OPA tuning range. CMs have the advantage of high throughput and a particularly robust setup, introducing a dispersion that is essentially insensitive to misalignment and depends only on the number of bounces. CMs introduce a dispersion that can only be varied in discrete steps (e.g., GDD ≈ -50 fs^2 in the visible range), so that fine dispersion tuning requires the insertion of some variable amount of material on the beam path, typically a pair of fused silica wedges. Finally, pulse compression can be achieved by adaptive optical systems, which employ pulse shapers to provide an arbitrary control of the frequency-dependent spectral phase. As shown in **Fig. 5(d)**, pulse shapers typically use the so-called 4-*f* arrangement, in which, after a dispersive element (a grating or a prism), a lens (or a spherical mirror) performs a spatial Fourier transform which converts the angular dispersion to a spatial separation at the back focal plane, where a phase (and/or intensity) modulator is located. Different phase modulators, based on liquid-crystals, acousto-optic modulators or deformable mirrors, can be employed. Alternatively, phase modulation can be achieved by an acousto-optic programmable dispersive filter ((AOPDF), panel (e)), in which the light wave interacts with a collinearly propagating acoustic wave.

Ultra-Broadband OPAs

Eq. (21) makes it clear that, in order to obtain broad phase matching bandwidths, one must achieve, for a given signal frequency $\overline{\omega_1}$, $\delta_{12}=0$, i.e., group velocity matching between signal and idler pulses. It will then become possible to amplify a broad bandwidth centered around $\overline{\omega_1}$. In an OPA using a collinear interaction geometry, the propagation direction inside the nonlinear crystal is selected to satisfy the phase-matching condition ($\Delta k=0$) for a given signal wavelength. In this configuration the signal and idler group velocities are in general not matched. Group velocity matching can be obtained in a Type I (or Type 0 in a periodically poled crystal) degenerate configuration, in which signal and idler have the same frequency ($\omega_1=\omega_2=\omega_3/2$) and the same polarization.

If the signal wavelength is tuned away from degeneracy, then the $\delta_{12}=0$ condition is generally not fulfilled in a collinear configuration, leading to narrow phase matching bandwidths. An additional degree of freedom can be introduced using a noncollinear geometry (Brida *et al.*, 2010; Gale *et al.*, 1995), in which pump and signal wave-vectors form an angle α (independent of signal wavelength) and the idler is emitted at an angle Ω with respect to the signal. In this case the phase matching condition is a vector equation, $\mathbf{k}_3=\mathbf{k}_1+\mathbf{k}_2$ that, when projected on directions parallel and perpendicular to the signal wave-vector, becomes

$$k_3\cos\alpha = k_1 + k_2\cos\Omega \tag{23a}$$

$$k_3\sin\alpha = k_2\sin\Omega \tag{23b}$$

Note that the angle Ω is not fixed, but it depends on the idler frequency according to Eq. (23b). If the signal frequency increases by $\Delta\omega$, the idler frequency decreases by $\Delta\omega$ and the wave-vector mismatches along the two directions parallel and perpendicular to the signal wave vector can be approximated, to the first order, as

$$\Delta k_{\mathrm{par}} \cong -\frac{\partial k_1}{\partial\omega_1}\Delta\omega + \frac{\partial k_2}{\partial\omega_2}\cos\Omega\,\Delta\omega - k_2\sin\Omega\frac{\partial\Omega}{\partial\omega_2}\Delta\omega \tag{24a}$$

$$\Delta k_{\mathrm{perp}} \cong \frac{\partial k_2}{\partial\omega_2}\sin\Omega\,\Delta\omega + k_2\cos\Omega\frac{\partial\Omega}{\partial\omega_2}\Delta\omega \tag{24b}$$

To achieve broadband phase matching, both Δk_{par} and Δk_{perp} must vanish. Upon multiplying Eq. (24a) by $\cos\Omega$ and Eq. (24b) by $\sin\Omega$ and adding the results, we get

$$\frac{\partial k_2}{\partial\omega_2} - \cos\Omega\frac{\partial k_1}{\partial\omega_1} = 0 \tag{25}$$

which is equivalent to

$$v_{g1} = v_{g2}\cos\Omega \tag{26}$$

Eq. (24) lends itself to a very simple geometrical interpretation (Riedle *et al.*, 2000): in a noncollinear configuration, broadband phase matching can be achieved for a signal-idler angle Ω such that the signal group velocity equals the projection of the idler group velocity along the signal direction. The so-called noncollinear optical parametric amplifier (NOPA), is a widely used device for the generation of few-optical-cycle pulses in the visible range. In a practical situation the pump-signal angle α is determined by the propagation direction of the seed beam, while the signal-idler angle Ω adjusts itself, according to Eq. (23b), to satisfy the

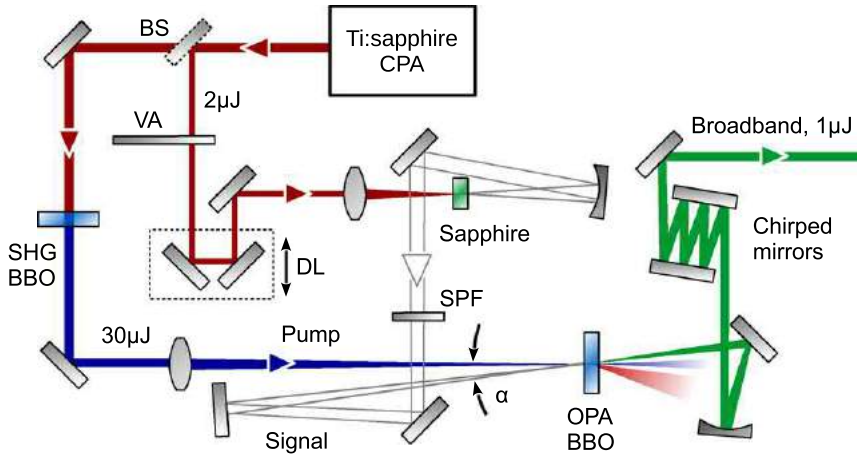

Fig. 6 Scheme of a β-barium borate (BBO)-based noncollinear optical parametric amplifier (NOPA) pumped at 400 nm. BS, beam splitter; CPA, chirped pulse amplification; DL, delay line; SHG, second harmonic generation; SPF, shortpass filter; VA, variable attenuator.

phase-matching condition. For this reason, the idler is emitted at a different angle for each wavelength, i.e., it is angularly dispersed and not easily usable.

In the following we will describe a typical visible NOPA design (Zavelani-Rossi *et al.*, 2001), the schematic of which is shown in **Fig. 6**. The system is pumped by an amplified Ti:sapphire laser generating 100-fs, 800-nm pulses at 1 kHz with energy up to 500 μJ. The energy is sufficient for simultaneously pumping several independent NOPAs. A fraction of the beam is used to generate the pump pulses at 400 nm by SHG in a 1-mm-thick BBO crystal; pulse energies up to 30 μJ are used to pump the first stage. Another small fraction of the beam, with energy of approximately 2 μJ, is focused into a 1-mm-thick sapphire plate to generate the single-filament WLC seed. The chirp of the visible portion of the white light is small and fairly linear with frequency. To avoid the introduction of additional chirp, only reflective optics are employed to guide the WLC to the amplification stage. Parametric gain is achieved in a 1-mm-thick BBO crystal, cut for type I phase matching ($\theta = 32$ degree), using a single-pass configuration to increase the gain bandwidth. The WLC seed is imaged by a spherical mirror in the BBO crystal, with a 100-μm spot size matching that of the pump beam. A thin short-pass filter removes the strong residual 800-nm component from the WLC, preventing its parasitic amplification.

When the BBO crystal is illuminated by the pump pulse and aligned perpendicularly to the pump beam, it emits a strong off-axis PSF in the visible in the form of a cone with apex angle of ≈ 6.2 degree (corresponding to an angle of 3.82 degree inside the crystal); this is the direction for which the group velocities of signal and idler are matched and therefore the gain bandwidth is maximized. The visible cone gives a visual aid to the identification of the optimum condition for broadband generation, which is found when the pump-signal angle matches the cone apex angle. In this condition, for optimum pump-seed delay, an ultrabroad gain bandwidth that extends over most of the visible (500–750 nm) is observed. The amplified pulses from a single stage have energy of approximately 1–2 μJ; much higher energies, up to ≈ 300 μJ, can be obtained by a second amplification stage. After the gain stage the amplified pulses are collimated by a spherical mirror and sent to the compressor.

Several compressor schemes have been implemented for the visible NOPA. Simple prism pairs can correct the second but not third-order dispersion and thus can only compress the pulses down to 10–15 fs; sub-10-fs pulses can be achieved by using either prism-grating or prism-CMs combinations, as well as adaptive compressors based on deformable mirrors. It is also possible to use exclusively CMs (Zavelani-Rossi *et al.*, 2001), greatly simplifying the system design, allowing for compactness, insensitivity to misalignment and high day-to-day reproducibility, which are of great importance in practical applications.

Optical Parametric CPA

In a broadband OPA both the pump and the seed are typically femtosecond pulses generated by a Ti:sapphire or Yb laser system. It is generally difficult and expensive to scale the energy of femtosecond pump lasers; on the other hand, it is easier to generate energetic picosecond pulses (5–50 ps duration) exploiting well-established gain media such as Nd: or Yb:doped crystals. With a long pump pulse, in order to achieve efficient energy extraction, it is necessary to temporally overlap the seed pulse with the pump by first stretching the seed pulse and then, after the amplification step, recompressing it back to nearly TL duration. This scheme, which is very similar to the chirped pulse amplification (CPA) (Backus *et al.*, 1998) occurring in a real gain medium, is known as OPCPA (Dubietis *et al.*, 1992) and is considered as the most promising route for energy scaling of few-optical-cycle pulses. A typical scheme of an OPCPA is shown in **Fig. 7**. The key technical hurdle in OPCPAs is the synchronization of the pump and seed pulses; it can be achieved either electronically, using a suitable circuitry, or optically, by using a part of the seed pulse spectrum to injection seed the pump laser. The optical approach allows a much lower pump-seed timing jitter but it is more challenging, since pump and seed are at different frequencies.

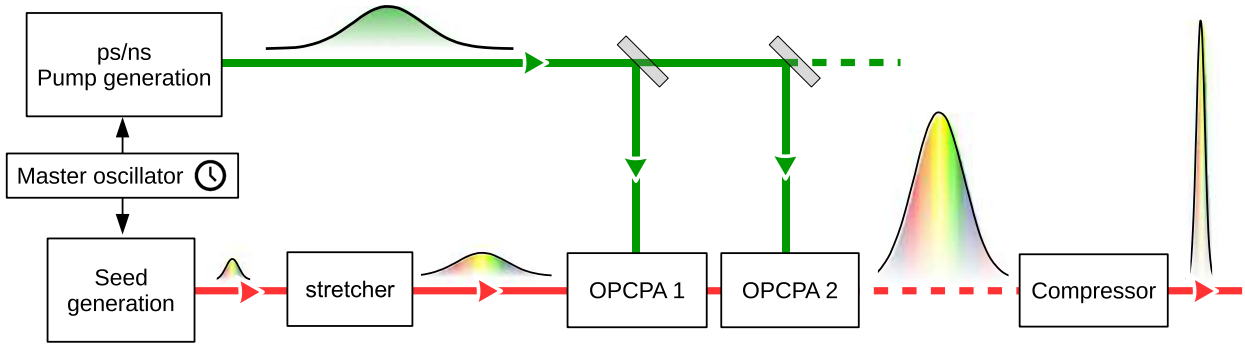

Fig. 7 General scheme of a multistage optical parametric chirped pulse amplifier (OPCPA). In this scheme the timing between the seed and the pump sources is optically provided by a master oscillator.

The OPA/OPCPA approaches offer some distinct advantages with respect to CPA for the generation of high peak power few-optical-cycle pulses:

1. The OPA/OPCPA has the capability of providing a high gain in a relatively short path, allowing a compact, tabletop amplifier setup, minimizing the linear and nonlinear phase distortions and ensuring an excellent temporal and spatial quality of the pulses.
2. With suitable selection of the phase-matching conditions (see Section Ultra-Broadband OPAs), the OPA process can provide gain bandwidths well in excess of those achievable with conventional amplifiers and can sustain pulse spectra corresponding to a TL duration of just a few cycles of the carrier frequency.
3. In an OPCPA, amplification occurs only during the pump pulse, so that amplified spontaneous emission and the consequent pre-pulse pedestal are greatly reduced.
4. Since the OPA/OPCPA is an instantaneous process of energy exchange between the interacting beams, no fraction of the pump photon energy is deposited in the medium; thus, if parasitic pump absorption can be neglected, thermal loading effects are absent, greatly reducing spatial aberrations of the beams and allowing to achieve much higher repetition rates avoiding the heat removal problems that limit the frequency scaling of CPA systems.

In addition to these clear advantages, the OPA/OPCPA concept also has some drawbacks:

1. an OPA/OPCPA requires pump pulses with short duration (a few picoseconds to tens of picoseconds), which are technologically challenging to generate, especially at high energy and/or high repetition rate; on the other hand the CPA, thanks to energy storage in the excited state, can use much longer nanosecond pump pulses (or even continuous wave pumping for materials with long excited state lifetime like Yb-doped crystals);
2. an OPA/OPCPA is very sensitive to the spatial and temporal quality of the pump, again putting a severe technological constraint on the pump laser.

Conclusions

Continuous progress in the development of light sources based on nonlinear optical effects for the generation of tunable ultrashort pulses ranging from the near-infrared throughout all the visible range provides very efficient tools to be exploited in time-resolved spectroscopic techniques and high field physics. Second-order nonlinear optical processes like OPA have demonstrated the capability of generating light pulses with durations down to few electric field cycles. Basic concepts and implementation schemes have been described, offering a comprehensive state of art tour in this rapidly evolving research field. Tunable few-cycle pulses constitute unprecedented tools for investing ultrafast processes in many sectors related to material science and biology. Pump-probe techniques with broadband pulses and more recently two-dimensional spectroscopic techniques allow to follow in great detail complex dynamical processes.

References

Backus, S., Durfee III, C.G., Murnane, M.M., Kapteyn, H.C., 1998. High power ultrafast lasers. Review of Scientific Instruments 69, 1207–1223.
Baumgartner, R.A., Byer, R., 1979. Optical parametric amplification. IEEE Journal of Quantum Electronics 15, 432–444.
Boyd, R.W., 2003. Nonlinear Optics. New York, NY: Academic Press.
Bradler, M., Baum, P., Riedle, E., 2009. Femtosecond continuum generation in bulk laser host materials with sub-μJ pump pulses. Applied Physics B 97, 561–574.
Brida, D., Manzoni, C., Cirmi, G., *et al.*, 2010. Few-optical-cycle pulses tunable from the visible to the mid-infrared by optical parametric amplifiers. Journal of Optics 12, 013001.

Cerullo, G., De Silvestri, S., 2003. Ultrafast optical parametric amplifiers. Review of Scientific Instruments 74, 1–8.
Dmitriev, V.G., Gurzadyan, G.G., Nikogosyan, D.N., Lotsch, H.K.V., 1999. Handbook of Nonlinear Optical Crystals. Springer Series in Optical Sciences, vol. 64. Berlin: Springer.
Dubietis, A., Jonušauskas, G., Piskarskas, A., 1992. Powerful femtosecond pulse generation by chirped and stretched pulse parametric amplification in BBO crystal. Optics Communications 88, 437–440.
Gaeta, A.L., 2000. Catastrophic collapse of ultrashort pulses. Physical Review Letters 84, 3582–3585.
Gale, G.M., Cavallari, M., Driscoll, T.J., Hache, F., 1995. Sub-20 fs tunable pulses in the visible from an 82 MHz optical parametric oscillator. Optics Letters 20, 1562–1564.
Giordmaine, J.A., Miller, R.C., 1965. Tunable coherent parametric oscillation in $LiNO_3$ at optical frequencies. Physical Review Letters 14, 973–976.
Harris, S.E., Oshman, M.K., Byer, R.L., 1967. Observation of tunable optical parametric fluorescence. Physical Review Letters 18, 732–734.
Hum, D.S., Fejer, M.M., 2007. Quasi-phasematching. C R Physique 8, 180–198.
Nikogosyan, D.N., 1991. Beta barium borate (BBO). Applied Physics A 52, 359–368.
Riedle, E., Beutter, M., Lochbrunner, S., *et al.*, 2000. Generation of 10 to 50 fs pulses tunable through all of the visible and the NIR. Applied Physics B 71, 457–465.
Zavelani-Rossi, M., Cerullo, G., De Silvestri, S., *et al.*, 2001. Pulse compression over a 170 THz bandwidth in the visible by use of only chirped mirrors. Optics Letters 26, 1155–1157.

Mode-Locked Lasers

Ladan Arissian and Jean-Claude Diels, University of New Mexico, Albuquerque, NM, United States

1 Introduction on Frequency Combs

A mode-locked laser is more than the typical source of ultrashort pulse; it is the source of the most accurate ruler in the frequency domain. It combines the properties of a narrow line single mode laser, with a femtosecond time resolution. The sketches of Fig. 1 summarize the properties of the output pulse train of a mode-locked laser, both in time and frequency domain. A simple description in time is that of a perfectly monochromatic wave, being sampled at regular time intervals τ_{rt}. There is in general no relation between the period of the optical wave $1/\nu$ and τ_{rt}. The phase of the harmonic wave at the "center" or "peak" of the temporal gate (pulse) is called the Carrier to Envelope Phase (CEP) (Fig. 1(a)). The difference between the CEP of two successive pulses $\Delta\phi_1$ and $\Delta\phi_2$ is constant over the (ideal) pulse train. The ratio $(\Delta\phi_2 - \Delta\phi_1)/(2\pi\tau_{rt}) = f_0$ is a frequency, called the carrier to frequency offset (CEO).

In the frequency domain, for an infinite pulse train, the laser spectrum consists of δ-functions equally spaced by $1/\tau_{rt}$, over a range of the order of the inverse of the gate (pulse) width (Fig. 1(b)). Extending this pulse train to close to zero frequency, one can show that f_0 corresponds to the first tooth of the extended comb [1].

While the CEO is a well defined measurable quantity, there is some fuzziness in the definition given for the CEP. Indeed, for a very long pulse compared to the carrier period $1/\nu$, it is at best difficult to pinpoint the "center" or "peak" of the pulse envelope as compared to a crest of the wave (Fig. 1(c)). For a very short pulse as in (Fig. 1(d)), in the frequency domain it is difficult to pinpoint which tooth of the comb is closest to the "center" or "peak" of the pulse spectrum. Clearly an alternate definition is desirable for the CEP. Techniques exist to measure the real electric field of a pulse in amplitude and phase, without a decomposition in "carrier" and "envelope". A complex representation of the field is obtained by taking the Fourier transform of this real measured field,

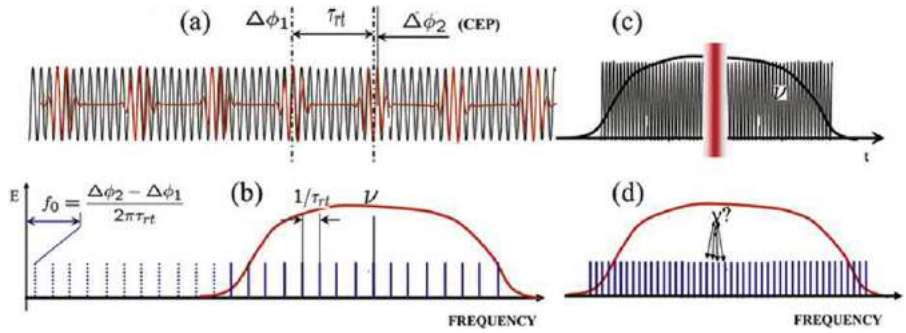

Fig. 1 (a) A set of pulses is periodically (period τ_{rt}) extracted from a cw wave. Each pulse envelope peaks for a particular phase of the cw wave, called Carrier to Envelope Phase (CEP). The ratio of CEP to repetition period τ_{rt} is the carrier to envelope offset (CEO). (b) The Fourier transform of (a) is a set of δ-functions equally spaced. The CEO is the frequency of the tooth closest to zero, in an extension of the comb. (c) For a long pulse, the exact location (within a fraction of the light period) of the center or peak of the envelope is difficult to define. (d) for an ultrashort pulse, the mode corresponding to the peak of the spectrum is equally difficult to define.

 doi:10.1016/B978-0-12-803581-8.09633-8

eliminating the negative frequency parts, before taking the inverse Fourier transform, which is now the "complex" electric field. A new definition of the CEP, independent of a decomposition in "carrier" and "envelope" of the optical pulses, is the difference between the value of the phase of the complex electric field at the peak of its amplitude, and the value of the phase taken at the peak of the real field [2,3].

It is often believed that the notions of CEP and CEO apply only to few cycle pulses. This is clearly disproved by experiments of intracavity phase interferometry performed with ps pulses [4]. In these experiments to be detailed in Section 4, the variation in CEO is used to perform phase detection with sensitivity better than 1 part in 10^8.

2 The Mode-Locked Laser as a Frequency Comb Generator

2.1 Typical Laser Configurations

The typical laser consists in a linear (Fabry-Perot) or ring resonator in which a broadband gain medium is inserted. In mode-locked operation, one or more pulses are circulating in the cavity, generating a train of pulses through an output mirror, spaced by the cavity round-trip time. One essential element required to create a train of identical ultrashort pulses at the output is a phase modulator inducing a frequency modulation or chirp on the circulating pulse. In the most common Ti:sapphire laser, the gain crystal modulates the phase through its nonlinear index $n_2 I$ (I being the pulse intensity in the crystal). This creates a positive frequency sweep or chirp across the circulating pulse. If a saturable absorber is used with a resonance centered at an optical frequency above that of the circulating pulse, the saturation of the resonant dispersion associated with the homogenously broadened line results in a downchirp [1]. Another essential element is a transparent optical system in which the light velocity is dependent on the optical frequency. This dispersive element can consist in a pair of prisms, as in **Fig. 2**, which can be tuned from a positive to a negative dispersion, dependent on the amount of glass traversed by the beam and the distance between prisms. Exact expressions for the dispersion associated with a pair of prisms can be found in reference [5]. Often the pair of prisms is replaced by mirrors of which the coating has been engineered to have dispersion (frequency dependent phase shift) of the desired sign and shape. The ratio of phase modulation to dispersion is a key element in having a mode-locked laser produce a perfect frequency comb, as will be detailed in Section 2.3.

In the case of a linear cavity (**Fig. 2(a)**), the laser consists in a gain section, a dispersive section (a pair of prisms in the example of the figure) and end-cavity mirrors. The various control knobs available to fix the parameters of the frequency comb (center of pulse spectrum, pulse duration, CEO and repetition rate) are detailed in Section 3. A slit, aperture or razor blade between the prism sequence and the end mirror determines the position of the spectral envelope of the individual pulse. The pulse duration is tuned by acting on the dispersion, translating one of the prism of the sequence, thus changing the amount of glass. The repetition rate can be adjusted independently of other parameters by tilting the end mirror after the prism sequence. The pump power provides a fine adjustment.

Cavity length adjustment with a piezo-element on one end mirror provides another tuning parameter for the comb CEO. The exact transfer function of all these controls can be found in reference [5]. While manipulation of dispersion and phase modulation is essential in the generation of a stable frequency comb, mode-locking requires some amplitude modulation to bring down the duration of noise fluctuation to the picosecond range before dispersive compression kicks in. These initial pulses can be provided by saturable absorbers, for instance multiple quantum wells used as end mirror of a linear cavity (**Fig. 2(a)**). The saturable absorber can have the additional function of locking the repetition rate (group velocity control) in the case of mode-locked lasers producing dual mode combs. It is then located in the middle of a linear cavity, or at 1/4 perimeter of the gain medium in a ring cavity (**Fig. 2 (b)**). This type of mode-locked operation with two pulses circulating in the cavity is dealt with in Section 4.

2.2 Mode-Locking as Orthodoncy

The resonance condition of any cavity is that the total phase shift $\sum k_i l_i$ experienced by the light in a round-trip time be a multiple of 2π. The sum is taken over all elements i of the cavity of index of refraction n_i and length l_i. An average index n_{av} can be defined

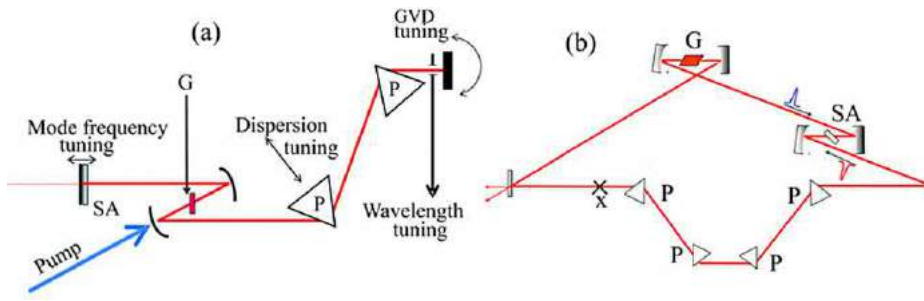

Fig. 2 Typical mode-locked laser configurations. SA: saturable absorber, P dispersive (Brewster angle) prism; G: gain medium (a) The linear laser where the different comb controls are indicated. (b) Typical bidirectional ring laser configuration, where two pulses circulate in opposite direction crossing in the saturable absorber (SA) and in position *X*.

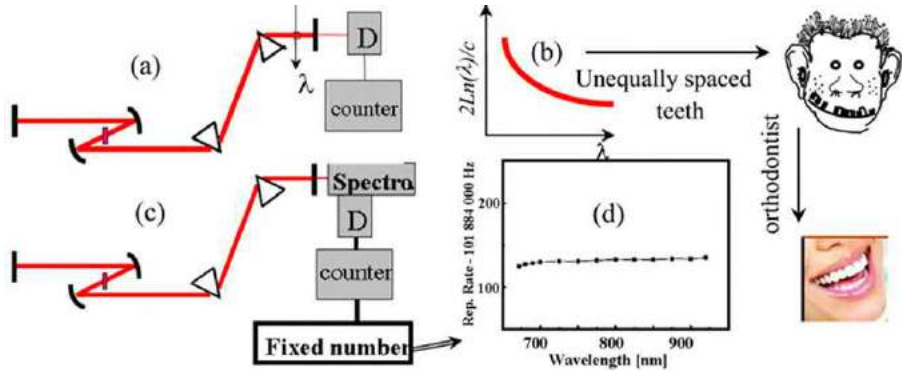

Fig. 3 (a) The aperture (or slit) at the output end of the laser is closed so as to let only a few modes oscillate [6,7]. A detector D records the mode beating, at the round-trip frequency of the cavity. That round-trip frequency, recorded as a function of the wavelength λ by a frequency counter, shows the variation of mode spacing across the gain spectrum (b). (c) The aperture is open, allowing mode-locked operation of the laser. A portion of the spectrum of the frequency comb is selected by a monochromator. The repetition rate is recorded by the same frequency counter as function of the wavelength (d), showing no variation across the spectrum. Locking the modes of the laser is the analogue of an orthodontist intervention on the comb.

by $\sum k_i l_i = k_{av}L = (2\pi/\lambda)n_{av}L = [\omega n_{av}/c]L$ where L is the cavity perimeter. The index of refraction n_{av} generally varies significantly over the frequency range corresponding to the bandwidth of a typical fs pulse. This implies that the modes spacing $n_{av}L/c$ varies across the bandwidth of the pulse. In the classical textbook describing mode-locking, it is said that the short pulse is the result of all the modes of the cavity emitting simultaneously and in phase.

This textbook description does not correspond to reality: it can be shown that the emission by a set of non-equidistant modes in phase does not result in an uniform pulse train [1]. Despite the non-uniformity of the spacing of the modes of the laser cavity, a simple experiment as sketched in **Fig. 3** demonstrates the difference between the modes of the laser cavity and the teeth of the comb. Mode-locked operation is prevented by closing the aperture at the output end of the laser cavity (**Fig. 3(a)**). The wavelength of the laser can be tuned by translating the aperture in the plane of the figure. An RF spectrum analyzer (or frequency counter) can be used to record the variation of mode-beating frequency $1/\tau_{rt}$ versus wavelength [6,7]. This variation is a measure of the average group velocity of the intracavity pulse $v_{av}=2L/\tau_{rt}$ where L is the cavity length (**Fig. 3(b)**). Opening or eliminating the aperture results in mode-locked operation (**Fig. 3(c)**). The teeth spacing of the comb can be measured across the output spectrum by selecting a wavelength range of interest with a spectrometer, and recording the round-trip frequency with a spectrum analyzer or a frequency counter (**Fig. 3(d)**). The very small variation of this number across the spectrum (<10 Hz) over a 250 nm wavelength range reflects the thermal expansion of the laser cavity during the course of the experiment [8]. Measurements with stabilized lasers have shown the comb teeth spacing of the order of hundred of MHz to be constant within mHz over the full spectral width of the laser output [9].

2.3 Phase Modulation and Dispersion - the Soliton Operation

To a coarse approximation, the simplest model of steady state operation can be reduced to a cascade of phase modulation by Kerr nonlinearity followed by quadratic dispersion. As sketched in **Fig. 4**, a transparent medium of thickness d with a time dependent index of refraction $n(t)$ will phase modulate a pulse sent through it. In a plane wave approximation, we describe a pulse of optical frequency ω propagating in the z direction by an electric field $E(t,z)=\tilde{\mathscr{E}}(t,z)\exp[i(\omega t-kz)]$, where $\tilde{\mathscr{E}}(t,z)=\tilde{\mathscr{E}}(t,z)\exp(i\varphi(t,z))$ (complex quantities are written here with a tilde). After propagating through an element with a time dependent index of refraction, the pulse will be phase modulated or "chirped" and have the complex envelope $\tilde{\mathscr{E}}(t,d)=\tilde{\mathscr{E}}(t,0)\exp[-ik(t)d]$ where $k(t)=\omega n(t)/c=(2\pi/\lambda)n(t)$, c is the speed of light and λ the wavelength in vacuum. Most often, the time dependent index is due to the Kerr effect and is proportional to the light intensity $n(t)=n_2I(t)$, resulting in an instantaneous frequency generally increasing with time or "upchirp", $\varphi(t)=[(\partial/\partial t)k(t)]d$ as sketched to the right of **Fig. 4**. In this temporal phase modulation process, the envelope shape $\mathscr{E}(t, z)$ remains unchanged.

Dispersion can be understood as phase modulation in the frequency domain. The input to a transparent medium with a frequency dependent index of refraction $n(\Omega)$ is the Fourier transform of the field defined as:

$$\tilde{E}(\Omega)=\mathcal{F}\{E(t)\}=\int_{-\infty}^{\infty}E(t)e^{-i\Omega t}dt=|\tilde{E}(\Omega)|e^{i\Phi(\Omega)} \tag{1}$$

In the definition (1), $|\tilde{E}(\Omega)|$ denotes the spectral amplitude and $\Phi(\Omega)$ is the spectral phase. If we refer to the central frequency of the pulse, $\Delta\Omega=\Omega-\omega$, the transformation of the pulse spectral envelope by propagation of a distance z through the dispersive medium is:

$$\tilde{\mathscr{E}}(\Delta\Omega,z)=\tilde{\mathscr{E}}(\Delta\Omega,0)e^{-ik(\Delta\Omega,z)} \tag{2}$$

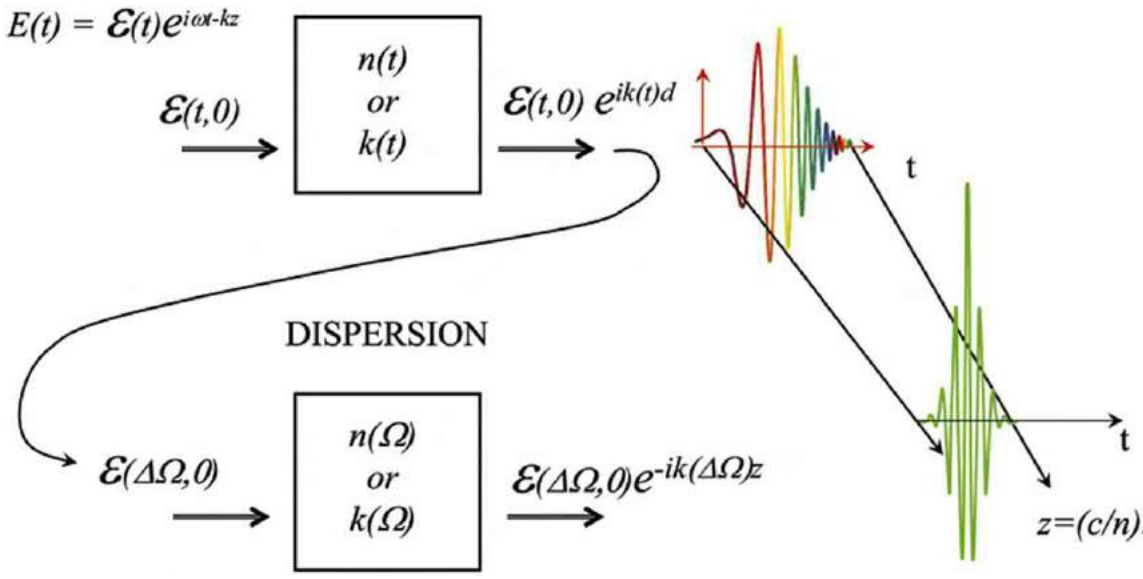

Fig. 4 Phase modulation and dispersion are the time-frequency correspondent of each other. In the time domain, phase modulation broadens the spectrum and introduce chirp (in general upchirp as sketched on the right) without affecting hte pulse temporal profile. In the frequency domain, dispersion leaves the spectrum of the pulse unchanged. In he case of negative dispersion applied to the chirped pulse of the figure, the faster higher frequency) leading edge catches up with the slower leading edge, leading to pulse compression.

In this spectral phase modulation process, the envelope shape $\tilde{\mathscr{E}}(\Delta\Omega, z)$ remains unchanged. If we expand the wavevector in series $k(\Delta\Omega) = k_0 + \Delta\Omega k' + \Delta\Omega^2 k''/2$ the first term k_0 corresponds to an irrelevant phase shift, the second term is simply the group delay, and the third term is the group velocity dispersion. As shown in the sketch on the right side of **Fig. 4**, if k'' is negative, the high frequency at the tail of the chirped pulse will propagate faster than the pulse front resulting in pulse compression. An exact balance is sought in mode-locked lasers between dispersion and Kerr phase modulation to achieve stable operation and minimum pulse duration.

Given the phase modulation and dispersion, one can derive a relation that predicts the values of pulse duration and intensity in steady state [3]. A question that remains is the nature of the mysterious orthodontist that transforms the unequal tooth spacing into a perfect comb with equally spaced teeth. The answer is in a frequency domain calculation similar to the one in time domain leading to solitons. In this calculation detailed in reference [3], the unequal mode spacing of the cavity is modeled by quadratic dispersion. The modification of the comb brought upon by self-phase modulation is calculated in the frequency domain. Imposing that the teeth of the comb are equally spaced leads to the condition between pulse duration τ, the cavity perimeter P ($=2L$ in the case of a linear cavity), the length of the Kerr medium ℓ, the nonlinear index n_2 the peak intensity I_0, the wavelength λ and the average dispersion of the cavity k''_{av}:

$$\tau^2 = \frac{\lambda |k''_{av}|}{2\pi n_2 I_0} \frac{P}{\ell} \tag{3}$$

It is remarkable this calculation leads to exactly the same prediction of pulse duration and intensity in steady state than the derivation of the soliton equation [3].

3 The Control Knobs for the Frequency Comb

While the CEO controls the frequency comb, it remains to be seen how this control is actuated, which depends on the specific goals to be achieved. If the aim is to create a frequency standard or an ultra-stable frequency comb, it is desirable to have coarse (usually slow) and fast (fine control) actuators on two parameters that define the comb — for instance the frequency of one mode, and the repetition rate.

Most of the controls easily accessible in a linear mode-locked laser are sketched in **Fig. 2(a)**. Translating a slit or aperture after the prism sequence [or between the prism sequence in the case of the ring laser of **Fig. 2(b)**] provides a control of the central wavelength of the pulses. Translating a prism orthogonally to its base controls the pulse duration, through tuning of the cavity dispersion. A slow control of the cavity length is provided typically by mounting the end cavity mirror on a piezo-element. A fast control of the repetition rate is often achieved by inserting a electro-optic modulator in the pump beam, using the repetition rate dependence on pump power. Tilting the end mirror after the two prism sequence can be used for coarse (slow) control of the repetition rate. It has been shown that this control — effected with piezo-elements — offers a control of the comb repetition rate

(or mode spacing), without affecting the position of the central mode [5]. The latter repetition rate control combined with an acousto-optic modulator provides a near perfect set of orthogonal controls. Indeed, an acousto-optic modulator positioned outside the laser cavity can shift the comb frequency without affecting the mode spacing.

Rather than offering orthogonal control, most of the cavity parameters modify simultaneously both parameters of the comb, as illustrated below in the case of a change in cavity length.

3.1 Cavity Length Change

The easiest parameter to change in a mode-locked laser is the optical path between end mirrors (or the perimeter in the case of a ring cavity). This operation can performed mechanically by putting an end mirror on a piezoelectric element in the case of a linear laser. In the case of a fiber laser, a portion of fiber will be wrapped around a piezoelectric cylinder.

There is no universal answer as to what is the result of a change in cavity length. On deduces easily from the sketch in **Fig. 1** that the frequency ν_N of a mode N of the comb is:

$$\nu_N = f_0 + N\frac{1}{\tau_{rt}} \tag{4}$$

where f_0 is the CEO and the cavity round-trip time is $\tau_{rt}=2L/v_{av}$. L is the length of the cavity (or 2L=P is the perimeter of a ring cavity), and v_{av} is the average velocity of the pulse circulating in the mode-locked laser cavity.

Let us imagine an adiabatic expansion of the resonator by an amount ΔL. We can expect that the number of modes N is unchanged. Assuming $f_0=0$ and is not affected by the cavity expansion (**Fig. 1(b)**), then the mode spacing (repetition rate) has to change. The accordion motion of the comb results in the following changes in mode frequency and repetition rate:

$$\Delta\nu_N = \nu_N \times \frac{\Delta L}{L} \tag{5}$$

$$\Delta\left(\frac{1}{\tau_{rt}}\right) = \frac{2\Delta L}{c} = \frac{\Delta P}{c} \tag{6}$$

where the second equality in Eq. (6) applies to a ring cavity. If we take as an example an elongation by a half wavelength $\lambda/2$ ($\lambda=800$ nm) of a $\tau_{rt}=10$ ns cavity, since $N=4\cdot10^6$, the change in central mode frequency is $\approx10^8$ Hz. The change in repetition rate is 50 Hz.

These considerations suggest that a change in intracavity optical path primarily translates the comb. The parameter that is modified by a particular control knob is not always obvious, and may depend on the way the control is applied. A different result is predicted in reference [10], where it is proposed to insert an electro-optic modulator (EOM) in a mode-locked cavity, regeneratively driven at the cavity repetition rate or a multiple thereof. The DC bias applied to the EOM would be used for fast control of the cavity length, or optical frequency of the comb. The intracavity pulse traversing the modulator experiences a positive, negative, or zero frequency shift to the pulse, depending on which part of the modulation waveform the pulse encountered. In soliton mode-locking, the cavity has negative dispersion $dk/d\Omega|_\omega<0$. In the modulator of length ℓ, a positive frequency shift $\Delta\Omega$ will cause a negative group delay τ_d, implying an increase of repetition rate. In this scenario, the phase of the wave applied to the EOM controls the repetition rate of the mode spacing of the comb, while the DC bias controls the mode position. A single actuator would thus be sufficient to stabilize the CEO f_0 and the teeth spacing of the comb.

3.2 Repetition Rate and Group Velocity

The mode spacing of the frequency comb — or repetition rate of the pulse train — is set by the average group velocity of the pulse circulating in the laser. It is generally taken for granted [11] that the group velocity v_g of a pulse circulating in a laser cavity is determined by the derivative of the k vector with respect to angular frequency:

$$v_g = \frac{1}{\frac{dk}{d\Omega}|_\omega} \tag{7}$$

where ω is the pulse carrier frequency. Contrary to the association of pulse envelope velocity and group velocity defined in Eq. (7), it is demonstrated in Ref. [12], that the envelope velocity of a pulse circulating in a mode-locked laser is generally not related to $dk/d\Omega$, for a k vector averaged in the cavity, but to the gain and loss dynamics inside the laser. For instance, it is well known that the pulse envelope propagates at superluminal velocities in a saturable gain medium [13]. This effect contributes to the fine tuning of the group velocity by modulation of the pump power in a Ti:sapphire laser. In a saturable absorber, the envelope velocity will be decreased by elimination of the pulse leading edge in favor of the trailing edge. In mode-locked fiber lasers, the average envelope velocity is a combination of group velocity, modal velocity, gain velocity and absorber velocity.

3.3 Comb Control with an Intracavity Etalon

Intracavity Fabry-Perot have been used for decades to narrow down the laser spectrum of cw and pulsed lasers [14,15]. It is less known that they are an effective tool to control the parameters of a mode-locked laser. A uncoated intracavity glass etalon will tune

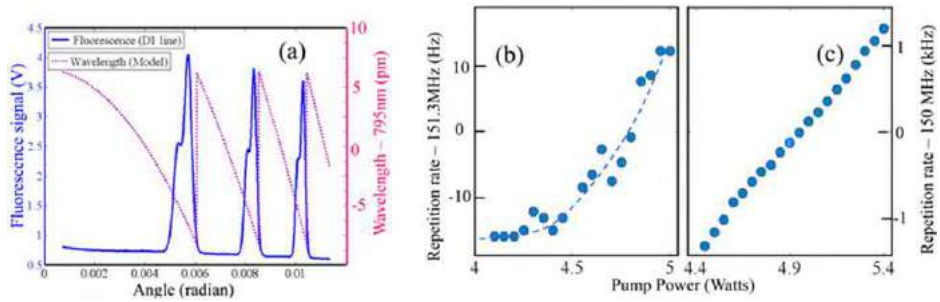

Fig. 5 Tuning a frequency comb by insertion of a fused silica etalon in a linear mode-locked laser cavity. (a) Solid line: fluorescence of ^{87}Rb irradiated by the frequency comb, as a function of the (internal) tilt angle of the etalon. The zero corresponds to normal incidence. Dashed line: calculated resonance wavelength of the etalon. The right ordinate indicates the difference (in pm) between the calculated wavelength and 795 nm. (b) Tuning of the repetition rate (mode-spacing) of the comb, as the pump power of the laser is ramped. The full tuning range is less than 30 Hz. (c) Tuning curve of the repetition rate (mode-spacing) after insertion of the etalon. The full tuning range is now 2.45 kHz over a pump power variaton of 1 W.

not only the mode spacing, but also the pulse repetition rate [16]. Unlike the action of the slit in **Fig. 2** or a birefringent filter, the etalon does not act on the average pulse frequency, but on the individual modes. As a demonstration, a 15.119 mm thickness uncoated fused-silica etalon is inserted in the linear cavity of **Fig. 2(a)**. The output of the laser is send to a cell containing ^{87}Rb, and the fluorescence of the vapor is monitored as the laser is tuned to the D1 line centered at 794.978851 nm. As the etalon is tilted away from the normal to the beam, the fluorescence of rubidium crosses successive resonances (**Fig. 5(a)**). The dotted line shows the wavelength corresponding to the etalon resonance condition. The 814.52 MHz hyperfine splitting of the upper state of the D1 transition is resolved in the structures of **Fig. 5(a)**. Such a measurement indicates that, despite the low finesse of the etalon, precise mode-tuning is achieved. This property can be explained by the fact that the modes of the etalon are locked to those of the comb. **Fig. 5(b)** is a measurement of the change in repetition rate, in the absence of intracavity etalon, as the pump power is changed. The tuning range of the $\approx$ 150 MHz repetition rate is only 30 Hz for a pump power change of 1 Watt. If the fused silica etalon is inserted in the cavity, the tuning range increases to 2.45 kW, over the same 1 W pump power increase, as shown in **Fig. 5(c)**. The insertion of an etalon inside the mode-locked cavity allows fine tuning over a large dynamic range of both the mode position (through the etalon angle) and the mode spacing (through the pump power). This property applies as well to linear lasers [16] as to ring lasers [12].

3.4 Locking the Repetition Rate

In most applications related to stabilization of frequency combs, the aim is to keep the CEO under tight control. There are however some applications where it is the CEO itself that is the free parameter being measured, while other characteristics of the comb (pulse duration, repetition rate, power) are to be kept constant. One application is intracavity phase interferometry, discussed in Section 4. In that technique, the mode locked laser is generating two correlated frequency combs, corresponding each to a different pulse circulating in the cavity. If the two combs have the same mode spacing, it is possible to interfere them. The interference between the two combs gives a signal at a frequency that is the difference between the two CEO f_{01} and f_{02} of each comb. Indeed, taking the difference of the frequency of corresponding modes (index N) of the comb:

$$\Delta\nu = \left(f_{02} + N\frac{1}{\tau_{rt,2}}\right) - \left(f_{01} + N\frac{1}{\tau_{rt,1}}\right) = f_{02} - f_{01} \tag{8}$$

since $\tau_{rt,2} = \tau_{rt,2}$ for the mode spacings to be equal. Note that the visibility of the beating between the two combs can be near 100%, since all pairs of modes contribute thanks to the property of the combs to have equal tooth spacing. The challenge to create this double comb is to lock the round-trip time of either pulse to the same value, and force the pulses to meet at the same location at each round-trip, *without perturbing* any other parameter of the comb. One obvious approach is to have both circulating pulses in the cavity created by gain switching, as is the case in optical parametric oscillators [17,18]. Unlike in inversion based lasers, in optical parametric oscillators the gain for a signal of intensity I_s at frequency ω_s is proportional to the product $I_s I_p$. It is therefore limited in duration to that of the pump pulse. In fs pulse synchronously pumped optical parametric oscillators, the pump laser is the clock that determines the repetition rate of either pulse that it generates in the optical parametric oscillators cavity.

Another method of locking the meeting point of two pulses in a mode-locked cavity is the saturable absorber. A saturable absorber (homogeneously broadened) is inserted either in a ring cavity, or in the middle of a linear cavity. Two pulses circulate in the cavity. The configuration of minimum losses is that where the two pulses meet in the absorber. It can indeed be shown that, for two pulses of equal intensity counter-propagating in the absorber, the effective saturation intensity is reduced by a factor 3, as compared to the saturation intensity for a single propagating pulse [1]. The mechanism by which the "colliding pulse" mode-locking is stable is explained in **Fig. 6**. Let us assume that the absorber, represented by the dot patterned rectangle, has moved to the right. The leading edge of pulse A will have no overlap with pulse B as it enters first the medium, and will be more attenuated

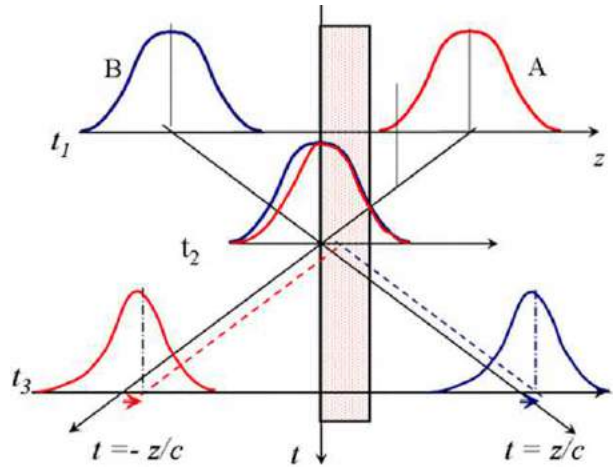

Fig. 6 Time sequence of two pulses *A* and *B* crossing in a saturable absorber. At the initial time t_1, pulses *A* is closer to the saturable absorber that *B*. At time t_2, the leading edge of A has passed the absorber, having been attenuated, while the leading edge of *B* overlapping with the trailing edge of *A* experiences less attenuation (mutual saturation).

than at the trailing edge (mutual saturation) where the two pulse overlap. Therefore, the center of gravity of pulse A will shift to the right, towards the absorber (dashed red line). The situation is opposite for pulse *B*, since pulse A will have left the absorber at the passage of the trailing edge of *B*. More attenuation of *B*'s trailing edge implies that its trajectory moves to the right (dashed blue line). The two dashed trajectories cross close to the center of the absorber, showing that, after successive round-trips, the pulse crossing point will be centered with the saturable absorber.

Saturable absorption is often the mechanism that sets the average group velocity in a discrete components laser, dominating the timing imposed by other elements. The situation can be different in other systems, such as fiber lasers. In contrast to observations in dye and solid state mode-locked lasers, bidirectional fiber lasers mode-locked by saturable absorber (single wall carbon nanotubes or CNT) produce pulses of different central wavelength, durations, and repetition rates [19]. This can be explained by the fact that fiber lasers cannot be adequately modeled by differential equations: the change in pulse characteristic per element is much larger than for discrete component lasers, which can result in a very large asymmetry in the evolution of the two pulses in the cavity.

4 Intracavity Phase Interferometry

Most applications of frequency combs involve fast and accurate stabilization electronics, to maintain the position of a tooth of a comb within a few Hz. Intracavity Phase Interferometry (IPI) is a technique that exploits the basic properties of mode-locking, to achieve phase detection better than 10^{-8} radian, without the need for stabilization. The IPI being referred to in this chapter is dealing with active (laser) resonators, not to be confused with "intracavity nonlinear spectroscopy" [20], in which actively stabilized ultra-high finesse passive Fabry-Perot cavities are used to transform a few ppm absorption into some % change in transmission. A broadband laser can be used as an amplitude sensor: the nonlinearity of laser close to threshold has been exploited to detect very small absorption [21]. More recently, it was realized that a mode-locked laser can be made an ultra-sensitive phase sensor [4]. A basic properties being exploited is that a laser is akin to a Fabry-Perot of infinite finesse. In addition, in a laser, a change in phase $\Delta\phi$ is converted into a change in frequency $\Delta\nu = \Delta\phi/(2\pi\tau_{rt})$, in contrast to passive interferometers where any change in phase is measured as a change in amplitude.

Implementation of IPI in the case of a ring laser is sketched in **Fig. 7**. Two pulses are counter-circulating in the ring cavity, with their average group velocity locked to the same value (using for instance a saturable absorber). At 1/4 cavity perimeter from their crossing (starting) point, one pulse passes through the gain medium, and the other has its phase modified by the sensor *S*. The latter is a generic term for the phase modification by the element to be measured, which could be an elongation, a change in index due to Faraday rotation, Kerr effect, air current, rotation etc... The laser outputs two frequency combs, which are recombined via a delay line. The interfering pulse trains recorded on a detector have a modulation at a frequency $\Delta\nu$ equal to the difference between the corresponding modes of order *m* of the comb. Since the combs have equal mode spacing over the spectrum, that difference is simply equal to the difference CEO frequency between the two combs: $\Delta\nu = f_{02} - f_{01}$. One would expect the bandwidth of the beat note $\Delta\nu$ to be equal to the sum of the bandwidths of the two CEOs. This is clearly incorrect: in an unstabilized laser such as considered in **Fig. 2**, the mode bandwidth is typical 1 MHz. But the measured bandwidth of the beat note $\Delta\nu$ is as small as 0.17 Hz [17]. The reason for the small noise in beat note is that the two output frequency combs are correlated. If the beat note is caused by a differential optical path ΔP seen by the two pulses, its frequency is simply the product of the light carrier frequency by the relative elongation: $\Delta\nu = \nu\Delta P/P$.

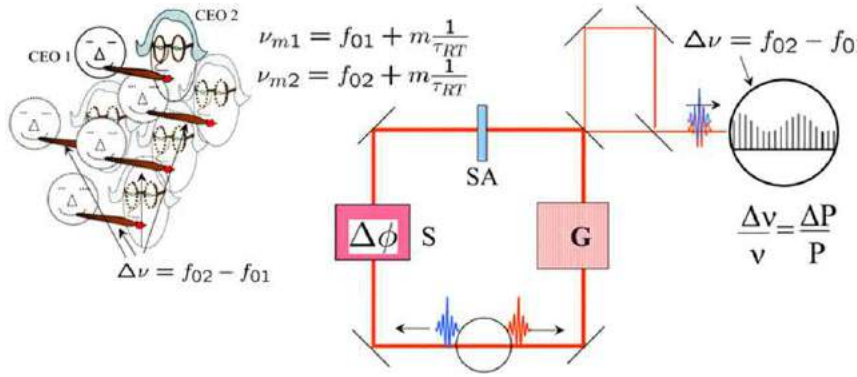

Fig. 7 Sketch of implementation of IPI with a ring cavity. Two pulses (red and blue) are circulating in opposite sense in the cavity, crossing in the circle at the bottom of the cavity, and in the saturable absorber SA. *G* is the gain medium, traversed alternatively at every half cavity round-trip by one of the pulses. S is the sample to be measured: an element that provides a differential phase shift proportional to the quantity to be measured. The two output pulse trains are overlapped through a delay line, and interfered on a detector. An oscilloscope will shows the pulse train modulated at the "beat note" frequency $\Delta\nu$ equal to the difference between the CEOs of the two pulse trains. The CEO frequencies of this unstabilized laser are strongly fluctuated. However, because the two trains are correlated, these fluctuations are not seen in the beat frequency $\Delta\nu$.

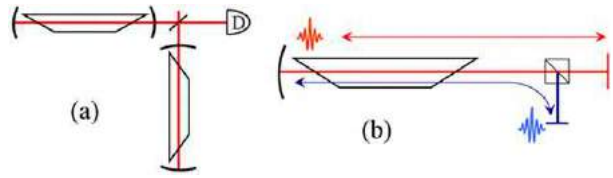

Fig. 8 (a) The active interferometer of Abramovici and Vager. Two gain section are inserted in each branch of a Michelson like interferometer. (b) IPI concept for a linear cavity [4]. The two pulses experience the same cavity and medium fluctuation with a half cavity round-trip time difference.

The small beat note sensitivity of IPI seems to contradict a theoretical proof by Abramovici and Vager [22] that active and passive cavity interferometers have comparable performances. The "active cavity interferometer" considered by Abramovici and Vager is as sketched in **Fig. 8(a)**, with gain media in each branch of a Michelson-like interferometer. Clearly, the noise of each gain section is uncorrelated. In the linear implementation of IPI, there is only one gain medium in which each pulse circulate alternatively at several ns interval. For longer times (compared to the round-trip time) the two pulse see exactly the same condition, which explains why the noise is correlated.

References

Diels, J.C., Rudolph, W., . Ultrashort laser pulse phenomena, second ed. Boston: Elsevier. ISBN 0-12-215492-4.
Diels, J.C., Luo, X., Xu, X., Masuda, K., Arissian, L., . Definitions and control of the CEP and CEO. Laser Physics 20, 1038–1043.
Arissian, L., Diels, J.C., . Investigation of carrier to envelope phase and repetition rate — fingerprints of mode-locked laser cavities. Journal of Physics B: At. Mol. Opt. Phys 42, 183001.
Arissian, L., Diels, J.C., . Intracavity phase interferometry: Frequency combsensors inside a laser cavity. Laser Photonics Rev 8, 799–826.
Arissian, L., Diels, J.C., . Carrier to envelope and dispersion control in a cavity with prism pairs. Physical Review A 75, 013814–013824.
Knox, W.H., . Femtosecond intracavity dispersion measurements. In: Martin, J.L., Migus, A., Mourou, G.A., Zewail, A.H. (Eds.), Ultrafast Phenomena VIII. Berlin: Springer, pp. 192–193.
Knox, W.H., . In situ measurement of complete intracavity dispersion in an operating Ti:Sapphire femtosecond laser. Optics Letters 17, 514–516.
Jones, R.J., Diels, J.C., Jasapara, J., Rudolph, W., . Stabilization of the frequency, phase, and repetition rate of an ultra-short pulse train to a Fabry-Perot reference cavity. Optics Communication 175, 409–418.
Udem, T., Reichert, J., Holzwarth, R., Hänsch, T., . Accurate measurement of large optical frequency differences with a mode-locked laser. Optics Letters 24, 881–883.
Hudson, D.D., Holman, K.W., Jones, R.J., Cundiff, S.T., Ye, J., . Mode-locked fiber laser frequency-controlled with an intracavity electro-optic modulator. Opt. Lett. 30, 2948–2950.
Yum, H.N., Salit, M., Yablon, J., Salit, K., Wang, Y., Shahriar, M.S., . Superluminal ring laser for hypersensitive sensing. Optics Express. 18, 17658.
Hendrie, J., Lenzner, M., Akhamiardakani, H., Diels, J.-C., Arissian, L., . Impact of resonant dispersion on the sensitivity of intracavity phase interferometry and laser gyros. Optics Express 24, 30402–304010.
Basov, N.G., Ambartsumyan, R.V., Zuev, V.S., Kryukov, P.G., Letokhov, V.S., . Nonlinear amplifications of light pulses. Soviet Physics JETP 23, 16–22.
Lecompte, C., Mainfray, G., Manus, C., Sanchez, F., . Laser temporal coherence effects on multiphoton ionization processes. Phys. Rev. A A-11, 1009.
Lompre, L.A., Mainfray, G., Manus, C., Thebault, J., . Multiphoton ionization of rare gases by a tunable-wavelength 30-psec laser pulse at 1.06 μm. Physical Review A 15, 1604–1612.
Masuda, K., Hendrie, J., Diels, J.C., Arissian, Ladan, . Envelope, group and phase velocities in a nested frequency comb. Journal of Physics B 49, 085402.

Velten, A., Schmitt-Sody, A., Diels, J.C., . Precise intracavity phase measurement in an optical parametric oscillator with two pulses per cavity round-trip. Optics Letters 35, 1181–1183.
Gowda, R., Nguyen, N., Diels, J.-C., Norwood, R., Peyghambarian, N., Kieu, K., . All-fiber bidirectional optical parametric oscillator for precision sensing. Optics Letters 40, 2033–2036.
Zeng, C., Liu, X., Yun, L., . Bidirectional fiber soliton laser mode-locked by single-wall carbon nanotubes. Optics Express 21, 18937–18942.
Ma, L.-S., Ye, J., Dube, P., Hall, J.L., . Ultrasensitive frequency-modulation spectroscopy enhanced by a high-finesse optical cavity. Journal of the Optical Society of America B 16, 2255.
Hänsch, T.W., Schawlow, A.L., Toschek, P.E., . Ultrasensitive response of a cw dye laser to selective extinction. IEEE Journal of Quantum Electronics 8, 802–808.
Abramovici, A., Vager, Z., . Comparison between active- and passive-cavity interferometers. Physical Review A 33, 3181–3184.

Few-Cycle and Attosecond Lasers

Francesca Calegari, Center for Free-Electron Laser Science, Hamburg, Germany; University of Hamburg, Hamburg, Germany; and Institute for Photonics and Nanotechnologies CNR-IFN, Milano, Italy
Caterina Vozzi, Institute for Photonics and Nanotechnologies CNR-IFN, Milano, Italy

Introduction

The tremendous advancements in the development of laser sources delivering few-optical-cycle pulses in the visible and near-infrared spectral ranges allowed to create even shorter light pulses, down to a few tens of attoseconds (1 as $=10^{-18}$ s), in the extreme ultraviolet (XUV) spectral region. This impressive progress in laser technology has given access to the electronic time-scale in matter (Calegari *et al.*, 2016). Attosecond pulses were first employed for the investigation of ultrafast electron dynamics in atomic systems, where for instance Auger decay, autoionization and photoemission delays have been measured in real time. In the past few years, attosecond science has successfully addressed physical process in more complex targets such as bio-relevant molecules and condensed matter.

In this article the main schemes for the generation and characterization of light pulses from few-femtosecond down to a few tens of attoseconds are described. The first part of the article introduces the femtosecond laser technology required for the generation of attosecond pulses. Section Hollow-Core Fiber Compression describes the hollow-core fiber approach as a post-compression stage for the generation of broadband laser pulses comprising only a few-optical cycles. A crucial requirement for the generation of isolated attosecond pulses with few-cycle pulses is a stable carrier-envelope-phase (CEP). Section Carrier-Envelope-Phase Stability provides a description of the most commonly used CEP stabilization techniques. The generation of tunable driving pulses in the infrared spectral region is briefly presented in Section OPA and OPCPA Technology, while the possibility to drive the process with single-cycle pulses is examined in Section Pulse Synthesis. An overview on the main techniques for the temporal characterization of such ultrashort driving pulses is presented in Section Ultrashort Pulse Characterization. The basis for the generation of attosecond pulses is then introduced, starting from the description of the high-order harmonic generation (HHG) process in Section High-Order Harmonic Generation. Techniques for gating HHG and isolate a single attosecond pulse from the attosecond pulse train are then presented in Section Gating Techniques. Section Attosecond Pulse Generation in the Water-Window Region discusses the possibility to extend the attosecond pulse generation process from the XUV to the soft-x energy region exploiting the driving laser sources described in Section OPA and OPCPA Technology. Finally a comprehensive description of all the temporal characterization methods for attosecond technology is reported in Section Attosecond Pulse Characterization.

Few-Cycle Lasers

Few-optical-cycle laser pulses with stabilized CEP are a fundamental prerequisite for the generation of attosecond pulses. Ti:sapphire lasers based on chirped pulse amplification (CPA) are the typical systems used to drive the HHG process. However, such laser systems routinely deliver 20 fs pulses and post compression techniques are required to shorten the pulse duration to sub-10 fs at 800-nm carrier wavelength. The most common technique for pulse compression of high-energy femtosecond pulses is based on propagation in a gas-filled hollow-core fiber in combination with ultra-broadband dispersion compensation. A promising alternative for scaling the generation of ultrashort laser pulses at the petawatt level is offered by the optical parametric chirped pulse amplification (OPCPA) technique, which combines optical parametric amplification (OPA) and CPA.

Hollow-Core Fiber Compression

In general, a post compression scheme for femtosecond pulses includes a phase modulator, which broadens the spectrum of the pulse, and a dispersion line, which compensates for the spectral phase thus making possible the achievement of near transform-limited pulse durations.

Self-phase modulation (SPM) in a Kerr medium produces an efficient spectral broadening. However, inhomogeneous spectral broadening affects the free space propagation of the pulses in a bulk medium because of the intensity profile of the beam. To overcome this limitation, guided propagation through a nonlinear medium can be used. Single-mode guiding line has to be employed otherwise intermodal dispersion distorts the pulse making its recompression impossible. This idea was first developed exploiting the propagation through optical fiber, where the nonlinear material is the fiber core itself, but the small diameter of single-mode optical fibers severely limited the maximum pulse energy to few tens of nJ in order to avoid material damage. Gas-filled hollow core fibers have been successfully demonstrated for pulse compression, overcoming this limitation. Rare gases are chosen as nonlinear medium, due to their high damage threshold and fast nonlinear response.

The hollow fiber consists of a fused silica capillary with an inner diameter of a few hundred microns. Light propagation in hollow fiber occurs by grazing incidence reflections at the dielectric inner surface of the fiber. The losses introduced by these reflections greatly limit the coupling of the incident radiation whit high order modes for long enough fibers. Thus, unlike in total internal reflection, in

this propagation mechanism the radiative mode-dependent losses select one fundamental mode. By a proper choice of the hollow fiber inner diameter, coupling efficiency for Gaussian beams can exceed 98%. Equations that describe the propagation of the laser pulse through the gas filled hollow fiber are the same as for propagation in standard optical fibers. The fundamental mode sustained by the hollow fiber is the hybrid EH_{11} mode, which corresponds to a truncated Bessel radial profile of the electric field.

The spectral broadening due to SPM in a hollow fiber of length L and radius a can be evaluated with the broadening factor F, which is defined as the ratio between the pulse bandwidth at the output of the fiber with respect to the same quantity at its input: $F=\Delta\omega/(\Delta\omega)_0$. A simple expression for this broadening factor can be derived as a function of the maximum phase shift $(\varphi)_m$ (Vozzi *et al.*, 2005).

$$F = \left(1 + \frac{4}{3\sqrt{3}}(\varphi)_m^2\right)^{\frac{1}{2}} \tag{1}$$

where $(\varphi)_m=\gamma P_0 L_{eff}$, being γ the nonlinear coefficient of the hollow core fiber, P_0 the peak power of the pulse and $L_{eff}=[1-e^{-\alpha L}]/\alpha$. $\frac{\alpha}{2}$ is the field attenuation constant of the fiber.

The nonlinear coefficient γ is given by the following relation:

$$\gamma = \frac{n_2\omega_0}{cA_{eff}} \tag{2}$$

where n_2 is the nonlinear index parameter, ω_0 is the central frequency of the pulse, c is the speed of light and A_{eff} is the effective mode area of the fiber, given by $A_{eff}\cong 0.48\pi a^2$. Thus for maximizing the broadening factor for a given incoming pulse with a certain energy and duration, one can increase the fiber length, increase the nonlinearity strength by increasing the gas pressure, or decrease the radius of the fiber. Several factors limit all these possibilities. The propagation losses of the fundamental mode, the distortion of the temporal pulse shape and some practical reasons limit the fiber length. Recently, new schemes based on a stretched flexible hollow fiber have been successfully implemented to overcome this last limitation. The pulse peak power should be lower than the critical power for self-focusing, in order to minimize the coupling of the fundamental transverse mode to the higher order modes which would introduce losses. The critical power for self-focusing is given by:

$$P_{crit} = \frac{\lambda_0^2}{2\kappa_2 p} \tag{3}$$

where κ_2 represents the ratio between the nonlinear coefficient and the gas pressure p and λ_0 is the central wavelength of the pulse.

A further limit on the fiber radius is set by the onset of ionization effects: The maximum pulse peak intensity, thus the minimum fiber radius a_{min}, must be chosen in order to have:

$$\frac{\Delta n_{kerr}}{\Delta n_p} \gg 1 \tag{4}$$

where $\Delta n_{kerr}=\kappa_2 pI$ is the refractive index variation due to the Kerr effect for a pulse of intensity I, while Δn_p is the plasma induced refractive index change, given by

$$\Delta n_p = \frac{\omega_p^2}{2\omega_0^2} \tag{5}$$

where $\omega_p = \sqrt{e^2\rho_e/(m_e\varepsilon_0)}$ is the plasma frequency, e and m_e are the electron charge and mass respectively and ρ_e is the free electron density in the gas.

The pulses emerging from the hollow core fiber have a group delay dispersion essentially due to the exit window of the cell that contains the fiber and to the air path after the fiber. In order to compensate for this group delay, chirped mirrors can be used, allowing compression down to sub 3-fs for mJ-energy pulses.

Carrier-Envelope-Phase Stability

The CEP ϕ of a light pulse is the phase delay between the highest half-cycle of the electric field associated to the pulse and the peak of the pulse envelope, as depicted in **Fig. 1**. Pulses propagating trough dispersive media experience a CEP change due to the difference between the phase velocity and the group velocity. CEP control matters for few-optical-cycle pulses in all those phenomena that depend directly on the electric field, such as multiphoton absorption, above-threshold ionization and high-harmonic generation. In all these cases, the reproducibility of the electric waveform is essential. In particular, CEP stability is a fundamental prerequisite for the generation of reproducible isolated attosecond pulses. The CEP of standard laser sources is intrinsically unstable, but it is possible to stabilize it in either active or passive ways.

In mode-locked oscillators, the carrier propagates with the phase velocity and the envelope propagates with the group velocity, thus a systematic change in the CEP $\Delta\phi$ is introduced for each round trip. A slow drift also affects this phase change $\Delta\phi$ due to the change in the oscillator parameters. At the output of the oscillator cavity a sequence of pulses is emitted in such a way that a pulse with CEP ϕ is followed by a pulse with CEP $\phi+\Delta\phi+2\ n\pi$. It is then possible to obtain two pulses with the same CEP at the oscillator output by waiting $2\pi/\Delta\phi$ round trips. Thus, the CEP period can be defined as:

$$T_{CEP} = \frac{2\pi}{\Delta\phi}T_R \tag{6}$$

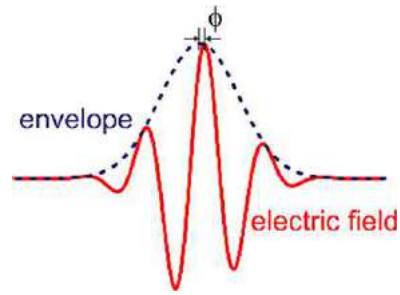

Fig. 1 Carrier-envelope-phase (CEP) ϕ of a light pulse.

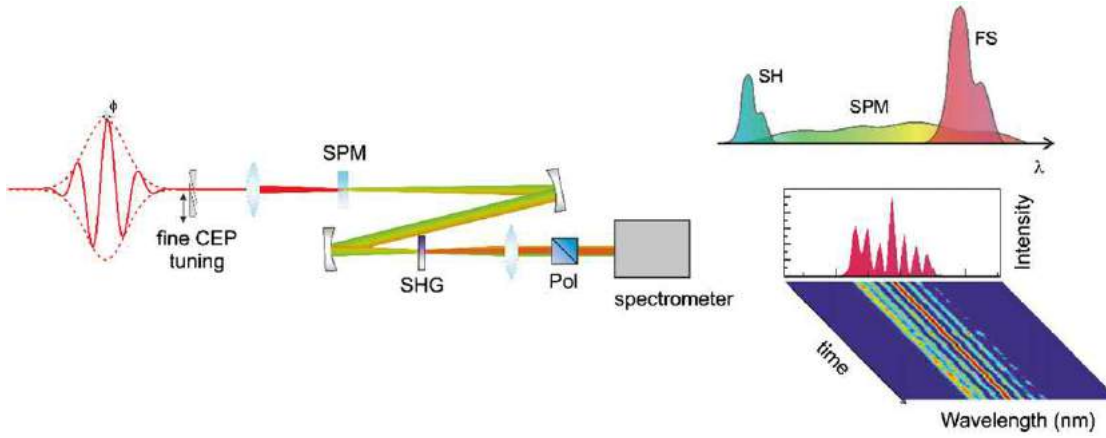

Fig. 2 Left panel: scheme of the interferometric measurement of the CEP drift by an f-to-2f interferometer. Right panel: spectral fringes. FS, fundamental spectrum; Pol, polarizer; SHG, second harmonic generation; SPM, self-phase modulation.

where, T_R is the period of a roundtrip in the laser cavity. The corresponding CEP frequency reads:

$$\omega_{CEP} = \frac{\Delta\phi}{2\pi}\omega_R \tag{7}$$

The last relation suggests that a train of pulses with the same value of CEP can be obtained by selecting the pulses at a frequency ω_{CEP} or at a fraction of it. The measurement of ω_{CEP} exploits the frequency comb emitted by the oscillator, which is the superposition of longitudinal modes with frequencies $\omega_n = \omega_{CEP} + n\omega_R$, with well-known metrology applications. In order to access the frequency ω_{CEP}, it is possible to measure the beating between the fundamental comb and its sum frequency in the spectral region where they overlap, assuming that the fundamental spectrum spans over an octave. In an f-to-2f interferometer, the fundamental comb with frequencies ω_n and a second harmonic comb at frequencies $\omega_{SH} = 2\omega_n = 2\omega_{CEP} + 2m\omega_R$ overlap and generate a beating for the modes with $n = 2m$ at the frequency ω_{CEP}. The frequency ω_{CEP} can then be stabilized by an active feedback on the oscillator, typically by changing the pump power with an acousto-optic modulator. In this way, one obtains a train of CEP-stable pulses at a fraction of the frequency ω_{CEP} that can be amplified in solid-state amplifiers or in optical parametric amplifiers.

In amplified systems with repetition rates of the order of a few kHz it is impossible to directly measure the comb frequencies. In this case, in order to characterize the CEP of the individual pulses it is possible to exploit interferometric techniques. Indeed the electric field associated with the pulse can be written as:

$$E(t) = A(t)e^{i(\omega t+\phi)} + c.c. = E_0(t)e^{i\phi} + c.c. = F\{\tilde{E}_0(\omega)\}e^{i\phi} + c.c. \tag{8}$$

If the pulse with fundamental frequency ω_0 and its m-th harmonic at frequency $m\omega_0$ overlap in some region of the pulse spectrum, in this spectral region an interference modulation appears

$$\cos[(m+1)\phi + \omega\tau + c] \tag{9}$$

where, τ is the delay between the two components at ω_0 and $m\omega_0$ and c is a constant. The shot-to-shot CEP shift can then be tracked by the direct observation of interference fringes. A typical setup for this kind of measurement is shown in Fig. 2. In this f-to-2f interferometer the fundamental pulse (F) is focused into a sapphire plate for generating white light with an octave spanning bandwidth. The fundamental pulse is then frequency doubled in a nonlinear crystal in such a way that the second harmonic (SH) spectrally overlaps the white light. These two spectral components are sent to a spectrometer trough a polarizer. The intensity

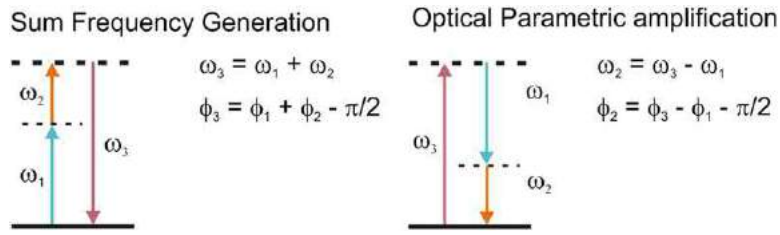

Fig. 3 Schematic view of how the CEP is affected by some common nonlinear phenomena.

measured by the spectrometer is modulated according to the following relation:

$$I(\omega) = I_F(\omega) + I_{SH}(\omega) + 2\sqrt{I_F(\omega)I_{SH}(\omega)}\cos(\omega\tau + \phi) \tag{10}$$

where τ is the delay between the fundamental pulse and its second harmonic. A drift in the CEP reflects in a shot-to-shot change of the position of the fringes in the interference pattern. This change can be used in a slow feedback loop for stabilizing slow-varying CEP noise in amplified systems. Commercially available amplified Ti:sapphire systems with active CEP stabilization show rms CEP fluctuations lower than 100 mrad. In order to measure the actual value of the CEP, more sophisticated instruments based on above threshold ionization (ATI) are necessary for the absolute calibration of the above mentioned interference trace.

An alternative approach to generate CEP stabilized pulses is the passive one, cancelling CEP fluctuations based on the nonlinear optical processes. The effect of these processes on the CEP can be considered as follows. In sum frequency generation (SFG), two pulses at frequency ω_1 and ω_2 and CEP ϕ_1 and ϕ_2 generate a third pulse at frequency $\omega_3 = \omega_1 + \omega_2$ with a CEP $\phi_3 = \phi_1 + \phi_2 + \frac{\pi}{2}$. Second harmonic generation is a particular case of SFG and thus from a pulse of frequency ω and CEP ϕ one obtains a pulse at frequency 2ω with CEP $2\phi + \frac{\pi}{2}$. In the difference frequency generation (DFG), two pulses at frequencies ω_1 and ω_2 and CEP ϕ_1 and ϕ_2 give a resulting frequency $\omega_3 = \omega_1 - \omega_2$ with CEP $\phi_3 = \phi_1 - \phi_2 - \frac{\pi}{2}$. Optical parametric amplification (OPA) is similar to DFG: the intense pump pulse at frequency ω_3 with CEP ϕ_3 amplifies the weak signal pulse at frequency ω_1 with CEP ϕ_1 and generate an idler pulse at frequency $\omega_2 = \omega_3 - \omega_1$. In the process the CEP of the signal is not affected, while the idler has a CEP $\phi_2 = \phi_3 - \phi_1 - \frac{\pi}{2}$ (**Fig. 3**). Self-phase modulation preserves the CEP and adds a constant phase shift of $\pi/2$.

Passive stabilization of the CEP is based on DFG between two pulses sharing the same CEP. The CEP of the DFG results as the difference between the phases of the two generating pulses, thus the shot-to-shot fluctuations of the CEP cancel. This approach allows the generation of CEP stable few cycle pulses in a broad spectral range from the visible to the mid-IR. This approach is all-optical and has the advantage of avoiding electronic feedback. Moreover, the train of CEP stable pulses is automatically generated at the same repetition rate of the driving field, without the need of picking up pulses at a fraction of the frequency ω_{CEP}. There are two possible configurations for implementing this method. In the *inter-pulse* configuration, two CEP-locked frequency-shifted pulses are synchronized by a delay line and then they are mixed in the nonlinear crystal for DFG. A drawback of this scheme is that the delay line can introduce CEP jitter due to mechanical instabilities. *Intra-pulse* schemes involve mixing between different frequencies of a single ultra-broadband pulse that generate the difference frequency. In this case, since the DFG occurs between spectral portions of the same pulse, the delay-induced CEP jitter is suppressed and the synchronization between the two components is automatically achieved by means of the control of the group delay of the pulse. A review of several laser systems implementing passive CEP stabilization is presented in Cerullo *et al.* (2010).

OPA and OPCPA Technology

Aside from the well established laser sources based on Ti:sapphire for HHG and attosecond pulse generation, innovative driving sources based on Optical Parametric Amplification (OPA) and Optical Parametric Chirped Pulse Amplification (OPCPA) have been proposed.

OPA is an excellent technique for the generation of high-energy, few-optical cycle pulses tunable in a wide spectral range. Passive stabilization of the CEP can be also implemented in OPAs. OPCPA-based sources promise to offer the possibility of generating ultrashort laser pulses with peak power in the PW range, overcoming some of the limitations on the scalability of OPA based lasers.

A review of recent developments in few-cycle OPAs and OPCPAs for HHG and attosecond generation is presented in Ciriolo *et al.* (2017).

Pulse Synthesis

Exploiting the above listed techniques, ultra-broadband pulses comprising just a few optical cycles are routinely produced. In order to achieve a pulse duration down to the single-cycle limit, coherent combination (or synthesis) of pulses having different colours is required.

One possible approach to achieve pulse synthesis consists of the generation of a super broadband spectrum (above one-octave) to be divided into several CEP-stable channels, which are individually compressed by custom-designed chirped mirrors (CMs) and then coherently recombined in a sub-cycle optical waveform. This passive approach is schematically shown in **Fig. 4**.

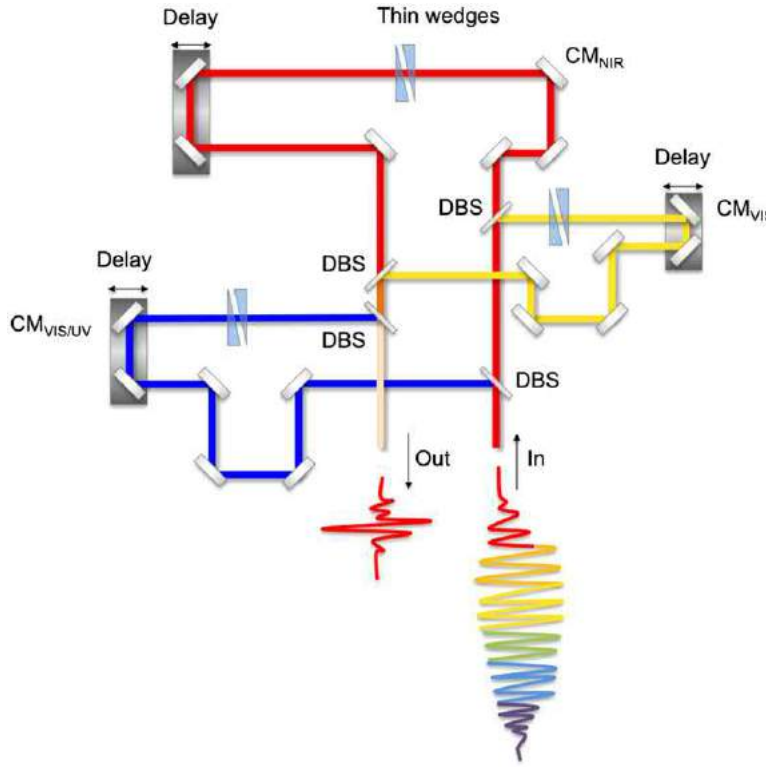

Fig. 4 Scheme of a three-channel light-field synthesizer. Dichroic beam splitters (DBS) are used to divide a supercontinuum into three channels: ultraviolet (UV), visible (VIS), near-infrared (NIR). Pulses of each channel are compressed using custom-designed chirped mirrors (CM) Fine-tuning of dispersion, CEP and delay is achieved with pairs of fused silica wedges and with delay stages in each channel, respectively.

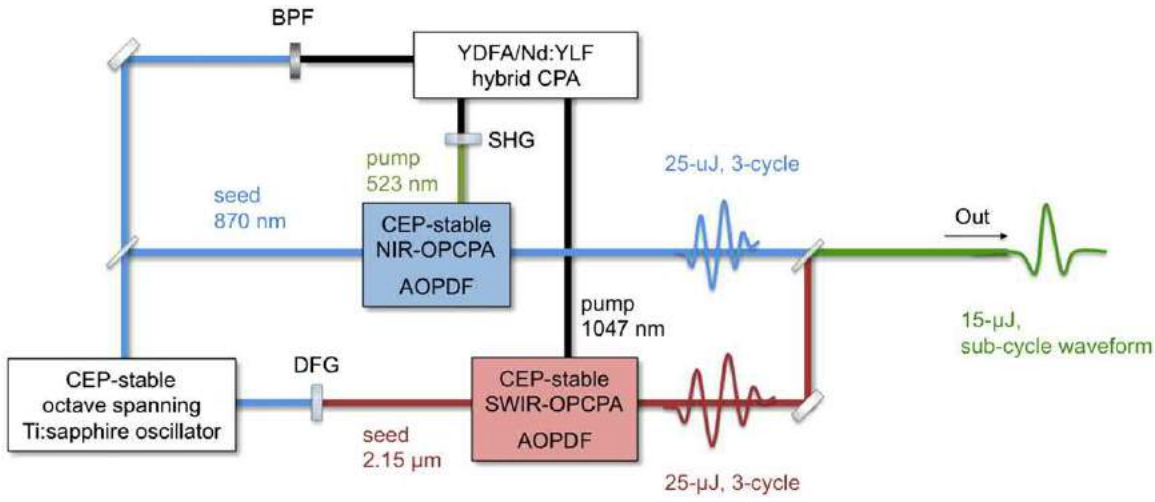

Fig. 5 Scheme of the OPCPA-based pulse synthesizer. A sub-cycle waveform spanning over 1.8 octaves is obtained by combining a NIR OPCPA and SWIR OPCPA. A common front-end is used for both OPCPAs. Waveform shaping is achieved by controlling the CEP of both channels. Control of the chirp is obtained with an acousto-optic programmable dispersive filter (AOPDF). BPF, band pass filter; DFG, difference frequency generation; SHG, second harmonic generation.

Coherent combination of ultra-broadband OPAs is another approach to single-cycle waveform synthesis. In this case, the synthesis is achieved by coherently combining two spectrally adjacent OPAs seeded by distinct portions of the same CEP-stable white light continuum. Energy scaling of this approach can be achieved by replacing the OPA with OPCPA technology. The design of an OPCPA-based pulse synthesizer is shown in Fig. 5. This design could be easily scaled to higher energies by implementing into the scheme higher-power pump lasers as for instance the new generation of cryo-Yb:YAG lasers.

A comprehensive review of the pulse synthesis methods is reported in Manzoni *et al.* (2015).

Ultrashort Pulse Characterization

The tremendous progress in the development of ultra-broadband pulses calls for a precise temporal characterization of these light transients. The time-dependent electric field of the laser pulse can be written as

$$E(t) = Re\left\{\sqrt{I(t)}exp(i\omega_0 t - i(\varphi)(t))\right\} \tag{11}$$

where, $I(t)$ and $(\varphi)(t)$ are the time-dependent intensity and phase and ω_0 is the carrier frequency.

The main technique used to temporally characterize laser pulses is the autocorrelation. An autocorrelator is a device that involves splitting the pulse into two variably delayed replica, subsequently spatially overlapped in a second harmonic generation (SHG) crystal. The second harmonic signal acquired as a function of the time-delay between the two pulses provides the autocorrelation trace. In order to estimate $I(t)$ this method relies on a "guess" for the pulse shape, which is the main source of error. Moreover no information can be retrieved on the pulse phase $(\varphi)(t)$.

A large number of schemes have been developed over the past two decades to provide a better measurement of ultrashort laser pulses. Among them, the two most commonly used methods are the frequency-resolved optical gating (FROG) and the spectral phase interferometry for direct electric-field reconstruction (SPIDER), both providing a full characterization ($I(t)$,$(\varphi)(t)$) of the laser pulse. FROG is an extension of the autocorrelation technique: when the second harmonic signal from an autocorrelator is acquired by a spectrometer, the autocorrelation becomes a two-dimensional spectrally-resolved trace (spectrogram). In general, the measured spectrogram can be expressed as:

$$S(\omega,\tau) = \left|\int E(t)g(t-\tau)exp(-i\omega t)dt\right|^2 = \left|\int E_{sig}(t)exp(-i\omega t)dt\right|^2 \tag{12}$$

where $g(t–\tau)$ is a gate function, which selects a portion of the electric field $E(t)$ around the delay τ. Extracting the laser field $E(t)$ from the measured FROG trace $S(\omega,\tau)$ is a two-dimensional phase retrieval problem, and iterative algorithms are required for finding a solution. Various versions of FROG exist, which rely on different nonlinear gating mechanisms.

Spectral interferometry is another approach to characterize a laser pulse. Ideally, the spectral interference between the pulse under test and a reference pulse with a well-known spectral phase should be used to extract the spectral phase of the unknown pulse. However, the reference pulse is usually not available. SPIDER is a technique introduced to overcome this limitation. The idea is to generate two upconverted spectra slightly shifted in frequency and to measure their spectral interference. The electric field of each individual pulse can be expressed as:

$$\begin{aligned} E_1(t) &= E(t)e^{i\omega_s t} \\ E_2(t) &= E(t-\tau)e^{i(\omega_s+\Omega)(t-\tau)} \end{aligned} \tag{13}$$

where ω_s and $\omega_s+\Omega$ are the two frequencies resulting from the non-linear mixing, Ω is called spectral shear, τ is the time delay between the two replicas and $E(t)$ is the electric field to be characterized. The measured spectrum can be written as:

$$S(\omega) = \left|\int_{-\infty}^{+\infty}(E_1(t)+E_2(t))e^{-i\omega t}dt\right|^2 = S_{DC}(\omega) + S^+(\omega)e^{-i\omega t} + S^-(\omega)e^{i\omega t} \tag{14}$$

For a sufficiently large delay, the term $S^+(\omega)e^{-i\omega t}$ can be extracted and its phase can be uniquely determined:

$$\phi(\omega) = (\varphi)(\omega-\omega_s-\Omega) - (\varphi)(\omega-\omega_s) - \omega\tau \tag{15}$$

If the delay τ is precisely calibrated with an independent method, the linear phase $\omega\tau$ can be subtracted in Eq. (15) and the spectral phase of the unknown pulse can be determined by integration. A simple way to implement this scheme is shown in **Fig. 6**.

A more detailed description of the ultrashort pulse characterization methods can be found in Manzoni *et al.* (2015).

Attosecond Technology

High-Order Harmonic Generation

High order harmonic generation (HHG) is a strongly nonlinear process occurring when atoms or molecules are exposed to laser radiation with intensity of the order of $I \sim 10^{14}$ W/cm^2. It represents a unique and table-top source of coherent XUV and soft X-ray radiation. This radiation has also very interesting temporal characteristics: it consists of a sequence of light bursts, with sub femtosecond duration.

The semi-classical three-step model (Corkum, 1994) gives a simple picture of the physical process leading to HHG, which is valid in the strong field regime and assuming a single active electron approximation. As sketched in **Fig. 7**, the external laser field ionizes the outermost electron in the atom. The freed electron is then accelerated by the external field and brought back to the parent ion where it can recombine, giving rise to the emission of and XUV photon. The energy of the emitted photon depends on the kinetic energy of the recombining electron, thus on the trajectory followed by the electron in the external field. If we consider

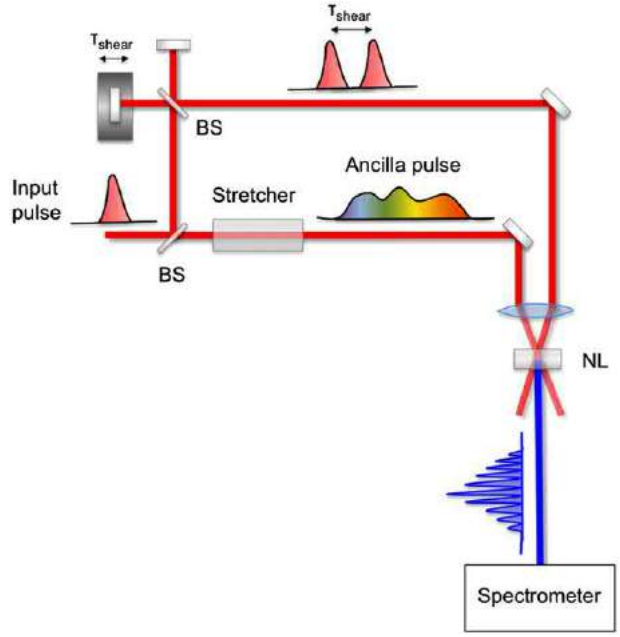

Fig. 6 Scheme of the SPIDER technique. Two delayed replicas of the input pulse are synchronized with two different frequency components of an ancilla pulse, separated by Ω. The upconverison process, occurring in a non-linear crystal (NL), frequency shifts the central frequency of the two pulses to ω_s and $\omega_s+\Omega$ respectively. A spectrometer detects the interference pattern. BS, beam splitter.

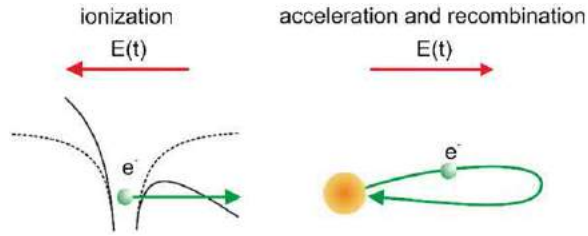

Fig. 7 Sketch of the HHG process as depicted in the three step model: on the left the active electron is ionized by the external laser field; on the right the electron is accelerated by the external field and can eventually recombine with the parent ion.

the interaction of an atom with a monochromatic electric field linearly polarized along the x direction $E(t)=E_0\cos(\omega_0 t)$, the electron ionized at the instant t', is accelerated by the external field and depending on the ionization time, it can be driven away or it returns to the parent ion. A photon of frequency $\omega=(I_p+E_k/\hbar)$ is emitted upon recombination, being I_p the atom ionization potential and E_k the kinetic energy gained by the electron in the continuum. The electron motion equation in the external field can be integrated assuming initial conditions $x(t')=0$ and $v(t')=0$ and for a given ionization time t', it is possible to calculate the recombination time t. For some ionization times, the electron revisits the parent ion several times, but as a first approximation, the successive recollisions can be neglected due to the quantum diffusion of the electron wave function, which reduces the recombination probability. This process is repeated periodically every half cycle of the external field and, due to the inversion symmetry of the atom, only the ionization and recollision events within this time interval must be considered. As shown in **Fig. 8**, for any photon energy, there are two solutions of the motion equation characterized by the different values of the time by the electron in the continuum. The solutions associated to the shorter time travel times are called short trajectories, whilst the other ones are referred to as long trajectories. The maximum kinetic energy gained by the electron in the continuum turns out to be (E_k) $max=3.17U_P$ where U_P is the ponderomotive energy given by

$$U_P=\frac{e^2E_0^2}{4m_e\omega_0} \tag{16}$$

Since the laser pulse contains a few optical cycles, in the temporal domain the XUV emission corresponds to a sequence of XUV pulses separated in time by half-cycle of the driving filed, as is depicted in **Fig. 9**. In the spectral domain, this emission corresponds to odd harmonics of the fundamental laser frequency. The harmonic spectrum contains two distinct regions. In the *plateau* region, the intensity of the harmonic is almost flat as a function of harmonic order. In the *cut-off* region, which comprises only a few harmonic orders, the harmonic yield decreases exponentially.

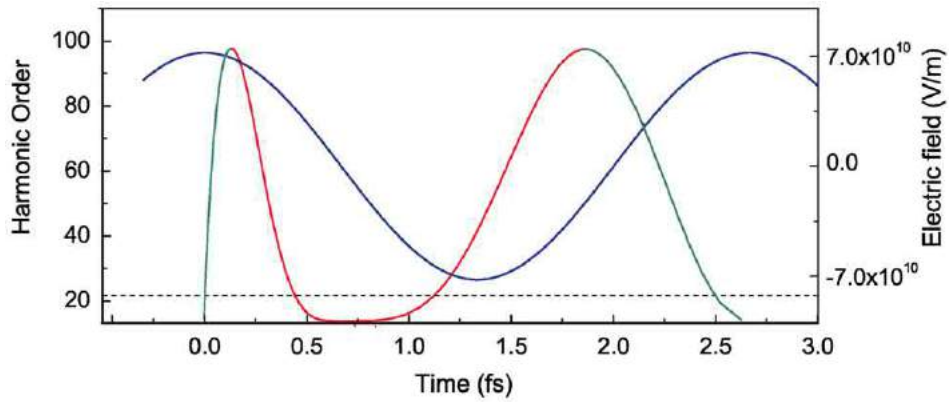

Fig. 8 Solution of the classical motion equation for the electron in a monochromatic external field of the form $E(t)=E_0\cos(\omega_0 t)$ with $E_0=7.2\ 10^{10}$ V/m and $\omega_0=2.36\ 10^{15}$ rad/s. The electric field evolution is shown as a blue curve. The left-hand side of the curve individuates the ionization times t′ for short (red) and long (green) electron trajectories. On the right-hand side of the curve, the recombination times for short (red) and long (green) electron trajectories are shown.

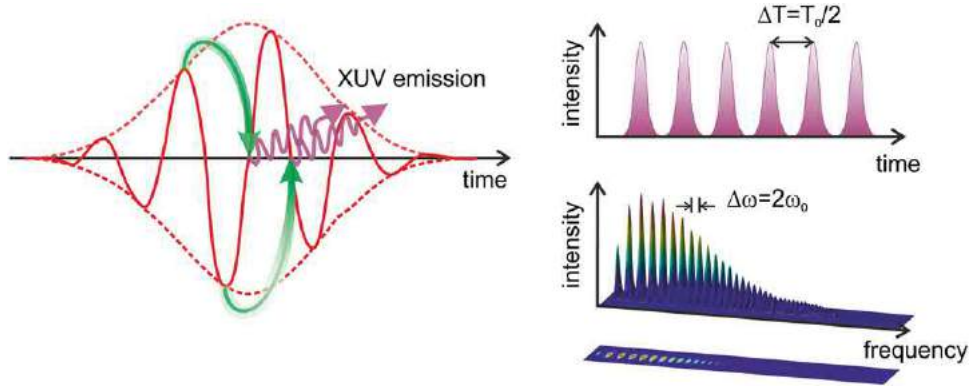

Fig. 9 Characteristics of harmonic emission. Left panel: the HHG process is repeated periodically every half cycle. Right panel: harmonic emission corresponds to a train of attosecond pulses in the temporal domain (upper figure) and to a sequence of odd harmonics of the fundamental driving frequency (lower figure).

The semi-classical model allows grasping most of the peculiarities of harmonic emission, nevertheless a full description of the laser-atom interaction requires a quantum model, such as the Lewenstein model (Lewenstein *et al.*, 1994). In this framework the harmonic emission spectrum is proportional to the Fourier transform of the atomic dipole moment:

$$x(t)=\langle\psi(t)|\boldsymbol{er}\cdot\boldsymbol{E}(t)|\psi(t)\rangle \tag{17}$$

where $|\psi(t)\rangle$ is the solution of Schrödinger equation and $\boldsymbol{er}$ is the dipole moment operator. A simple analytical expression for the atomic dipole moment can be derived with the assumptions of single active electron (SAE – the atom is considered as a hydrogen-like system and multiple ionization is neglected) and strong field approximation (SFA – the influence of the atomic Coulomb potential on the motion of the electron in the continuum is neglected and the external electric field has no influence on the atomic ground state wave function):

$$x(t)=i\int_{-\infty}^{t}dt'\int d^3p E(t')\,d_x^*[p-A(t)]d_x[p-A(t')]e^{-iS(p,t,t')}+c.c. \tag{18}$$

where atomic units have been used and $E(t')$ is the external electric field; p is the canonical momentum given by $p=v+A(t)$, v is the electron velocity, $A(t)$ is the vector potential associated to the external field, $d_x[p-A(t)]$ is the expectation value for the dipole moment transition between the atomic ground state and the continuum state associated to electron velocity $v=p-A(t)$, $S(p,t,t')$ is the quasi-classical action defined as:

$$S\left(p,t,t'\right)=\int_{t'}^{t}dt''\left(\frac{\left[p-A(t'')\right]^2}{2}+I_P\right) \tag{19}$$

which corresponds to the phase accumulated by the free electron during its propagation in the continuum.

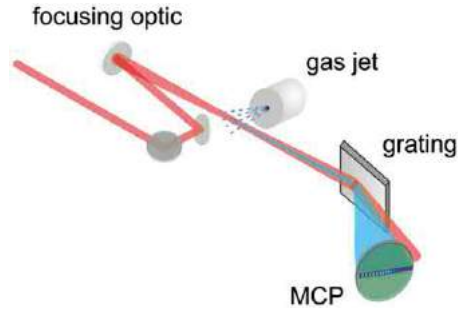

Fig. 10 Typical setup for harmonic generation and detection. MCP, micro channel plate.

The single atom harmonic spectrum can be calculated as:

$$\tilde{x}(\omega) \propto \int_0^{T_0} dt' x(t) e^{i\omega t} \tag{20}$$

Thus the dipole moment at the time t is given by the contribution of all the electrons injected in the continuum at the time t' with a momentum p. $E(t')d_x[p - A(t')]$ is the probability that an electron is ionized by the external field at time t and appears in the continuum with a momentum p. The electrons accelerated by the external field accumulate a phase $S(p,t,t')$ and eventually recombine with parent ions at the time t with a probability $d_x^*[p - A(t)]$.

The relevant trajectories contributing to the Lewenstein integral can be selected with the saddle point approximation (SPA). The trajectories found with the three-step model correspond to the classical limit of the saddle-point quantum paths. The real part of such quantum paths is mostly similar to the classical electron trajectories while the imaginary part is related to the quantum process of tunnel ionization.

The harmonic spectrum results from the coherent superposition of single-atom contributions. Thus for the physical interpretation of experimental results the propagation of harmonic radiation in the atomic gas has to be taken into account. For maximizing the harmonic conversion efficiency, phase-matching conditions between the driving pulse and the co-propagating harmonic radiation must be fulfilled. Several factors contribute to phase matching in HHG, such as the phase $S(p,t,t')$ accumulated by the electron in the continuum, the time-dependent refractive index due to the plasma induced by the fundamental pulse and the geometrical phase factor related to the focusing geometry of the fundamental field (Guoy phase). In the case of a Gaussian driving pulse, when the generating medium is located around the laser focus – see **Fig. 10** for a sketch of the typical experimental setup – the phase matching condition is fulfilled for harmonic radiation generated off-axis and the structure of harmonic beam is annular. On the other hand, if the atomic medium is located after the laser focus the harmonic beam is efficiently generated along the driving pulse propagation direction. Furthermore, as the dipole phase term depends on the quantum trajectory, phase-matching conditions can enhance or depress the contribution of short or long trajectories. For a Gaussian beam, when the atomic medium is placed after the laser focus, the contribution of the long quantum paths is suppressed while when the atomic medium is placed in correspondence of the laser focus the contribution of the long quantum paths increases.

The optimization of phase-matching conditions is essential for maximizing the HHG yield and several strategies have been demonstrated so far, such as loose focusing geometry, hollow waveguide for the generating medium or periodic modulation of the gas density.

Gating Techniques

For several time-resolved applications, it is required to generate a single attosecond pulse instead of an attosecond pulse train. Various schemes have been developed to confine the harmonic generation process to a single event. These schemes can be grouped into two main categories: amplitude gating and temporal gating. In the amplitude gating scheme the selection of a single attosecond pulse is done by spectrally filtering the cutoff energy region of the harmonic spectrum, where only the most intense half cycle of the driving pulse contributes to the emission (left panel of **Fig. 11**). This generation scheme requires the use of intense few-optical-cycle driving pulses, with controlled CEP. Isolated attosecond pulses with a temporal duration down to 80 as have been generated using this approach (Goulielmakis *et al.*, 2008).

In the case of temporal gating, HHG is confined to a single emission event by properly manipulating the driving electric field in time, without limitation in the generated XUV pulse bandwidth (right panel of **Fig. 11**). One possible temporal gating approach is called polarization gating (PG) and it takes advantage of the strong sensitivity of the HHG process on the degree of ellipticity of the fundamental field. If the polarization of the driving field is temporally manipulated from circular to linear and back to circular, the XUV emission is limited to the temporal window in which the driving pulse is linearly polarized. Top panel of **Fig. 12** shows a possible scheme for the implementation of PG: a first quartz plate is used to separate the linearly polarized input pulse into two orthogonally polarized delayed pulses, so that the outgoing pulse polarization evolves from linear to circular, and back to linear. A zero-order broadband quarter waveplate is then used to obtain a pulse linearly polarized at its centre and circularly polarized in its

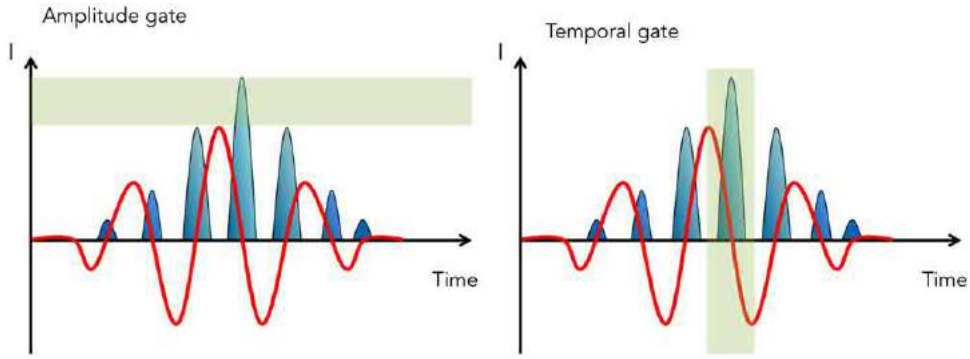

Fig. 11 Operating principle of the main gating schemes for the generation of isolated attosecond pulses. The red line represents the electric field of the driving pulse and the blue shaded area represents the corresponding ionization times. A single attosecond pulse can be selected either by spectrally filtering the cut-off energies of the harmonic spectrum (left panel) or by temporally gating the recombination process (right panel).

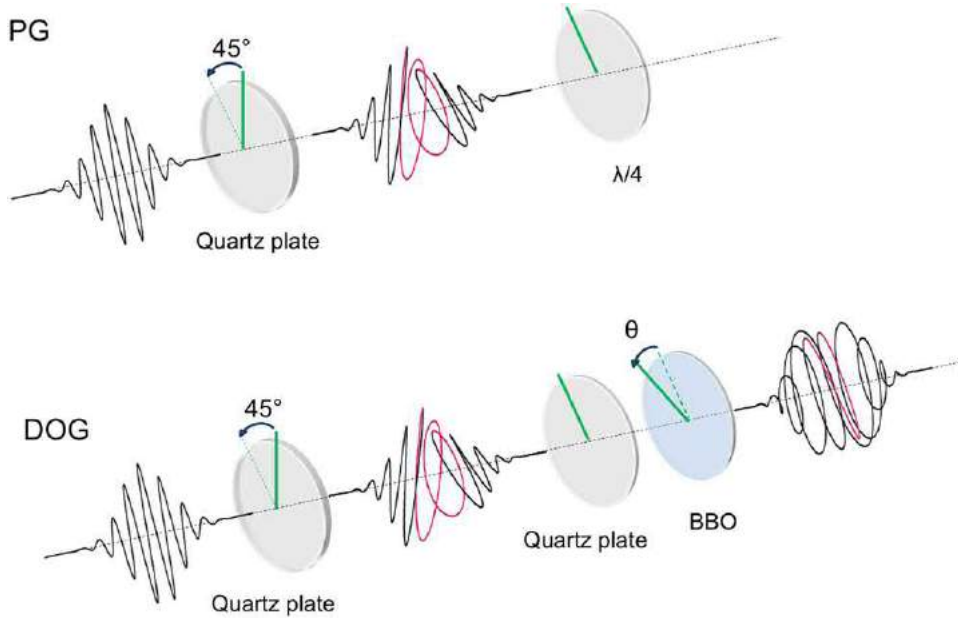

Fig. 12 Optical scheme for polarization gating (top panel) and for double optical gating (bottom panel).

wings. The duration of the temporal gate δt_g where the polarization is linear and HHG can occur can be estimated as:

$$\delta t_g = 0.3 \frac{\varepsilon_{thr} {\tau_p}^2}{\tau_d} \tag{21}$$

where τ_d is the temporal delay between the two orthogonally polarized pulses at the output of the first quartz plate, τ_p is the pulse duration and ϵ_{thr} is a threshold value of ellipticity for which the harmonic signal is decreased to 50%. By using this approach isolated attosecond pulses as short as 130 as have been generated and characterized (Sansone *et al.*, 2006).

In the PG approach, the gate window should be approximately $T/2$, where T is the period of the driving field. By using Eq. (21), it is possible to demonstrate that such a narrow gate requires that $\tau_p = 2.5T$, which corresponds to less than 7 fs at 800 nm carrier wavelength. The double optical gating (DOG) technique is an extension of the PG method, which relaxes the requirement on the pulse duration. The idea is to add the second harmonic of the driving field to the fundamental pulse to increase the separation between two XUV emission events by a factor of two. This allows one to apply a one-optical-cycle gate instead of a half-cycle gate. The optical scheme for the DOG method is depicted in bottom panel of Fig. 12 and it differs from the PG scheme only for the thickness of the quartz plates and the presence of a BBO crystal for second harmonic generation. In this case, the second quartz plate and the BBO crystal act together as a quarter wave plate. The DOG technique has been successfully implemented to generate 67-as isolated pulses. A generalized version of the DOG (GDOG) also exists, allowing for the generation of isolated attosecond pulses from multi-cycles pulses to be achieved.

Another possible approach for temporally confining the XUV emission is dubbed ionization gating (IG) and it exploits high-intensity driving fields. For high excitation intensities and ultrashort pulse durations, the plasma density rapidly increases on the leading edge of the pulse, thus creating a single-atom sub-cycle ionization response that can be exploited for temporal gating. With

this technique, isolated attosecond pulses with pulse duration of 155 as and pulse energy on target of 2.1 nJ were generated in xenon (Ferrari *et al.*, 2009).

An alternative approach to temporal gating is spatial gating, in which the selection of a single attosecond pulse is obtained by spatially filtering the EUV radiation. The idea is to drive HHG using ultrashort pulses with a non-negligible pulse-front tilt, so that each attosecond pulse is emitted in a slightly different direction, which corresponds to the instantaneous direction of propagation of the driving field at the instant of generation. Assuming that the angular separation between two consecutive XUV pulses is larger than their divergence, in the farfield it is possible to spatially select each individual attosecond pulse. Since isolated attosecond pulses are emitted in different directions, the effect has been dubbed attosecond lighthouse. Angular separation between two-consecutive attosecond pulses can be also achieved by focusing two non-collinear laser beams at the position of the gas target, the method is called noncollinear optical gating (NOG).

Finally it is worth mentioning that most of the gating techniques require CEP stability of the driving few-optical-cycle pulse. CEP-control can be also exploited to tailor the XUV emission: by changing the CEP value it is indeed possible to spectrally tune the XUV emission as well as to select the emission of an isolated attosecond pulse or a pair of attosecond pulses. A more detailed description, with appropriate references, to all the above-mentioned gating methods is reported in Calegari *et al.* (2016).

Attosecond Pulse Generation in the Water-Window Region

According to the three-step model, the harmonic cut-off scales proportionally to the ponderomotive energy, thus it increases with the driving wavelength squared: $\hbar\omega_{cut\text{-}off} \propto I\lambda^2$, where I and λ are the peak intensity and the wavelength of the driving field respectively. This result has been confirmed experimentally and it implies that mid-IR driving sources can be very promising for the generation of XUV photons with very high energy, since for the same peak intensity of 800-nm Ti:Sapphire lasers they allow a huge extension of the harmonic cut-off toward the X-rays.

Unfortunately for longer driving pulses the excursion time of the electron in the continuum increases and the probability of recombination with the parent ion strongly decreases due to the spatial spread of the electron wave function. For the single atom, the harmonic power spectrum yield decreases with the driving wavelength roughly as λ^{-6}.

Since the possibility of generating coherent radiation in the water window spectral region is very intriguing, specific generation geometries have been developed for mid-IR driving sources. In particular, the unfavourable scaling of the single-atom response can be compensated by a consistent increase of the generating medium gas pressure (Popmintchev *et al.*, 2012). In this way, a cutoff energy of 1.6 keV was observed in HHG driven by 3.9 μm driving pulses in a helium-filled capillary.

Another important feature of HHG is the attochirp of the emitted XUV bursts. This is the natural dispersion of photon energies along the temporal duration of a single XUV attosecond burst emitted during HHG: recalling the three-step model, different flight times (trajectories) of the electron correspond to different kinetic energies, hence, different photon energies are emitted at different times. The attochirp scales as λ^{-1}, thus it decreases for increasing the driving wavelength. Indeed the attochirp is proportional to the ratio between the period of the fundamental pulse and the harmonic cutoff energy. At constant intensity, the former scales as λ and the latter scales as λ^2, thus the attochirp scales as the inverse of the driving wavelength. The scaling of the attochirp with the driving wavelength also implies that shorter attosecond pulses could be emitted when harmonics are driven by mid-IR sources.

Attosecond Pulse Characterization

Femtosecond metrology allows for the complete characterization of ultrashort light pulses in the visible or near-visible range. As mentioned above, such a complete characterization requires the use of at least one time-nonstationary filter, typically obtained with a nonlinear effect. Transferring this idea to attosecond metrology is exceedingly difficult primarily due to limitations imposed by the low XUV photon flux. For this reason, other approaches have been implemented for characterizing attosecond pulses, which can be mainly divided in two categories. In the first approach, called *ex situ*, the photoelectrons produced in a target medium by the attosecond pulse are perturbed by the presence of a synchronized laser field, while, the second approach, called *in situ*, is based on the gentle perturbation of the electron trajectories during the attosecond pulse generation process itself.

Among the *ex situ* techniques, the Reconstruction of Attosecond Beating By Interference of Two-photon Transitions (RABBITT) was the first proposed method (Paul *et al.*, 2001). In the RABBITT scheme, the XUV pulse train is synchronized with a weak IR laser field to produce electrons by photoionization of a gas target. As shown in **Fig. 13**, the XUV pulse train generate a photoelectron spectrum, which is a replica of the harmonic spectrum: the peaks are separated by $2\hbar\omega$, where ω is the carrier frequency of the IR field. When the IR field is properly synchronized with the XUV field, additional peaks at $\pm\hbar\omega$ are produced in the photoelectron spectrum, thus leading to the appearance of the so-called sidebands. These peaks are due to the absorption or emission of one (or more) IR photons together with the absorption of one XUV photon. Typically the IR intensity is chosen such that only one IR photon can be absorbed or emitted. As depicted in the left panel of **Fig. 13**, the same sideband can originate from two different paths: (i) the absorption of one XUV photon corresponding to harmonic $q+1$ and the emission of one IR photon, or (ii) the absorption of one XUV photon corresponding to harmonic $q-1$ plus one IR photon. The two indistinguishable paths lead to interference, and the amplitude of the sidebands (SB) is periodically modulated as a function of the delay τ between the XUV and the IR pulses as:

$$SB = A_f \cos\left(2\omega\tau - \Delta(\varphi) - \Delta(\varphi)_f\right) \tag{22}$$

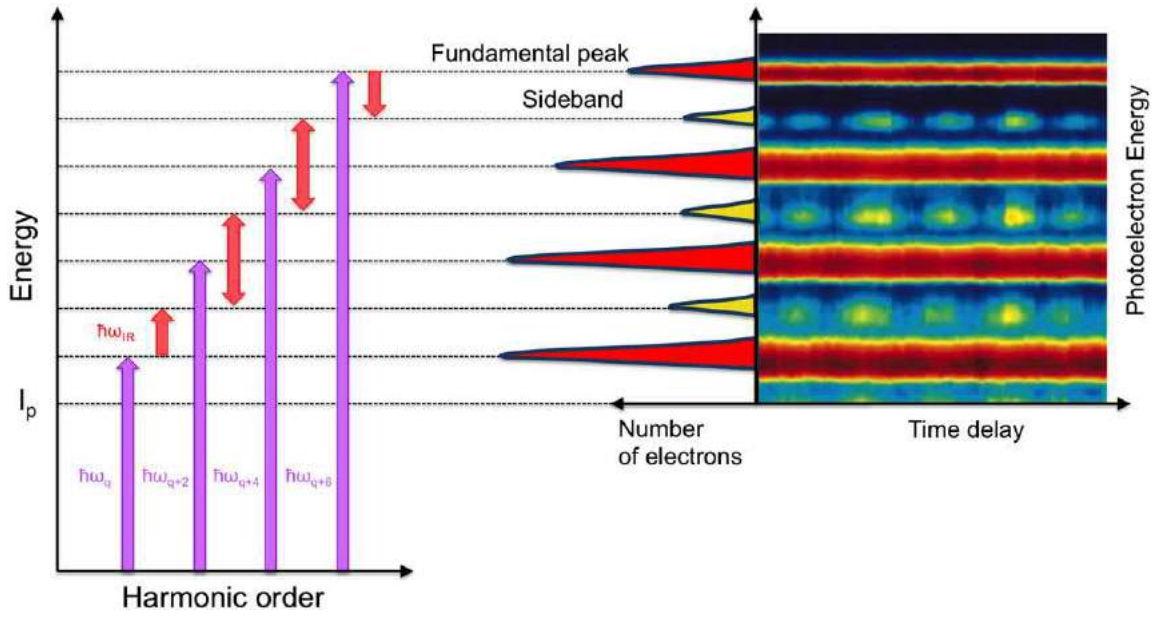

Fig. 13 RABBITT scheme. Left panel: odd order harmonics ionize the medium and create a photoelectron signal (red peaks). Further absorption/emission of an IR photon (red arrows) creates sideband peaks in the photoelectron spectrum (yellow peaks). Right panel: sideband amplitude oscillation at twice the frequency of the IR field.

where A_f is the amplitude, $\Delta(\varphi)=(\varphi)_{q+1}-(\varphi)_{q-1}$ is the phase difference between harmonics $q+1$ and $q-1$, and $\Delta(\varphi)_f$ is the intrinsic atomic phase difference between the matrix elements corresponding to photo-ionization from the $q+1$ and $q-1$ harmonics. This equation clearly indicates that the interference results in amplitude oscillations at twice the IR frequency upon changing the delay τ (see right panel of **Fig. 13**). If a suitable gas is used as a target, the intrinsic phase $\Delta(\varphi)_f$ can be calculated and from the time-delay scan the phase difference between two consecutive harmonics can be extracted.

In a typical optical scheme for RABBITT, a 800-nm laser is split in two arms: a portion of the beam is used to generate the harmonics in a noble-gas jet, while the remaining part of the IR beam is properly delayed and then collinearly focused together with the XUV beam in a second gas jet, producing photoionization. The resulting photoelectron spectrum is generally analysed with a time-of-flight (TOF) electron spectrometer as a function of the relative delay between the IR and XUV pulses.

As for trains of attosecond pulses, isolated attosecond pulses can be characterized using a method based on the measurement of the electrons photoionized by the XUV pulse in the co-presence of an IR field. The method is dubbed attosecond streak-camera (Mairesse and Quéré, 2005). The experimental setup is very similar to what is used for RABBITT, except that the IR pulse (streaking pulse) intensity in this case is low enough not to ionize atoms but high enough to impart substantial momentum to the photoelectrons liberated by the XUV pulse. If the dipole transition matrix element does not vary significantly (in phase and amplitude) over the XUV energy range, the photoelectron wave packet can be considered as a replica of the attosecond pulse. The streaking field modulates in time the phase of the electron wave packet as:

$$\Phi(t) = -\int_t^{+\infty} dt' \left[\vec{v} \cdot \vec{A}(t') + \vec{A}^2(t')/2\right] = -\int_t^{+\infty} dt' U_p(t') + \frac{\sqrt{8WU_p}}{\omega}\cos\theta\ \cos\omega t\ -(U_p/2\omega)\ \sin 2\omega t \tag{23}$$

where ω is the frequency of the IR field, U_p is the ponderomotive potential of the IR pulse, θ is the angle between the electron velocity $\vec{v}$ and the vector potential $\vec{A}$, and $W=p^2/2$ is the kinetic energy of the emitted electron. As shown in bottom panel of **Fig. 14**, a modulation in the temporal phase corresponds to an energy shift of the photoelectron spectrum, which follows the shape of the vector potential of the IR pulse. Thus, by measuring the streaking effect on the photoelectron distribution for different time delays, it is possible to estimate the temporal duration of the XUV pulse.

The attosecond streak camera has a strong analogy to the FROG technique used to characterize femtosecond optical pulses. Indeed, the phase modulation induced by the IR streaking pulse can be seen as a phase gate allowing for a FROG-like measurement. The extension of FROG to the attosecond domain is called FROG-CRAB (FROG for Complete Reconstruction of Attosecond Bursts). The analogy with the FROG technique can be easily identified by comparing Eq. (12) with the streaking spectrogram measured in a given observation direction:

$$S(\vec{v},\tau) = \left| \int exp(i\Phi(t))\, \vec{d}_{\vec{p}-\vec{A}(t)} \cdot \vec{E}_{XUV}(t-\tau) exp(i(W+I_p)t) dt \right|^2 \tag{24}$$

where $\Phi(t)$ is the phase reported in Eq. (22), $\vec{d}$ is the dipole matrix element, I_p is the ionization potential of the medium, and $\vec{E}_{XUV}$ is the field to be characterized. The streaking spectrogram thus corresponds to a FROG trace where the temporal gate is a

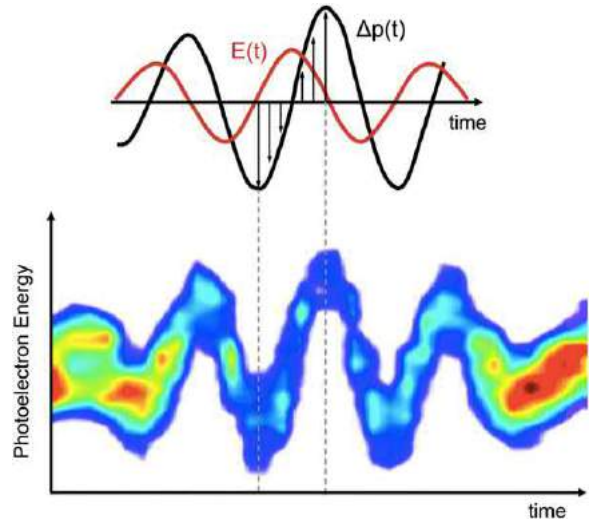

Fig. 14 Attosecond streak camera. Top panel: a photoelectron is released after XUV ionization and the IR probe pulse transfers to it an additional momentum Δp, which depends on the phase and amplitude of the IR electric field $E(t)$. Bottom panel: streaked photoelectron spectra acquired as a function of the XUV-IR time delay. The acquired spectrogram follows the shape of the vector potential of the IR pulse.

pure phase gate: $g(t)=\exp(i\Phi(t))$. Iterative inversion algorithms such as the very efficient Principal Component Generalized Projections Algorithm (PCGPA) can be used extract from the CRAB trace the spectral phase of the attosecond pulse. Alternatively to the attosecond streak camera, the Phase Retrieval by Omega Oscillation Filtering (PROOF) can be used. This method takes advantage of the same approach as RABBITT, using a weak perturbing IR field. The reconstruction procedure does not require an iterative method and it does not rely on the central momentum approximation as in the case of FROG-CRAB, thus making the technique particularly suitable for characterizing ultra-broadband attosecond pulses (Chini *et al.*, 2010).

Finally, *in situ* techniques can be used for the temporal characterization of the attosecond pulse. These techniques rely on an all-optical method – solely based on the measurement of the XUV photon spectrum – for the reconstruction of the phase of the attosecond pulse, thus overcoming the technical challenge of measuring photoelectron spectra. The typical approach to implement an *in situ* characterization is to use a weak second-harmonic perturbing field to slightly alter the XUV generation mechanisms (Kim *et al.*, 2014). By implementing methods based on this idea, both trains and isolated attosecond pulses have been fully characterized.

See also: Attosecond Spectroscopy

References

Calegari, F., *et al.*, 2016. Journal of Physics B: Atomic, Molecular and Optical Physics 49, 062001.
Cerullo, G., *et al.*, 2010. Laser and Photonics Reviews. 1–29.
Chini, M., *et al.*, 2010. Optical Express 18, 13006.
Ciriolo, A.G., *et al.*, 2017. Applied Science 2017 (7), 265.
Corkum, P., 1994. Physical Review Letters 71 (1993),
Ferrari, F., *et al.*, 2009. Nature Photonics 4, 875.
Goulielmakis, E., *et al.*, 2008. Science 320, 1614.
Kim, K.T., *et al.*, 2014. Nature Photonics 8, 187.
Lewenstein, M., *et al.*, 1994. Physical Review A 49, 2117.
Mairesse, Y., Quéré, F., 2005. Physical Review A 71, 011401(R).
Manzoni, C., *et al.*, 2015. Laser Photonics Reviews 9 (2), 129.
Paul, P., *et al.*, 2001. Science 292, 1689.
Popmintchev, T., *et al.*, 2012. Science 336, 1287.
Sansone, G., *et al.*, 2006. Science 314, 443.
Vozzi, C., *et al.*, 2005. Applied Physics B 80, 285.

Chirped Pulse Amplification

GA Mourou, University of Michigan, Ann Arbor, MI, USA

In the five years after the invention of the laser in 1960, tabletop lasers advanced in a series of technological leaps to reach a power of one gigawatt (10^9 W). For the next 20 years, progress was stymied and the maximum power of tabletop laser systems did not grow. The sole way to increase power was to build ever larger lasers. Trying to operate beyond the limiting intensity would create unwanted nonlinear effects in components of the laser, impairing the beam quality and even damaging the components. Only in 1985 was this optical damage problem circumvented with the introduction of chirped pulse amplification (CPA), a technique developed by the research group led by the author. Tabletop laser powers then leaped ahead by factors of 10^3 to 10^5. CPA is used in all ultrashort pulse amplifiers from the very small, such as the fiber-based amplifier to the extremely large, such as the ones used in inertial confinement fusion, for fast ignition.

'Chirping' a signal or a wave means stretching it in time, arranging the frequencies from blue to red or vice versa. In chirped pulse amplification, the first step is to produce a short pulse with an oscillator and stretch it, usually to 10^3 to 10^5 times as long, by using a pair of diffraction gratings (see **Fig. 1**). This operation decreases the intensity of the pulse by the same amount. Standard laser amplification techniques can now be applied to increase the laser pulse energy by up 10^3 to 10^5 times what we could achieve if we were not stretching the pulse. Finally, when the amplifier energy is fully extracted a sturdy device, such as a pair of diffraction gratings – sometimes in vacuum – recompresses the pulse to its original duration, by regrouping all the frequencies – increasing its power 10^3 to 10^5 times beyond the amplifier s limit. A typical example would begin with a seed pulse lasting 100 femtoseconds and having 0.2 nanojoule of energy. We stretch it by a factor of 10^4 to a nanosecond (reducing its power from about two kilowatts to 0.2 watt) and amplify it by ten orders of magnitude to two joules and two gigawatts. Recompressing the pulse to 100 femtoseconds increases the power to 20 terawatts. Without this method, sending the original two-kilowatt pulse through a tabletop amplifier would have destroyed the amplifier – unless we increased the amplifier's cross-sectional area 10^4 times and dispersed the beam across it. The CPA technique makes it possible to use conventional laser amplifiers and to stay below the onset of nonlinear effects.

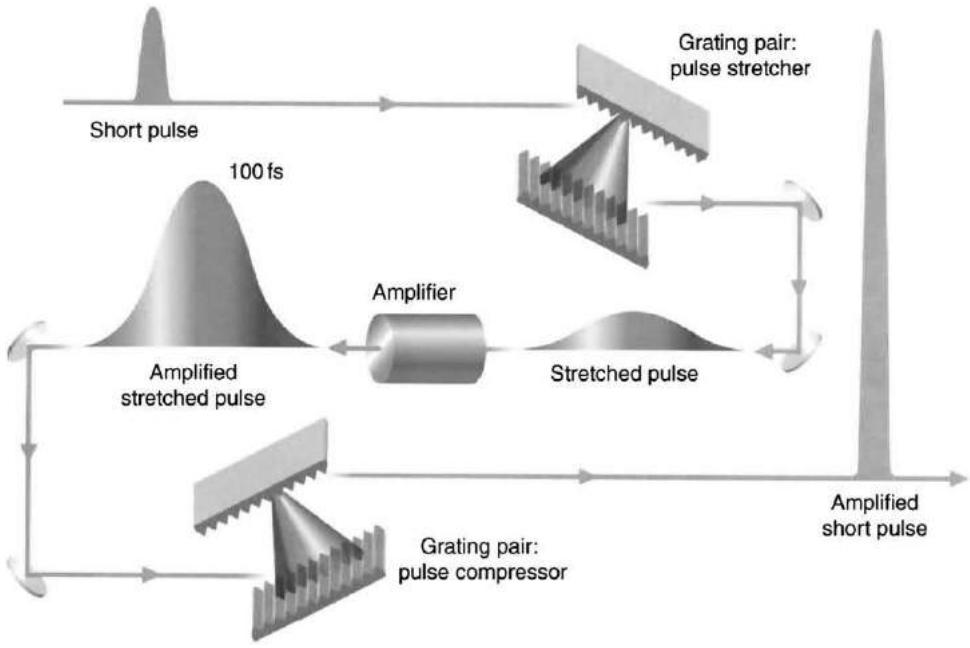

Fig. 1 The key to tabletop ultrahigh-intensity lasers is a technique called chirped pulse amplification. An initial short laser pulse is stretched out (chirped) by a factor of about 10^4, for instance, by a pair of diffraction gratings. The stretched pulse s intensity is low, allowing it to be amplified by a small laser amplifier. A second pair of gratings recompresses the pulse, boosting it to 10^4 times the peak intensity that the amplifier could have withstood.

 doi:10.1016/B978-0-12-809283-5.00775-8

Carbon Dioxide Laser

CR Chatwin, University of Sussex, Brighton, UK

Nomenclature

α The direct excitation of carbon dioxide (CO_2) ground state molecules (sec^{-1})
β The direct de-excitation of nitrogen (N_2) (sec^{-1})
γ The direct excitation of nitrogen (N_2) ground state molecules (sec^{-1})
η The direct de-excitation of carbon dioxide (CO_2) (sec^{-1})
λ wavelength (m)
σ absorption coefficient (cm^2)
τ Electrical current pulse length (μs)
τ_{sp} Spontaneous emission life time of the upper laser level (sec)
$A=(\tau_{sp}^{-1}=0.213(sec^{-1})$ The Einstein 'A' coefficient for the laser transition
c velocity of light (cm sec^{-1})
C_c Coupling capacitance (nF)
CO_2 Carbon dioxide
e subscript 'e' refers to the fact that the populations in the square brackets are the values for thermodynamic equilibrium
E Electric Field (V cm^{-1})
F_{CO_2} Fraction of the input power (IP) coupled into the excitation of the energy level E_{00^01}
F_{N_2} Fraction of the input power (IP) coupled into the excitation of the energy level $E_{v=1}$
g gain (cm^{-1})
g_1 and g_2 energy level degeneracy's of levels 1 and 2
He Helium
I beam irradiance (W cm^{-2})
I_0 beam irradiance at the center of a Gaussian laser beam (W cm^{-2})
I_p Photon population density (photons cm^{-3})
I_p Input current (A)
IP Electrical input power (W cm^{-3})
J the rotational quantum number
K_{51}, K_{15} Resonant energy transfer between the CO_2 (00^01) and N_2 ($v=2$) energy levels proceeds via excited molecules colliding with ground state molecules (sec^{-1})
K_{132}, K_{2131}, K_{2231} are vibration/vibration transfer rates (sec^{-1}) between energy levels 1 and 32; 21 and 31; 22 and 31, respectively (see Fig. 1)
K_{320} is a vibration/translation transfer rate (sec^{-1}) between energy levels 32 and 0 (see Fig. 1)
K_{sp}, A Spontaneous emission rate (sec^{-1})
L The distance between the back and the front mirrors, which have reflectivity's of R_B and R_F (cm)
n molecular population (molecules cm^{-3})
N_2 Nitrogen
P Pressure (Torr)
P_{CO_2}, P_{He} and P_{N_2} The respective gas partial pressures (Torr)
P_{in} Electrical input power (kW)
r radius of laser beam (cm)
R_B Back mirror reflectivity
R_F Front mirror reflectivity
t time (sec)
T Temperature (deg K)
T_0 the photon decay time (sec)
w the radial distance where the power density is decreased to $1/e^2$ of its axial value

Introduction

This article gives a brief history of the development of the laser and goes on to describe the characteristics of the carbon dioxide laser and the molecular dynamics that permit it to operate at comparatively high power and efficiency. It is these commercially attractive features and its low cost that has led to its adoption as one of most popular industrial power beams. This outline also describes the main types of carbon dioxide laser and briefly discusses their characteristics and uses.

Brief History

In 1917 Albert Einstein developed the concept of stimulated emission which is the phenomenon used in lasers. In 1954 the MASER (Microwave Amplification by Stimulated Emission of Radiation) was the first device to use stimulated emission. In that year Townes and Schawlow suggested that stimulated emission could be used in the infrared and optical portions of the electromagnetic spectrum. The device was originally termed the optical maser, this term being dropped in favor of LASER, standing for: Light Amplification by Stimulated Emission of Radiation. Working against the wishes of his manager at Hughes Research Laboratories, the electrical engineer Ted Maiman created the first laser on the 15 May 1960. Maiman's flash lamp pumped ruby laser produced pulsed red electromagnetic radiation at a wavelength of 694.3 nm. During the most active period of laser systems discovery Bell Labs made a very significant contribution. In 1960, Ali Javan, William Bennet and Donald Herriot produced the first Helium Neon laser, which was the first continuous wave (CW) laser operating at 1.15 μm. In 1961, Boyle and Nelson

developed a continuously operating Ruby laser and in 1962, Kumar Patel, Faust, McFarlane and Bennet discovered five noble gas lasers and lasers using oxygen mixtures. In 1964, C.K.N. Patel created the high-power carbon dioxide laser operating at 10.6 μm. In 1964, J.F. Geusic and R.G. Smith produced the first Nd:Yag laser using neodymium doped yttrium aluminum garnet crystals and operating at 1.06 μm.

Characteristics

Due to its operation between low lying vibrational energy states of the CO_2 molecule, the CO_2 laser has a high quantum efficiency, ~40%, which makes it extremely attractive as a high-power industrial materials processing laser (1 to 20 kW), where energy and running costs are a major consideration. Due to the requirement for cooling to retain the population inversion, the efficiency of electrical pumping and optical losses – commercial systems have an overall efficiency of approximately 10%. Whilst this may seem low, for lasers this is still a high efficiency. The CO_2 laser is widely used in other fields, for example, surgical applications, remote sensing, and measurement. It emits infrared radiation with a wavelength that can range from 9 μm up to 11 μm. The laser transition may occur on one of two transitions: $(00^01)\rightarrow(10^00)$, $\lambda=10.6$ μm; $(00^01)\rightarrow(02^00)$, $\lambda=9.6$ μm, see **Fig. 1**. The 10.6 μm transition has the maximum probability of oscillation and gives the strongest output; hence, this is the usual wavelength of operation, although for specialist applications the laser can be forced to operate on the 9.6 μm line.

Fig. 1 illustrates an energy level diagram with four vibrational energy groupings that include all the significantly populated energy levels. The internal relaxation rates within these groups are considered to be infinitely fast when compared with the rate of energy transfer between these groups. In reality the internal relaxation rates are at least an order of magnitude greater than the rates between groups.

Excitation of the upper laser level is usually provided by an electrical glow discharge. However, gas dynamic lasers have been built where expanding a hot gas through a supersonic nozzle creates the population inversion; this creates a nonequilibrium region in the downstream gas stream with a large population inversion, which produces a very high-power output beam (135 kW – Avco Everett Research Lab). For some time the gas dynamic laser was seriously considered for use in the space-based Strategic Defence Initiative (SDI-USA). The gas mixture used in a CO_2 laser is usually a mixture of carbon dioxide, nitrogen, and helium. The proportions of these gases varies from one laser system to another, however, a typical mixture is 10%-CO_2; 10%-N_2; 80%-He. Helium plays a vital role in the operation of the CO_2 laser in that it maintains the population inversion by depopulating the lower laser level by nonradiative collision processes. Helium is also important for stabilization of the gas discharge; furthermore it greatly improves the thermal conductivity of the gas mixture, which assists in the removal of waste heat via heat exchangers.

Small quantities of other gases are often added to commercial systems in order to optimize particular performance characteristics or stabilize the gas discharge; for brevity we only concern ourselves here with this simple gas mixture.

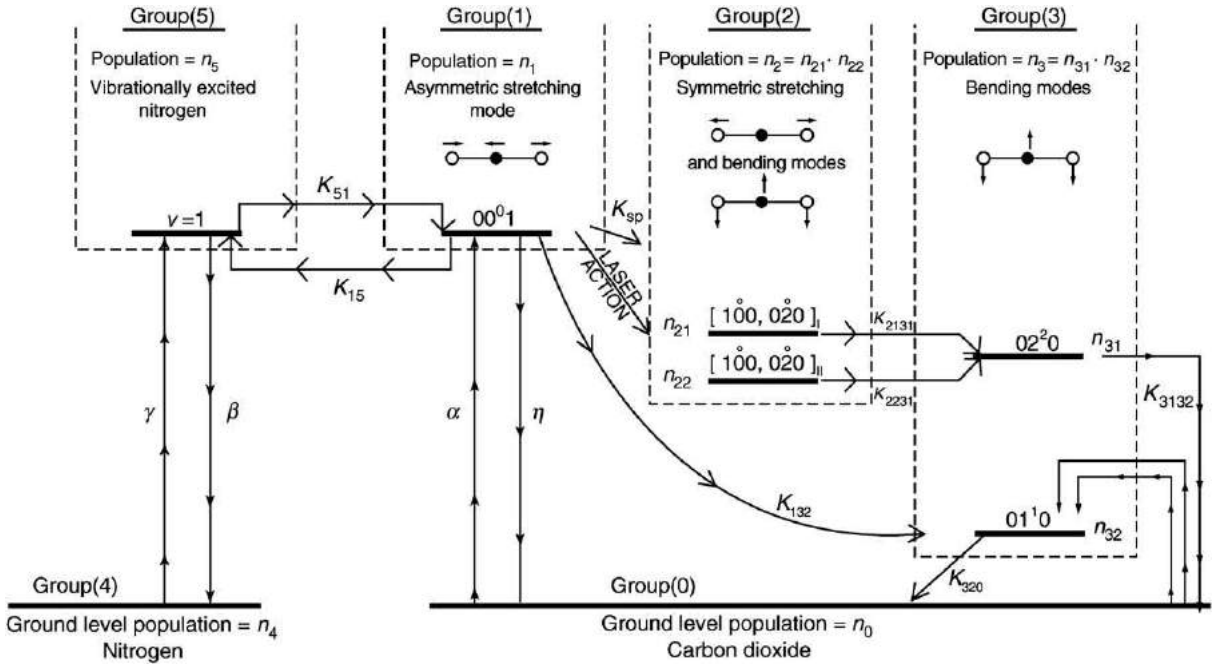

Fig. 1 Six level model used for the theoretical description of CO_2 laser action.

Molecular Dynamics

Direct Excitation and De-excitation

It is usual for excitation to be provided by an electrical glow discharge. The direct excitation of carbon dioxide (CO_2) and nitrogen (N_2) ground state molecules proceeds via inelastic collisions with fast electrons. The rates of kinetic energy transfer are α and γ, respectively, and are given by Eqs. (1) and (2):

$$\alpha = \frac{F_{CO_2} \times IP}{E_{00^01} \times n_0} (sec^{-1}) \tag{1}$$

$$\gamma = \frac{F_{N_2} \times IP}{E_{v=1} \times n_4} (sec^{-1}) \tag{2}$$

where:

F_{CO_2} = Fraction of the input power (IP) coupled into the excitation of the energy level E_{00^01}, n_0 is the CO_2 ground level population density;
F_{N_2} = Fraction of the input power (IP) coupled into the excitation of the energy level $E_{v=1}$; and n_4 is the N_2 ground level population density.

The reverse process of the above occurs when molecules lose energy to the electrons and the electrons gain an equal amount of kinetic energy; the direct de-excitation rates are given by η and β, Eqs. (3) and (4), respectively:

$$\eta = \alpha \times \exp\left(\frac{E_{00^01}}{E_e}\right) (sec^{-1}) \tag{3}$$

$$\beta = \gamma \times \exp\left(\frac{E_{v=1}}{E_e}\right) (sec^{-1}) \tag{4}$$

where E_e is the average electron energy in the discharge.

E_e, F_{CO_2} and F_{N_2} are obtained by solution of the Boltzmann transport equation (BTE); the average electron energy can be optimized to maximize the efficiency (F_{CO_2}, F_{N_2}) with which electrical energy is utilized to create a population inversion. Hence, the discharge conditions required to maximize efficiency can be predicted from the transport equation. Fig. 2 shows one solution of the BTE for the electron energy distribution function.

Resonant Energy Transfer

Resonant energy transfer between the CO_2 (00^01) and N_2 $v=2$ energy levels (denoted 1 and 5 in Fig. 1) proceeds via excited molecules colliding with ground state molecules. A large percentage of the excitation of the upper laser level takes place via collisions between excited N_2 molecules and ground state CO_2 molecules. The generally accepted rate of this energy transfer is

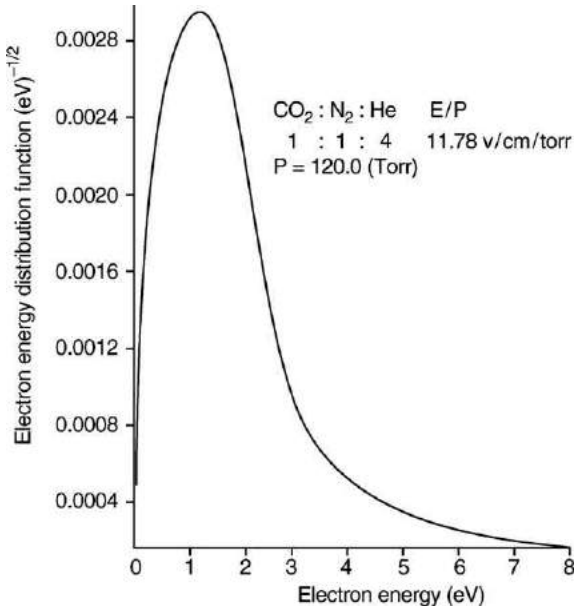

Fig. 2 Electron energy distribution function.

given by Eqs. (5) and (6):

$$K_{51} = 19\,000 P_{CO_2}(\text{sec}^{-1}) \tag{5}$$

$$K_{15} = 19\,000 P_{N_2}(\text{sec}^{-1}) \tag{6}$$

where P_{CO_2} and P_{N_2} are the respective gas partial pressures in Torr.

Hence, CO_2 molecules are excited into the upper laser level by both electron impact and impact with excited N_2 molecules. The contribution from N_2 molecules can be greater than 40% depending on the discharge conditions.

Collision–Induced Vibrational Relaxation of the Upper and Lower Laser Levels

The important vibrational relaxation processes are illustrated by Fig. 1 and can be evaluated from Eqs. (7–10); where the subscripts refer to the rate between energy levels 1 and 32; 21 and 31; 22 and 31; 32 and 0, respectively:

$$K_{132} = 367 P_{CO_2} + 110 P_{N_2} + 67 P_{He}(\text{sec}^{-1}) \tag{7}$$

$$K_{2131} = 6 \times 10^5 P_{CO_2}(\text{sec}^{1}) \tag{8}$$

$$K_{2231} = 5.15 \times 10^5 P_{CO_2}(\text{sec}^{-1}) \tag{9}$$

$$K_{320} = 200 P_{CO_2} + 215 P_{N_2} + 3270 P_{He}(\text{sec}^{-1}) \tag{10}$$

K_{132}, K_{2131}, and K_{2231} are vibration/vibration transfer rates and K_{320} is a vibration/translation transfer rate. Note the important effect of helium on Eq. (10); helium plays a major role in depopulating the lower laser level, thus enhancing the population inversion. P_{He} is the partial pressure of helium in Torr.

Radiative Relaxation

Spontaneous radiative decay is not a major relaxation process in the CO_2 laser but it is responsible for starting laser action via spontaneous emission. The Einstein 'A' coefficient for the laser transition is given by Eq. [11]:

$$A = (\tau_{sp})^{-1} = 0.213(\text{sec}^{-1}) \tag{11}$$

Gain

The gain (g) is evaluated from the product of the absorption coefficient (σ) and the population inversion, Eq. (12):

$$g = \sigma\left(n_{00^0 1} - \frac{g_1}{g_2} n_{10^0 0}\right)\text{cm}^{-1} \tag{12}$$

For most commercial laser systems the absorption coefficient is that for high-pressure collision-broadening ($P > 5.2$ Torr) where the intensity distribution function describing the line shape is Lorentzian. The following expression describes the absorption coefficient, Eq. (13):

$$\sigma = \frac{692.5}{T n_{CO_2}\left(1 + 1.603\frac{n_{N_2}}{n_{CO_2}} + 1.4846\frac{n_{He}}{n_{CO_2}}\right)}(\text{cm}^2) \tag{13}$$

where T is the absolute temperature and n refers to the population density of the gas designated by the subscript.

This expression takes account of the different constituent molecular velocity distributions and different collision cross-sections for $CO_2 \rightarrow CO_2$, $N_2 \rightarrow CO_2$ and $He \rightarrow CO_2$ type collisions. Eq. (13) also takes account of the significant line broadening effect of helium.

Neglecting the unit change in rotational quantum number, the energy level degeneracies g_1 and g_2 may be dropped. $n_{10^0 0}$ is partitioned such that $n_{10^0 0} = 0.145 n_2$ and Eq. (12) can be re-cast as Eq. (14):

$$g = \sigma(n_1 - 0.1452 n_2)\text{cm}^{-1} \tag{14}$$

where n_1 and n_2 are the population densities of energy groups '1' and '2' respectively.

Stimulated Emission

Consider a laser oscillator with two plane mirrors, one placed at either end of the active gain medium, with one mirror partially transmitting (see Fig. 4). Laser action is initiated by spontaneous emission that happens to produce radiation whose direction is normal to the end mirrors and falls within the resonant modes of the optical resonator. The rate of change of photon population density (I_p within the laser cavity can be written as Eq. (15):

$$\frac{dI_P}{dt} = I_p c g - \frac{I_p}{T_0} \tag{15}$$

where the first term on the right-hand side accounts for the effect of stimulated emission and the second term represents the number of photons that decay out of the laser cavity, T_0 is the photon decay time, given by Eq. (16), and is defined as the average time a photon remains inside the laser cavity before being lost either through the laser output window or due to dispersion; if dispersion is ignored, I_p/T_0, is the laser output:

$$T_0 = \frac{2L}{c \log_e\left(\frac{1}{R_B R_F}\right)} \tag{16}$$

where L is the distance between the back and the front mirrors, which have reflectivities of R_B and R_F, respectively. The dominant laser emission occurs on a rotational–vibrational P branch transition $P(22)$, that is $(J=21) \rightarrow (J=22)$ line of the $(00^01) \rightarrow (10^00)$, $\lambda = 10.6$ μm transition, where J is the rotational quantum number. The rotational level relaxation rate is so rapid that equilibrium is maintained between rotational levels so that they feed all their energy through the $P(22)$ transition. This model simply assumes constant intensity, basing laser performance on the performance of an average unit volume. By introducing the stimulated emission term into the molecular rate equations, which describe the rate of transfer of molecules between the various energy levels illustrated in Fig. 1, a set of molecular rate Eqs. (17–21) can be written that permit simulation of the performance of a carbon dioxide laser:

$$\frac{dn_1}{dt} = \alpha n_0 - \eta n_1 + K_{51} n_5 - K_{15} n_1 - K_{sp} n_1 - K_{132}\left(n_1 - \left[\frac{n_1}{n_{32}}\right]_e n_{32}\right) - I_p cg \tag{17}$$

$$\frac{dn_2}{dt} = K_{sp} n_1 + I_p cg - K_{2131}\left(n_{21} - \left[\frac{n_{21}}{n_{31}}\right]_e n_{31}\right) - K_{2231}\left(n_{22} - \left[\frac{n_{22}}{n_{31}}\right]_e n_{31}\right) \tag{18}$$

$$\frac{dn_3}{dt} = 2K_{2131}\left(n_{21} - \left[\frac{n_{21}}{n_{31}}\right]_e n_{31}\right) - 2K_{2231}\left(n_{22} - \left[\frac{n_{22}}{n_{31}}\right]_e n_{31}\right) + K_{132}\left(n_1 - \left[\frac{n_1}{n_{32}}\right]_e n_{32}\right) - K_{320}\left(n_{32} - \left[\frac{n_{32}}{n_0}\right]_e n_0\right) \tag{19}$$

$$\frac{dn_5}{dt} = \gamma n_4 - \beta n_5 - K_{51} n_5 + K_{15} n_1 \tag{20}$$

$$\frac{dI_P}{dt} = I_p cg - \frac{I_p}{T_0} \tag{21}$$

The terms in square brackets ensure that the system maintains thermodynamic equilibrium; subscript 'e' refers to the fact that the populations in the square brackets are the values for thermodynamic equilibrium. The set of five simultaneous differential equations can be solved using a Runge–Kutta method. They can provide valuable performance prediction data that is helpful in optimizing laser design, especially when operated in the pulsed mode. Fig. 3(a) illustrates some simulation results for the transverse flow laser shown in Fig. 9. The results illustrate the effect of altering the gas mixture and how this can be used to control the gain switched spike that would result in unwelcome work piece plasma generation if allowed to become too large. Fig. 3(b) shows an experimental laser output pulse from the high pulse repetition frequency (prf–5kHz) laser illustrated in Fig. 9. This illustrates that even a quite basic physical model can give a good prediction of laser output performance.

Optical Resonator

Fig. 4 shows a simple schematic of an optical resonator. This simple optical system consists of two mirrors (the full reflector is often a water cooled gold coated copper mirror), which are aligned to be orthogonal to the optical axis that runs centrally along the length of the active gain medium in which there is a population inversion. The output coupler is a partial reflector (usually dielectrically coated zinc selenide – ZnSe–that may be edge cooled) so that some of the electromagnetic radiation can escape as an output beam. The ZnSe output coupler has a natural reflectivity of about 17% at each air–solid interface. For high power lasers (2 kW) 17% is sufficient for laser operation; however, depending on the required laser performance the inside surface is often given a reflective coating. The reflectivity of the inside face depends on the balance between the gain (Eq. (14)), the output power and the power stability requirements. The outside face of the partial reflector must be anti-reflection (AR) coated.

Spontaneous emission occurs within the active gain medium and radiates randomly in all directions; a fraction of this spontaneous emission will be in the same direction as the optical axis, perpendicular to the end mirrors, and will also fall into a resonant mode of the optical resonator. Spontaneous emission photons interact with CO_2 molecules in the excited upper laser level, excited state (00^01), which stimulates these molecules to give up a quanta of vibrational energy as photons via the radiative transition $(00^01 \rightarrow (10^00$, $\lambda = 10.6$μm. The radiation given up will have exactly the same phase and direction as the stimulating radiation and thus will be coherent with it. The reverse process of absorption also occurs, but so long as there is a population inversion there will be a net positive output. This process is called light amplification by stimulated emission of radiation (LASER). The mirrors continue to redirect the photons parallel to the optical axis and so long as the population inversion is not depleted, more and more photons are stimulated by stimulated emission which dominates the process and also dominates the spontaneous emission, which is important to initiate laser action.

Light emitted by lasers contains several optical frequencies, which are a function of the different modes of the optical resonator; these are simply the standing wave patterns that can exist within the resonator structure. There are two types of resonator modes:

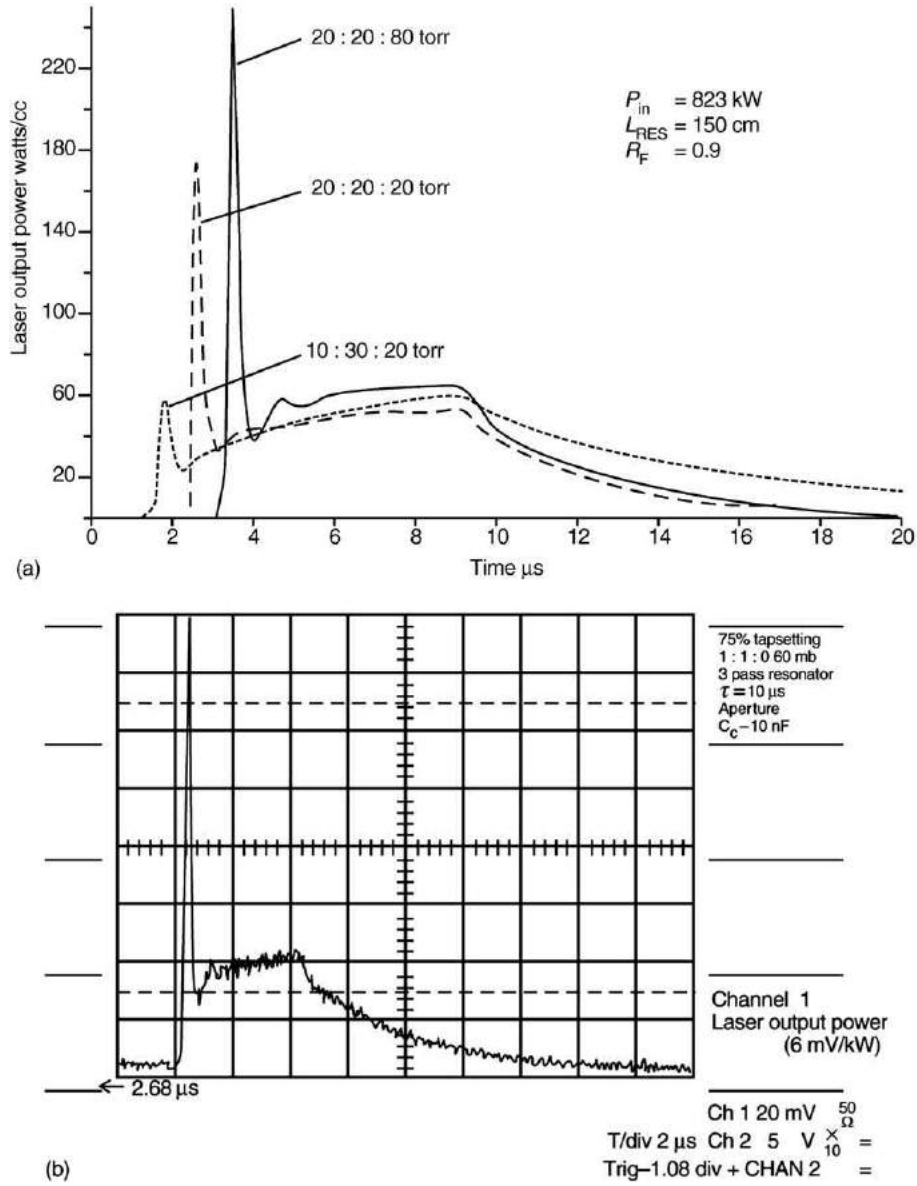

Fig. 3 (a) Predicted output pulses for transverse flow CO_2 laser for different gas mixtures, (b) Experimental output pulse from transverse flow CO_2 laser.

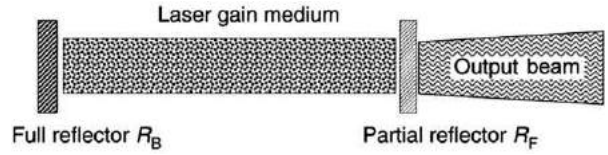

Fig. 4 Optical resonator.

longitudinal and transverse. Longitudinal modes differ from one another in their frequency of oscillation whereas transverse modes differ from one another in their oscillation frequency and field distribution in a plane orthogonal to the direction of propagation. Typically CO_2 lasers have a large number of longitudinal modes; in CO_2 laser applications these are of less interest than the transverse modes, which determine the transverse beam intensity and the nature of the beam when focused. In cylindrical coordinates the transverse modes are labelled TEM_{pl}, where subscript '*p*' is the number of radial nodes and '*l*' is the number of angular nodes. The lowest order mode is the TEM_{00}, which has a Gaussian-like intensity profile with its maximum on the beam axis. A light beam emitted from an optical resonator with a Gaussian profile is said to be operating in the 'fundamental

mode' or the TEM_{00} mode. The decrease in irradiance (I) with distance 'r' from the axis (I_0) of the Gaussian beam is described by Eq. (22):

$$I(r) = I_0 \exp(2r^2/w^2) \tag{22}$$

where w is the radial distance, where the power density is decreased to $1/e^2$ of its axial value. Ideally a commercial laser should be capable of operation in the fundamental mode as, with few exceptions, this results in the best performance in applications. Laser cutting benefits from operation in the fundamental mode; however, welding or heat treatment applications may benefit from operation with higher-order modes. Output beams are usually controlled to be linearly or circularly polarized, depending upon the requirements of the application. For materials processing applications the laser beam is usually focused via a water cooled ZnSe lens or, for very high power lasers, a parabolic gold coated mirror. Welding applications will generally use a long focal length lens and cutting applications will use a short focal length, which generates a higher irradiance at the work piece than that necessary for welding. The beam delivery optics are usually incorporated into a nozzle assembly that can deliver cooling water and assist gases for cutting and anti-oxidizing shroud gases for welding or surface engineering applications.

Laser Configuration

CO_2 lasers are available in many different configurations and tend to be classified on the basis of their physical form and the gas flow arrangement, both of which greatly affect the output power available and beam quality. The main categories are: sealed off lasers, waveguide lasers, slow axial flow, fast axial flow, diffusion cooled, transverse flow, transversely excited atmospheric lasers, and gas dynamic lasers.

Sealed-Off and Waveguide Lasers

Depopulation of the lower laser level is via collision with the walls of the discharge tube, so the attainable output power scales with the length of the discharge column and not its diameter. Output powers are in the range 5 W to 250 W. Devices may be constructed from concentric glass tubes with the inner tube providing the discharge cavity and the outer tube acting to contain water-cooling of the inner discharge tube. The inner tube walls act as a heat sink for the discharge thermal energy (see Fig. 5). The DC electrical discharge is provided between a cathode and anode situated at either end of the discharge tube. A catalyst must be provided to ensure regeneration of CO_2 from CO. This may be accomplished by adding about 1% of H_2O to the gas mixture, or alternatively, recombination can be achieved via a hot (300 °C) Ni cathode, which acts as a catalyst. RF-excited all metal sealed-off tube systems can deliver lifetimes greater than 45 000 hours.

Diffusion-cooled slab laser technology will also deliver reliable sealed operation for 20 000 hrs. Excitation of the laser medium occurs via RF excitation between two water-cooled electrodes. The water-cooled electrodes dissipate (diffusion cooled) the heat generated in the gas discharge. An unstable optical resonator provides the output coupling for such a device (see Fig. 6). Output powers are in the range 5 W to 300 W and can be pulsed from 0 to 100 kHz. These lasers are widely used for marking, rapid prototyping, and cutting of nonmetals (paper, glass, plastics, ceramics) and metals.

Waveguide CO_2 lasers use small bore tubes (2–4 mm) made of BeO or SiO_2 where the laser radiation is guided by the tube walls. Due to the small tube diameter, a gas total pressure of 100 to 200 Torr is necessary, hence the gain per unit length is high. This type of laser will deliver 30 W of output power from a relatively short (50 cm long) compact sealed-off design; such a system is useful for microsurgery and scientific applications. Excitation can be provided from a longitudinal DC discharge or from an RF source that is transverse to the optical axis; RF excitation avoids the requirement for an anode and cathode and results in a much lower electrode voltage.

Slow Axial Flow Lasers

In slow flow lasers the gas mixture flows slowly through the laser cavity. This is done to remove the products of dissociation that will reduce laser efficiency or prevent it from operating at all, and the main contaminant is CO. The dissociated gases (mainly CO and O_2) can be recombined using a catalyst pack and then reused via continuous recirculation. Heat is removed via

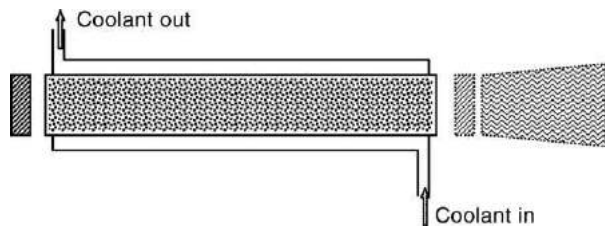

Fig. 5 Schematic of sealed-off CO_2 laser, approximately 100 W per meter of gain length, gas cooled by diffusion to the wall.

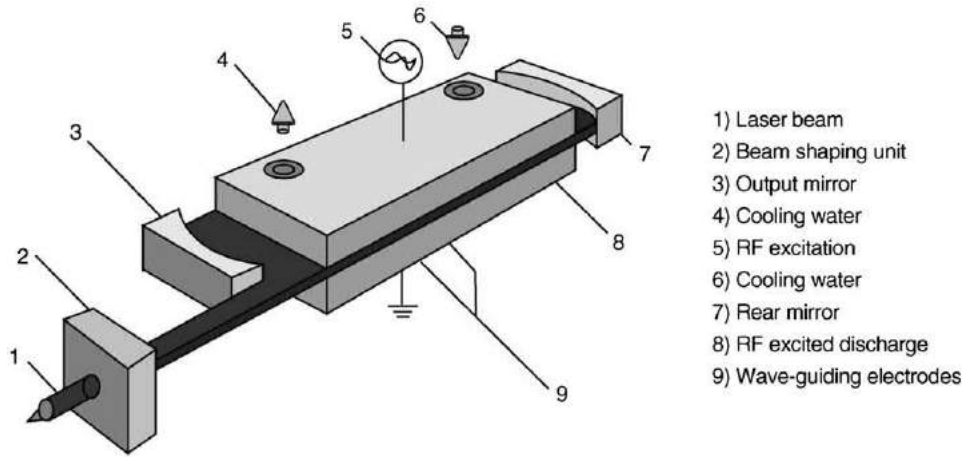

Fig. 6 Schematic for sealed-off slab laser and diffusion cooled laser with RF excitation (courtesy of Rofin).

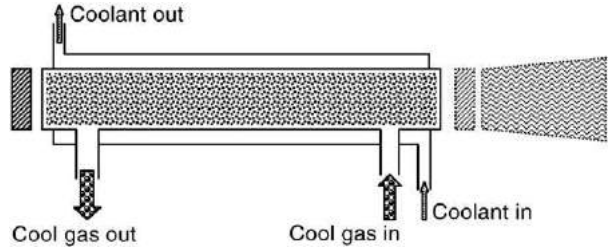

Fig. 7 Slow flow CO_2 laser, approximately 100 W per meter of gain length, gas cooled by diffusion to the wall.

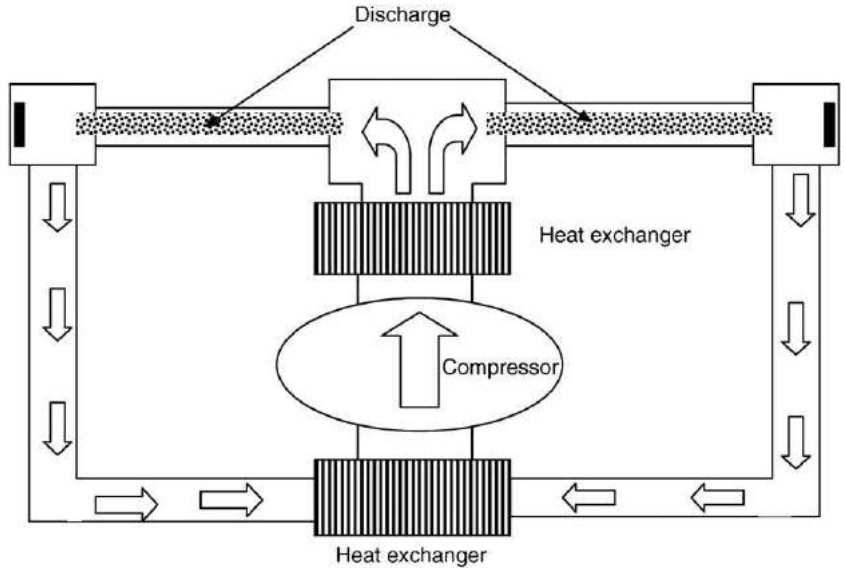

Fig. 8 Fast axial flow carbon dioxide laser.

diffusion through the walls of the tube containing the active gain medium. The tube is frequently made of Pyrex glass with a concentric outer tube to facilitate water-cooling of the laser cavity (see **Fig. 7**). Slow flow lasers operate in the power range 100 W to 1500 W, and tend to use a longitudinal DC electrical discharge which can be made to run continuously or pulsed if a thyratron switch is build into the power supply; alternatively, electrical power can be supplied via transverse RF excitation. The power scales with length, hence high power slow flow lasers have long cavities and require multiple cavity folds in order to reduce their physical size.

Fast Axial Flow Lasers

The fast axial flow laser, Fig. 8, can provide output powers from 1 kW to 20 kW; it is this configuration that dominates the use of CO_2 lasers for industrial applications. Industrial lasers are usually in the power range 2–4 kW. The output power from these devices scales with mass flow, hence the gas mixture is recycled through the laser discharge region at sonic or supersonic velocities. Historically this was achieved using Rootes blowers to compress the gas upstream of the laser cavity. Rootes compressors are inherently inefficient and the more advanced laser systems utilize turbine compressors, which deliver greater efficiency and better laser stability. Rootes compressors can be a major source of vibration. With this arrangement heat exchangers are required to remove heat after the laser discharge region and also after the compressor stage, as the compression process heats up the laser gases. Catalyst packs are used to regenerate gases but some gas replacement is often required. These laser systems have short cavities and use folded stable resonator designs to achieve higher output powers with extremely high-quality beams that are particularly suitable for cutting applications. They also give excellent results when used for welding and surface treatments.

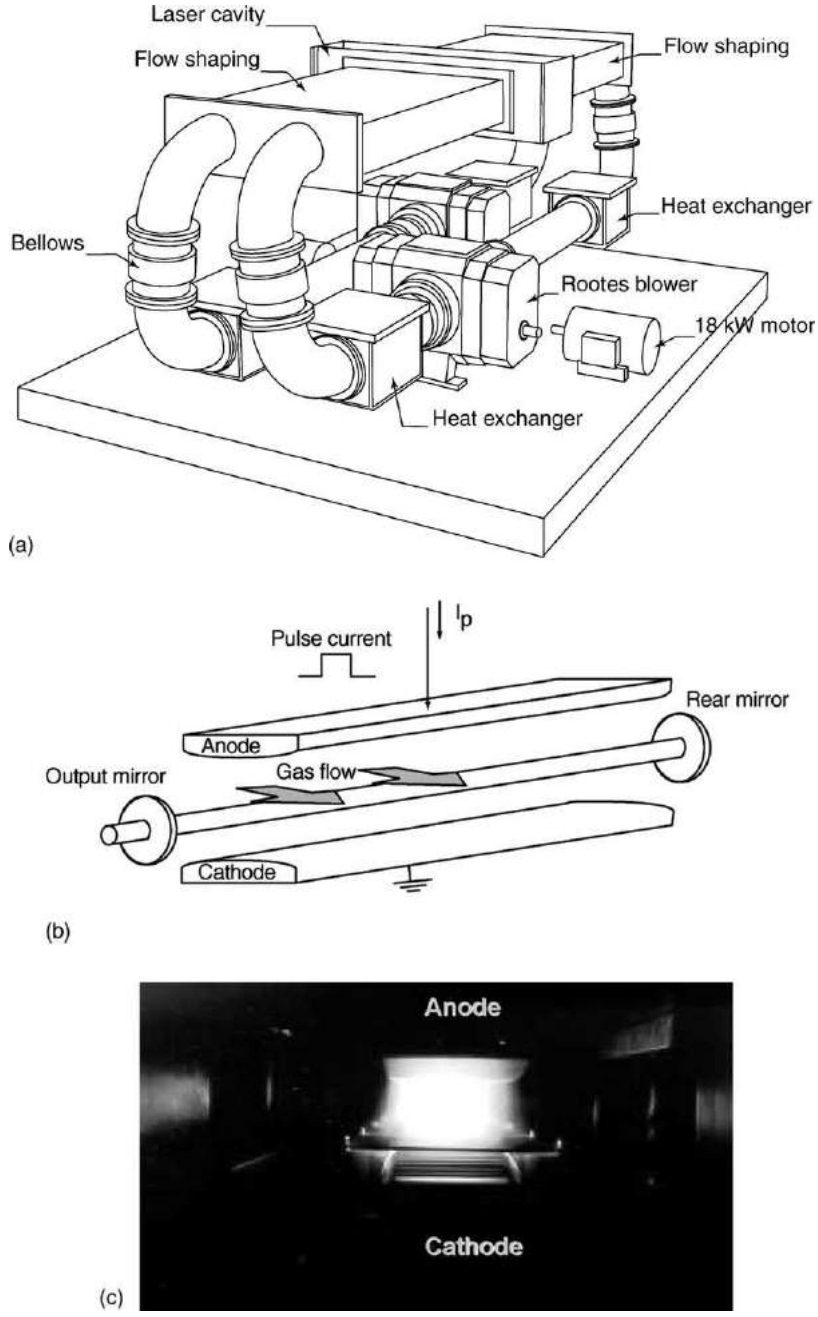

Fig. 9 (a) Transverse flow carbon dioxide laser gas recirculator, (b) Transverse flow carbon dioxide electrodes, (c) Transverse flow carbon dioxide gas discharge as seen from the output window.

Fast axial flow lasers can be excited by a longitudinal DC discharge or a transverse RF discharge. Both types of electrical excitation are common. For materials processing applications it is often important to be able to run a laser in continuous wave (CW) mode or as a high pulse repetition rate (prf) pulsed laser and to be able to switch between CW and pulsed in real time; for instance, laser cutting of accurate internal corners is difficult using CW operation but very easy using the pulsed mode of operation. Both methods of discharge excitation can provide this facility.

Diffusion Cooled Laser

The diffusion-cooled slab laser is RF excited and gives an extremely compact design capable of delivering 4.5 kW pulsed from 8 Hz to 5 kHz prf or CW with good beam quality (see **Fig. 6**). The optical resonator is formed by the front and rear mirrors and two parallel water cooled RF-electrodes. Diffusion cooling is provided by the RF-electrodes, removing the requirement for conventional gas recirculation via Rootes blowers or turbines. This design of laser results in a device with an extremely small footprint that has low maintenance and running costs. Applications include: cutting, welding, and surface engineering.

Fast Transverse Flow Laser

In the fast transverse flow laser (**Fig. 9(a)**) the gas flow, electrical discharge, and the output beam are at right angles to each other (**Fig. 9(b)**). The transverse discharge can be high voltage DC, RF, or pulsed up to 8 kHz (**Fig. 9(c)**). Very high output power per unit discharge length can be obtained with an optimal total pressure (P) of $\sim$100 Torr; systems are available delivering 10 kW of output power, CW or pulsed (see **Figs. 3(a)** and **(b)**). The increase in total pressure requires a corresponding increase in the gas discharge electric field, E, as the ratio E/P must remain constant, since this determines the temperature of the discharge electrons, which have an optimum mean value (optimum energy distribution, **Fig. 2**) for efficient excitation of the population inversion. With this high value of electric field, a longitudinal-discharge arrangement is impractical (500 kV for a 1 m discharge length); hence, the discharge is applied perpendicular to the optical axis. Fast transverse flow gas lasers provided the first multikilowatt outputs but tend to be expensive to maintain and operate. In order to obtain a reasonable beam quality, the output coupling is often obtained using a multipass unstable resonator. As the population inversion is available over a wide rectangular cross-section, this is a disadvantage of this arrangement and beam quality is not as good as that obtainable from fast axial flow designs. For this reason this type of laser is suitable for a wide range of welding and surface treatment applications.

Transversely Excited Atmospheric (TEA) Pressure

If the gas total pressure is increased above $\sim$100 Torr it is difficult to sustain a stable glow discharge, because above this pressure instabilities degenerate into arcs within the discharge volume. This problem can be overcome by pulse excitation; using sub-microsecond pulse duration, instabilities do not have sufficient time to develop; hence, the gas pressure can be increased above atmospheric pressure and the laser can be operated in a pulsed mode. In a mode locked format optical pulses shorter than 1 ns can be produced. This is called a TEA laser and with a transverse gas flow is capable of producing short high power pulses up to a few kHz repetition frequency. In order to prevent arc formation, TEA lasers usually employ ultraviolet or e-beam preionization of the gas discharge just prior to the main current pulse being applied via a thyratron switch. Output coupling is usually via an unstable resonator. TEA lasers are used for marking, remote sensing, range-finding, and scientific applications.

Conclusions

It is 40 years since Patel operated the first high power CO_2 laser. This led to the first generation of lasers which were quickly exploited for industrial laser materials processing, medical applications, defense, and scientific research applications; however, the first generation of lasers were quite unreliable and temperamental. After many design iterations, the CO_2 laser has now matured into a reliable, stable laser source available in many different geometries and power ranges. The low cost of ownership of the latest generation of CO_2 laser makes them a very attractive commercial proposition for many industrial and scientific applications. Commercial lasers incorporate many novel design features that are beyond the scope of this article and are often peculiar to the particular laser manufacturer. This includes gas additives and catalysts that may be required to stabilize the gas discharge of a particular laser design; it is this optimization of the laser design that has produced such reliable and controllable low-cost performance from the CO_2 laser.

Further Reading

Anderson, J.D., 1976. Gasdynamic Lasers: An Introduction. New York: Academic Press.
Chatwin, C.R., McDonald, D.W., Scott, B.F., 1991. Design of a High P.R.F. Carbon Dioxide Laser for Processing High Damage Threshold Materials. In: Selected Papers on Laser Design, SPIE Milestone Series., Washington: SPIE Optical Engineering Press.
Cool, A.C., 1969. Power and gain characteristics of high speed flow lasers. Journal of Applied Physics 40 (9), 3563.
Crafer, R.C., Gibson, A.F., Kent, M.J., Kimmit, M.F., 1969. Time-dependent processes on CO_2 laser amplifiers. British Journal of Applied Physics 2 (2), 183.

Gerry, E.T., Leonard, A.D., 1966. Measurement of 10.6-μ CO_2 laser transition probability and optical broadening cross sections. Applied Physics Letters 8 (9), 227.
Gondhalekar, A., Heckenberg, N.R., Holzhauer, E., 1975. The mechanism of single-frequency operation of the hybrid CO_2 laser. IEEE Journal of Quantum Electronics QE-11 (3), 103.
Gordiets, B.F., Sobolev, N.N., Shelepin, L.A., 1968. Kinetics of physical processes in CO_2 lasers. Soviet Physics JETP 26 (5), 1039.
Herzberg, G., 1945. Molecular Spectra & Molecular Structure, Infra-red and Raman Spectra of Polyatomic Molecules, vol. 2. New York: Van Nostrand.
Hoffman, A.L., Vlases, G.C., 1972. A simplified model for predicting gain, saturation and pulse length for gas dynamic lasers. IEEE Journal of Quantum Electronics 8 (2), 46.
Johnson, D.C., 1971. Excitation of an atmospheric pressure CO_2–N_2–He laser by capacitor discharges. IEEE Journal of Quantum Electronics Q.E.-7 (5), 185.
Koechner, W., 1988. Solid State Laser Engineering. Berlin: Springer Verlag.
Kogelnik, H., Li, T., 1966. Laser beams and resonators. Applied Optics 5, 1550–1567.
Levine, A.K., De Maria, A.J., 1971. Lasers, vol. 3. New York: Marcel Dekkar, Chapter 3.
Moeller, G., Ridgen, J.D., 1965. Applied Physics Letters 7, 274.
Moore, C.B., Wood, R.E., Bei-Lok, H., Yardley, J.T., 1967. Vibrational energy transfer in CO_2 lasers. Journal of Chemical Physics 11, 4222.
Patel, C.K.N., 1964. Physics Review Letters 12, 588.
Siegman, A., 1986. Lasers. Mill Valley, California: University Science.
Smith, K., Thomson, R.M., 1978. Computer Modeling of Gas Lasers. New York and London: Plenum Press.
Sobolev, N.N., Sokovikov, V.V., 1967. CO_2 lasers. Soviet Physics USPEKHI 10 (2), 153.
Svelto, O., 1998. Principles of Lasers, 4th edn New York: Plenum.
Tychinskii, V.P., 1967. Powerful lasers. Soviet Physics USPEKHI 10 (2), 131.
Vlases, G.C., Money, W.M., 1972. Numerical modelling of pulsed electric CO_2 lasers. Journal of Applied Physics 43 (4), 1840.
Wagner, W.G., Haus, H.A., Gustafson, K.T., 1968. High rate optical amplification. IEEE Journal of Quantum Electronics Q.E.-4, 287.
Witteman, W., 1987. The CO_2 Laser, Springer Series in Optical Sciences, vol. 53. .

Dye Lasers

FJ Duarte, Eastman Kodak Company, New York, NY, USA
A Costela, Consejo Superior de Investigaciones Cientificas, Madrid, Spain

Introduction

Background

Dye lasers are the original tunable lasers. Discovered in the mid-1960s these tunable sources of coherent radiation span the electromagnetic spectrum from the near-ultraviolet to the near-infrared (Fig. 1). Dye lasers spearheaded and sustained the revolution in atomic and molecular spectroscopy and have found use in many and diverse fields from medical to military applications. In addition to their extraordinary spectral versatility, dye lasers have been shown to oscillate from the femtosecond pulse domain to the continuous wave (cw) regime. For microsecond pulse emission, energies of up to hundreds of joules per pulse have been demonstrated. Further, operation at high pulsed repetition frequencies (prfs), in the multi-kHz regime, has provided average powers at kW levels. This unrivaled operational versatility is summarized in Table 1.

Dye lasers are excited by coherent optical energy from an excitation, or pump, laser or by optical energy from specially designed lamps called flashlamps. Recent advances in semiconductor laser technology have made it possible to construct very compact all-solid-state excitation sources that, coupled with new solid-state dye laser materials, should bring the opportunity to build compact tunable laser systems for the visible spectrum. Further, direct diode-laser pumping of solid-state dye lasers should prove even more advantageous to enable the development of fairly inexpensive tunable narrow-linewidth solid-state dye laser systems for spectroscopy and other applications requiring low powers. Work on electrically excited organic gain media might also provide new avenues for further progress.

The literature of dye lasers is very rich and many review articles have been written describing and discussing traditional dye lasers utilizing liquid gain media. In particular, the books *Dye Lasers*, *Dye Laser Principles*, *High Power Dye Lasers*, and *Tunable Lasers Handbook* provide excellent sources of authoritative and detailed description of the physics and technology involved. In this article we offer only a survey of the operational capabilities of the dye lasers using liquid gain media in order to examine with more attention the field of solid-state dye lasers.

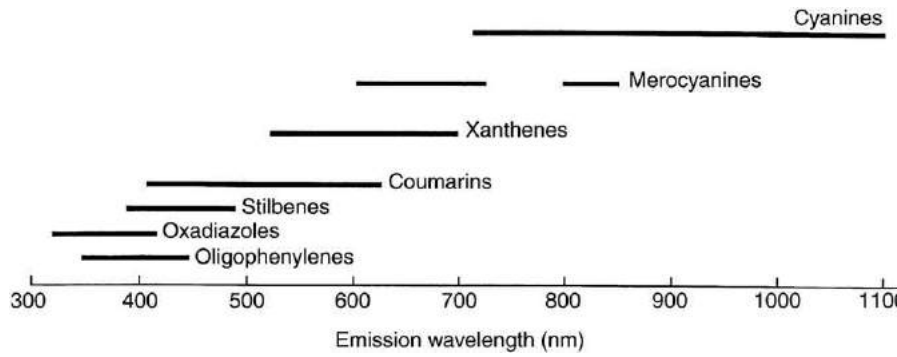

Fig. 1 Approximate wavelength span from the various classes of laser dye molecules. Reproduced with permission from Duarte FJ (1995)

Table 1 Emission characteristics of liquid dye lasers

Dye laser class	*Spectral coverage*[a]	*Energy per pulse*[b]	*Prf*[b]	*Power*[b]
Laser-pumped pulsed dye lasers	350–1100 nm	800 J[c]	13 kHz[d]	2.5 kW[d]
Flashlamp-pumped dye lasers	320–900 nm	400 J[e]	850 Hz[f]	1.2 kW[f]
CW dye lasers	370–1000 nm			43 W[g]

[a]Approximate range.
[b]Refers to maximum values for that particular emission parameter.
[c]Achieved with an excimer-laser pumped coumarin dye laser by Tang and colleagues, in 1987.
[d]Achieved with a multistage copper-vapor-laser pumped dye laser using rhodamine dye by Bass and colleagues, in 1992.
[e]Reported by Baltakov and colleagues, in 1974. Uses rhodamine 6G dye.
[f]Reported by Morton and Dragoo, in 1981. Uses coumarin 504 dye.
[g]Achieved with an Ar^+ laser pumped folded cavity dye laser using rhodamine 6G dye by Baving and colleagues, in 1982.

Tunable Laser Handbook. New York: Academic Press.

 doi:10.1016/B978-0-12-809283-5.00840-5

Brief History of Dye Lasers

1965: Quantum theory of dyes is discussed in the context of the maser (R. P. Feynman).
1966: Dye lasers are discovered (P. P. Sorokin and J. R. Lankard; F. P. Schäfer and colleagues).
1967: The flashlamp-pumped dye laser is discovered (P. P. Sorokin and J. R. Lankard; W. Schmidt and F. P. Schäfer).
1967–1968: Solid-state dye lasers are discovered (B. H. Soffer and B. B. McFarland; O. G. Peterson and B. B. Snavely).
1968: Mode-locking, using saturable absorbers, is demonstrated in dye lasers (W. Schmidt and F. P. Schäfer).
1970: The continuous-wave (cw) dye laser is discovered (O. G. Peterson, S. A. Tuccio, and B. B. Snavely).
1971: The distributed feedback dye laser is discovered (H. Kogelnik and C. V. Shank).
1971–1975: Prismatic beam expansion in dye lasers is introduced (S. A. Myers; E. D. Stokes and colleagues; D. C. Hanna and colleagues).
1972: Passive mode-locking is demonstrated in cw dye lasers (E. P. Ippen, C. V. Shank, and A. Dienes).
1972: The first pulsed narrow-linewidth tunable dye laser is introduced (T. W. Hänsch).
1973: Frequency stabilization of cw dye lasers is demonstrated (R. L. Barger, M. S. Sorem, and J. L. Hall).
1976: Colliding-pulse-mode locking is introduced (I. S. Ruddock and D. J. Bradley).
1977–1978: Grazing-incidence grating cavities are introduced (I. Shoshan and colleagues; M. G. Littman and H. J. Metcalf; S. Saikan).
1978–1980: Multiple-prism grating cavities are introduced (T. Kasuya and colleagues; G. Klauminzer; F. J. Duarte and J. A. Piper).
1981: Prism pre-expanded grazing-incidence grating oscillators are introduced (F. J. Duarte and J. A. Piper).
1982: Generalized multiple-prism dispersion theory is introduced (F. J. Duarte and J. A. Piper).
1983: Prismatic negative dispersion for pulse compression is introduced (W. Dietel, J. J. Fontaine, and J-C. Diels).
1987: Laser pulses as short as six femtoseconds are demonstrated (R. L. Fork, C. H. Brito Cruz, P. C. Becker, and C. V. Shank).
1994: First narrow-linewidth solid-state dye laser oscillator (F. J. Duarte).
1999–2000: Distributed feedback solid-state dye lasers are introduced (Wadsworth and colleagues; Zhu and colleagues).

Molecular Energy Levels

Dye molecules have large molecular weights and contain extended systems of conjugated double bonds. These molecules can be dissolved in an adequate organic solvent (such as ethanol, methanol, ethanol/water, and methanol/water) or incorporated into a solid matrix (organic, inorganic, or hybrid). These molecular gain media have a strong absorption generally in the visible and ultraviolet regions, and exhibit large fluorescence bandwidths covering the entire visible spectrum. The general energy level diagram of an organic dye is shown in Fig. 2. It consists of electronic singlet and triplet states with each electronic state containing a multitude of overlapping vibrational–rotational levels giving rise to broad continuous energy bands. Absorption of visible or ultraviolet pump light excites the molecules from the ground state S_0 into some rotational–vibrational level belonging to an upper excited singlet state, from where the molecules decay nonradiatively to the lowest vibrational level of the first excited singlet state S_1 on a picosecond time-scale. From S_1 the molecules can decay radiatively, with a radiative lifetime on the nanosecond time-scale, to a higher-lying vibrational–rotational level of S_0. From this level they rapidly thermalize into the lowest vibrational–rotational levels of S_0. Alternatively, from S_1, the molecules can experience nonradiative relaxation either to the triplet state T_1 by an intersystem crossing process or to the ground state by an internal conversion process. If the intensity of the pumping radiation is high enough a population inversion between S_1 and S_0 may be attained and stimulated emission occurs. Internal conversion and intersystem crossing compete with the fluorescence decay mode of the molecule and therefore reduce the efficiency of the laser emission. The rate for internal conversion to the electronic ground state is usually negligibly small so that the most important loss process is intersystem crossing into T_1 that populates the lower metastable triplet state. Thus, absorption on the triplet–triplet allowed transitions could cause considerable losses if these absorption bands overlap the lasing band, inhibiting or even halting the lasing process. This triplet loss can be reduced by adding small quantities of appropriate chemicals that favor nonradiative transitions that shorten the effective lifetime of the T_1 level. For pulsed excitation with nanosecond pulses, the triplet–triplet absorption can be neglected because for a typical dye the intersystem crossing rate is not fast enough to build up an appreciable triplet population in the nanosecond time domain.

Dye molecules are large (a typical molecule incorporates 50 or more atoms) and are grouped into families with similar chemical structures. A survey of the major classes of laser dyes is given later. Solid-state laser dye gain media are also considered later.

Liquid Dye Lasers

Laser-Pumped Pulsed Dye Lasers

Laser-pumped dye lasers use a shorter wavelength, or higher frequency, pulsed laser as the excitation or pump source. Typical pump lasers for dye lasers are gas lasers such as the excimer, nitrogen, or copper lasers. One of the most widely used solid-state laser pumps is the frequency doubled Nd:YAG laser which emits at 532 nm.

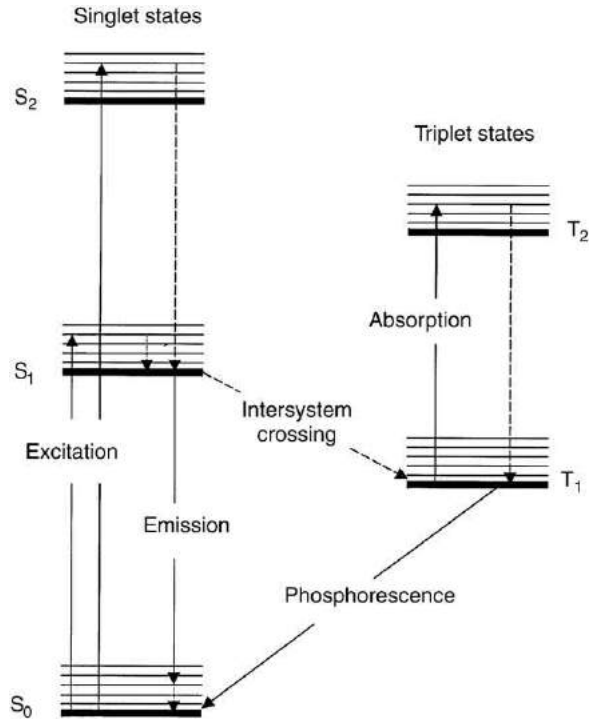

Fig. 2 Schematic energy level diagram for a dye molecule. Full lines: radiative transitions; dashed lines: nonradiative transitions; dotted lines: vibrational relaxation.

In a laser-pumped pulsed dye laser the active medium, or dye solution, is contained in an optical cell often made of quartz or fused silica, which provides an active region typically some 10 mm in length and a few mm in width. The active medium is then excited either longitudinally, or transversely, via a focusing lens using the pump laser. In the case of transverse excitation the pump laser is focused to a beam $\sim$10 mm in width and $\sim$0.1 mm in height. Longitudinal pumping requires focusing of the excitation beam to a diameter in the 0.1–0.15 mm range. For lasers operated at low prfs (a few pulses per second), the dye solution might be static. However, for high-prf operation (a few thousand pulses per second) the dye solution must be flowed at speeds of up to a few meters per second in order to dissipate the heat. A simple broadband optically pumped dye laser can be constructed using just the pump laser, the active medium, and two mirrors to form a resonator. In order to achieve tunable, narrow-linewidth, emission, a more sophisticated resonator must be employed. This is called a dispersive tunable oscillator and is depicted in **Fig. 3**. In a dispersive tunable oscillator the exit side of the cavity is comprised of a partial reflector, or an output coupler, and the other end of the resonator is composed of a multiple-prism grating assembly. It is the dispersive characteristics of this multiple-prism grating assembly and the dimensions of the emission beam produced at the gain medium that determine the tunability and the narrowness, or spectral purity, of the laser emission.

In order to selectively excite a single vibrational–rotational level of a molecule such as iodine (I_2), at room temperature, one needs a laser linewidth of $\Delta\nu \approx 1.5$ GHz (or $\Delta\lambda \approx 0.0017$ nm at $\lambda = 590$ nm). The hybrid multiple-prism near-grazing-incidence (HMPGI) grating dye laser oscillator illustrated in **Fig. 4** yields laser linewidths in the 400 MHz$\leq\Delta\nu\leq$650 MHz range at 4–5% conversion efficiencies whilst excited by a copper-vapor laser operating at a prf of 10 kHz. Pulse lengths are $\sim$10 ns at full-width half-maximum (FWHM). The narrow-linewidth emission from these oscillators is said to be single-longitudinal-mode lasing because only one electromagnetic mode is allowed to oscillate.

The emission from oscillators of this class can be amplified many times by propagating the tunable narrow-linewidth laser beam through single-pass amplifier dye cells under the excitation of pump lasers. Such amplified laser emission can reach enormous average powers. Indeed, a copper-vapor-laser pumped dye laser system at the Lawrence Livermore National Laboratory (USA), designed for the laser isotope separation program, was reported to yield average powers in excess of 2.5 kW, at a prf of 13.2 kHz, at a better than 50% conversion efficiency as reported by Bass and colleagues in 1992. Besides high conversion efficiencies, excitation with copper-vapor lasers, at $\lambda = 510.554$ nm, has the advantage of inducing little photodegradation in the active medium thus allowing very long dye lifetimes.

Flashlamp-Pumped Dye Lasers

Flashlamps utilized in dye laser excitation emit at black-body temperatures in the 20 000 K range, thus yielding intense ultraviolet radiation centered around 200 nm. One further requirement for flashlamps, and their excitation circuits, is to deliver light pulses with a fast rise time. For some flashlamps this rise time can be less than a few nanoseconds.

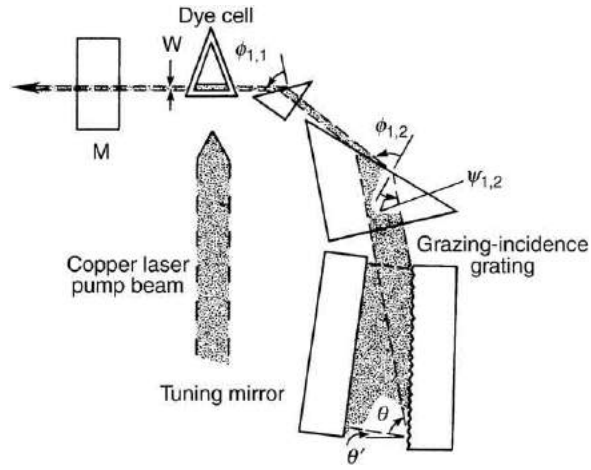

Fig. 3 Copper-vapor-laser pumped hybrid multiple-prism near grazing incidence (HMPGI) grating dye laser oscillator. Adapted with permission from Duarte FJ and Piper JA (1984) Narrow linewidth high prf copper laser-pumped dye-laser oscillators. *Applied Optics* 23: 1391–1394.

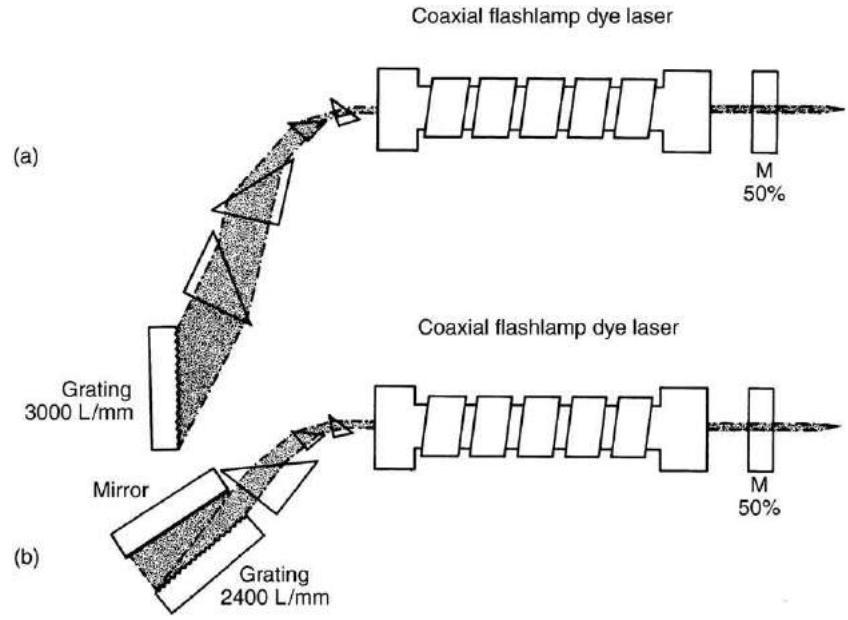

Fig. 4 Flashlamp-pumped multiple-prism grating oscillators. From Duarte FJ, Davenport WE, Ehrlich JJ and Taylor TS (1991) Ruggedized narrow-linewidth dispersive dye laser oscillator. *Optics Communications* 84: 310–316. Reproduced with permission from Elsevier.

Flashlamp-pumped dye lasers differ from laser-pumped pulsed dye lasers mainly in the pulse energies and pulse lengths attainable. This means that flashlamp-pumped dye lasers, using relatively large volumes of dye, can yield very large energy pulses. Excitation geometries use either coaxial lamps, with the dye flowing in a quartz cylinder at the center of the lamp, or two or more linear lamps arranged symmetrically around the quartz tube containing the dye solution. Using a relatively weak dye solution of rhodamine-6G (2.2×10^{-5} M), a coaxial lamp, and an active region defined by a quartz tube 6 cm in diameter and a length of 60 cm, Baltakov and colleagues, in 1974, reported energies of 400 J in pulses 25 μs long at FWHM.

Flashlamp-pumped tunable narrow-linewidth dye laser oscillators described by Duarte and colleagues, in1991, employ a cylindrical active region 6 mm in diameter and 17 cm in length. The dye solution is made of rhodamine 590 at a concentration of 1×10^{-5} M. This active region is excited by a coaxial flashlamp. Using multiple-prism grating architectures (see **Fig. 4**) these authors achieve a diffraction limited TEM_{00} laser beam and laser linewidths of $\Delta\nu \approx 300$ MHz at pulsed energies in the 2–3 mJ range. The laser pulse duration is reported to be $\Delta t \approx 100$ ns. The laser emission from this class of multiple-prism grating oscillator is reported to be extremely stable. The tunable narrow-linewidth emission from these dispersive oscillators is either used directly in spectroscopic, or other scientific applications, or is utilized to inject large flashlamp-pumped dye laser amplifiers to obtain multi-joule pulse energies with the laser linewidth characteristics of the oscillator.

Continuous Wave Dye Lasers

CW dye lasers use dye flowing at linear speeds of up to 10 meters per second which are necessary to remove the excess heat and to quench the triplet states. In the original cavity reported by Peterson and colleagues, in 1970, a beam from an Ar^+ laser was focused on to an active region which is contained within the resonator. The resonator comprised dichroic mirrors that transmit the blue-green radiation of the pump laser and reflect the red emission from the dye molecules. Using a pump power of about 1 W, in a TEM_{00} laser beam, these authors reported a dye laser output of 30 mW. Subsequent designs replaced the dye cell with a dye jet, an introduced external mirror, and integrated dispersive elements in the cavity. Dispersive elements such as prisms and gratings are used to tune the wavelength output of the laser. Frequency-selective elements, such as etalons and other types of interferometers, are used to induce frequency narrowing of the tunable emission. Two typical cw dye laser cavity designs are described by Hollberg, in 1990, and are reproduced here in **Fig. 5**. The first design is a linear three-mirror folded cavity. The second one is an eight-shaped ring dye laser cavity comprised of mirrors M_1, M_2, M_3, and M_4. Linear cavities exhibit the effect of *spatial hole burning* which allows the cavity to lase in more than one longitudinal mode. This problem can be overcome in ring cavities (**Fig. 5(b)**) where the laser emission is in the form of a traveling wave.

Two aspects of cw dye lasers are worth emphasizing. One is the availability of relatively high powers in single longitudinal mode emission and the other is the demonstration of very stable laser oscillation. First, Johnston and colleagues, in 1982, reported 5.6 W of stabilized laser output in a single longitudinal mode at 593 nm at a conversion efficiency of 23%. In this work eleven dyes were used to span the spectrum continuously from ~400 nm to ~900 nm. In the area of laser stabilization and ultra narrow-linewidth oscillation it is worth mentioning the work of Hough and colleagues, in 1984, that achieved laser linewidths of less than 750 Hz employing an external reference cavity.

Ultrashort-Pulse Dye Lasers

Ultrashort-pulse, or femtosecond, dye lasers use the same type of technology as cw dye lasers configured to incorporate a saturable absorber region. One such configuration is the ring cavity depicted in **Fig. 6**. In this cavity the gain region is established between mirrors M_1 and M_2 whilst the saturable absorber is deployed in a counter-propagating arrangement. This arrangement is necessary to establish a collision between two counter-propagating pulses at the saturable absorber thus yielding what is known as colliding-pulse mode locking (CPM) as reported by Ruddock and Bradley, in 1976. This has the effect of creating a transient grating, due to interference, at the absorber thus shortening the pulse. Intracavity prisms were incorporated in order to introduce negative dispersion, by Dietel and colleagues in 1983, and thus subtract dispersion from the cavity and ultimately provide the compensation needed to produce femtosecond pulses.

The shortest pulse obtained from a dye laser, 6 fs, has been reported by Fork and colleagues, in 1987, using extra-cavity compression. In that experiment, a dye laser incorporating CPM and prismatic compensation was used to generate pulses that were amplified by a copper-vapor laser at a prf of 8 kHz. The amplified pulses, of a duration of 50 fs, were then propagated through two grating pairs and a four-prism sequence for further compression.

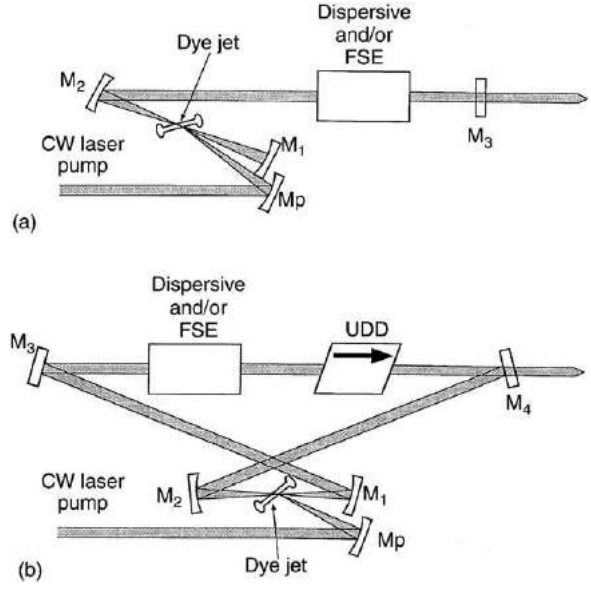

Fig. 5 CW laser cavities: (a) linear cavity and (b) ring cavity. Adapted from Hollberg LW (1990) CW dye lasers. In: Duarte FJ and Hillman LW (eds) *Dye Laser Principles*, pp. 185–238. New York: Academic Press. Reproduced with permission from Elsevier.

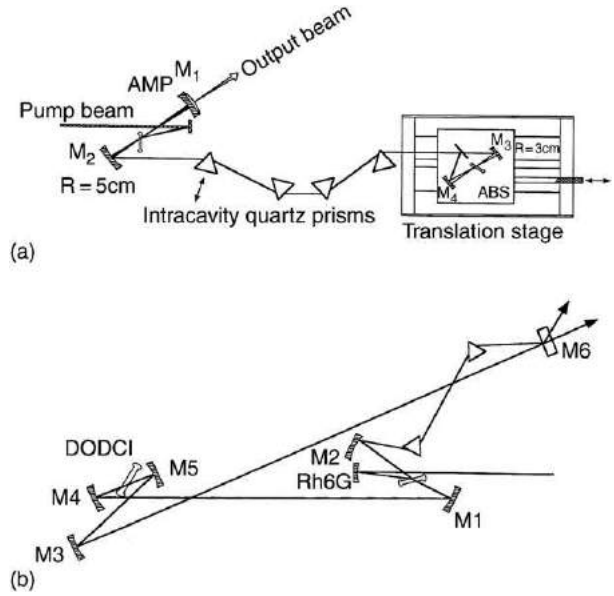

Fig. 6 Femtosecond dye laser cavities: (a) linear femtosecond cavity and (b) ring femtosecond cavity. Adapted from Diels J-C (1990) Femtosecond dye lasers. In: Duarte FJ and Hillman LW (eds) *Dye Laser Principles*, pp. 41–132. New York: Academic Press. Reproduced with permission from Elsevier.

Solid-State Dye Laser Oscillators

In this section the principles of linewidth narrowing in dispersive resonators are outlined. Albeit the discussion focuses on multiple-prism grating solid-state dye laser oscillators, in particular, the physics is applicable to pulsed high-power dispersive tunable lasers in general.

Multiple-Prism Dispersion Grating Theory

The spectral linewidth in a dispersive optical system is given by

$$\Delta\lambda \approx \Delta\theta(\nabla_\lambda\theta)^{-1} \tag{1}$$

where $\Delta\theta$ is the light beam divergence, $\nabla_\lambda = \partial/\partial\lambda$, and $\nabla_\lambda\theta$ is the overall dispersion of the optics. This identity can de derived either from the principles of geometrical optics or from the principles of generalized interferometry as described by Duarte, in 1992.

The cumulative single-pass generalized multiple-prism dispersion at the mth prism, of a multiple-prism array as illustrated in Fig. 7, as given by Duarte and Piper, in 1982,

$$\nabla_\lambda\phi_{2,m} = H_{2,m}\nabla_\lambda n_m + (k_{1,m}k_{2,m})^{-1} \times \left(H_{1,m}\nabla_\lambda n_m \pm \nabla_\lambda\phi_{2,(m-1)}\right) \tag{2}$$

In this equation

$$k_{1,m} = \cos\psi_{1,m}/\cos - \phi_{1,m} \tag{3a}$$

$$k_{2,m} = \cos\phi_{2,m}/\cos - \psi_{2,m} \tag{3b}$$

$$H_{1,m} = \tan\phi_{1,m}/n_m \tag{4a}$$

$$H_{2,m} = \tan\phi_{2,m}/n_m \tag{4b}$$

Here, $k_{1,m}$ and $k_{2,m}$ represent the physical beam expansion experienced by the incident and the exit beams, respectively.

Eq. (2) indicates that $\nabla_\lambda\phi_{2,m}$, the cumulative dispersion at the mth prism, is a function of the geometry of the mth prism, the position of the light beam relative to this prism, the refractive index of the prism, and the cumulative dispersion up to the previous prism $\nabla_\lambda\phi_{2,(m-1)}$.

For an array of r identical isosceles, or equilateral, prisms arranged symmetrically, in an additive configuration, so that the angles of incidence and emergence are the same, the cumulative dispersion reduces to

$$\nabla_\lambda\phi_{2,r} = r\nabla_\lambda\phi_{2,1} \tag{5}$$

Under these circumstances the dispersions add in a simple and straightforward manner. For configurations incorporating right-angle prisms, the dispersions need to be handled mathematically in a more subtle form.

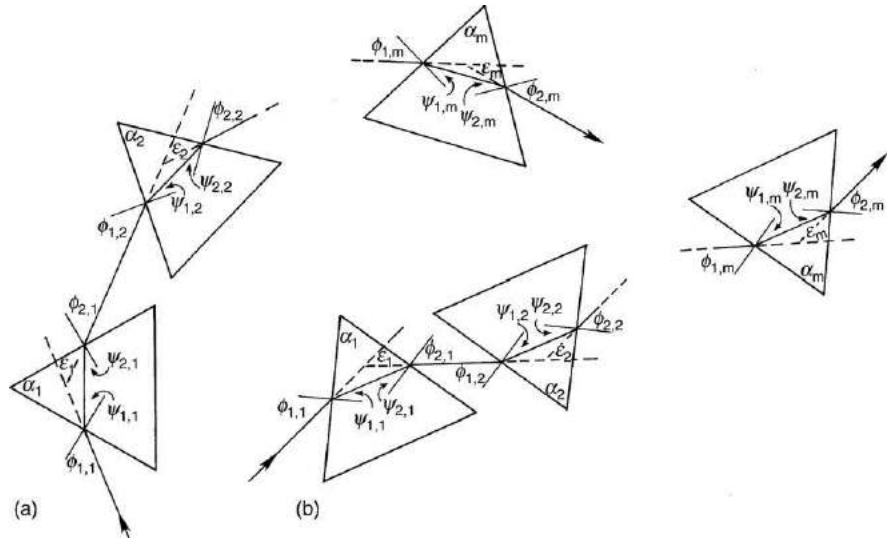

Fig. 7 Generalized multiple-prism arrays in (a) additive and (b) compensating configurations. From Duarte FJ (1990) Narrow-linewidth pulsed dye laser oscillators. In: Duarte FJ and Hillman LW (eds) *Dye Laser Principles*, pp. 133–183. New York: Academic Press. Reproduced with permission from Elsevier.

The generalized double-pass, or return-pass, dispersion for multiple-prism beam expanders was introduced by Duarte, in 1985:

$$\nabla_\lambda \Phi_P = 2M_1M_2 \sum_{m=1}^{r} (\pm 1) H_{1,m} \left(\prod_{j=m}^{r} k_{1,j} \prod_{j=m}^{r} k_{2,j} \right)^{-1} \nabla_\lambda n_m + 2 \sum_{m=1}^{r} (\pm 1) H_{2,m} \left(\prod_{j=1}^{m} k_{1,j} \prod_{j=1}^{m} k_{2,j} \right) \nabla_\lambda n_m \tag{6}$$

Here, M_1 and M_2 are the beam magnification factors given by

$$M_1 = \prod_{m=1}^{r} k_{1,m} \tag{7a}$$

$$M_2 = \prod_{m=1}^{r} k_{2,m} \tag{7b}$$

For a multiple prism expander designed for an orthogonal beam exit, and Brewster's angle of incidence, **Eq. (6)** reduces to the succinct expression given by Duarte, in 1990:

$$\nabla_\lambda \Phi_P = 2 \sum_{m=1}^{r} (\pm 1) (n_m)^{m-1} \nabla_\lambda n_m \tag{8}$$

Eq. (6) can be used to either quantify the overall dispersion of a given multiple-prism beam expander or to design a prismatic expander yielding zero dispersion, that is, $\nabla_\lambda \Phi_P = 0$, at a given wavelength.

Physics and Architecture of Solid-State Dye-Laser Oscillators

The first high-performance narrow-linewidth tunable laser was introduced by Hänsch in 1972. This laser yielded a linewidth of $\Delta\nu \approx 2.5$ GHz (or $\Delta\lambda \approx 0.003$ nm at $\lambda \approx 600$ nm) in the absence of an intracavity etalon. Hänsch demonstrated that the laser linewidth from a tunable laser was narrowed significantly when the beam incident on the tuning grating was expanded using an astronomical telescope. The linewidth equation including the intracavity beam magnification factor can be written as

$$\Delta\lambda \approx \Delta\theta (M \nabla_\lambda \Theta_G)^{-1} \tag{9}$$

From this equation, it can be deduced that a narrow $\Delta\lambda$ is achieved by reducing $\Delta\theta$ and increasing the overall intracavity dispersion ($M\nabla_\lambda \Theta_G$). The intracavity dispersion is optimized by expanding the size of the intracavity beam incident on the diffractive surface of the tuning grating until it is totally illuminated.

Hänsch used a two-dimensional astronomical telescope to expand the intracavity beam incident on the diffraction grating. A simpler beam expansion method consists in the use of a single-prism beam expander as disclosed by several authors (Myers, 1971; Stokes and colleagues, 1972; Hanna and colleagues, 1975). An extension and improvement of this approach was the introduction of multiple-prism beam expanders as reported by Kasuya and colleagues, in 1978, Klauminzer, in 1978, and Duarte and Piper, in 1980. The main advantages of multiple-prism beam expanders, over two-dimensional telescopes, are simplicity, compactness, and the fact that the beam expansion is reduced from two dimensions to one dimension. Physically, as explained previously, prismatic beam expanders also introduce a dispersion component that is absent in the case of the astronomical telescope. Advantages of

multiple-prism beam expanders over single-prism beam expansion are higher transmission efficiency, lower amplified spontaneous emission levels, and the flexibility to either augment or reduce the prismatic dispersion.

In general, for a pulsed multiple-prism grating oscillator, Duarte and Piper, in 1984, showed that the return-pass dispersive linewidth is given by

$$\Delta\lambda = \Delta\theta_R (MR\nabla_\lambda \Theta_G + R\nabla_\lambda \Phi_P)^{-1} \tag{10}$$

where R is the number of return-cavity passes. The grating dispersion in this equation, $\nabla_\lambda \Theta_G$, can be either from a grating in Littrow or near grazing-incidence configuration. The multiple-return-pass equation for the beam divergence was given by Duarte, in 2001:

$$\Delta\theta_R = (\lambda/\pi w)\left(1 + (L_R/B_R)^2 + (A_R L_R/B_R)^2\right)^{1/2} \tag{11}$$

Here, $L_R = (\pi w^2/\lambda)$ is the Rayleigh length of the cavity, w is the beam waist at the gain region, while A_R and B_R are the corresponding multiple-return-pass elements derived from propagation matrices.

At present, very compact and optimized multiple-prism grating tunable laser oscillators are found in two basic cavity architectures. These are the multiple-prism Littrow (MPL) grating laser oscillator, reported by Duarte in 1999 and illustrated in **Fig. 8**, and the hybrid multiple-prism near grazing-incidence (HMPGI) grating laser oscillator, introduced by Duarte in 1997 and depicted in **Fig. 9**. In early MPL grating oscillators the individual prisms integrating the multiple-prism expander were deployed in an additive configuration thus adding the cumulative dispersion to that of the grating and thus contributing to the overall dispersion of the cavity. In subsequent architectures the prisms were deployed in compensating configurations so as to yield zero dispersion and thus allow the tuning characteristics of the cavity to be determined by the grating exclusively. In this approach the principal role of the multiple-prism array is to expand the beam incident on the grating thus augmenting significantly the overall dispersion of the cavity as described in Eqs. [9] or [10]. In this regard, it should be mentioned that beam magnification factors of up to 100, and beyond, have been reported in the literature. Using solid-state laser dye gain media, these MPL and HMPGI grating laser oscillators deliver tunable single-longitudinal-mode emission at laser linewidths 350 MHz $\leq \Delta\nu \leq$ 375 MHz and pulse lengths in the 3–7 ns (FWHM) range. Long-pulse operation of this class of multiple-prism grating oscillators has been reported by Duarte and colleagues, in 1998. In these experiments laser linewidths of $\Delta\nu \approx 650$ MHz were achieved at pulse lengths in excess of 100 ns (FWHM) under flashlamp-pumped dye laser excitation.

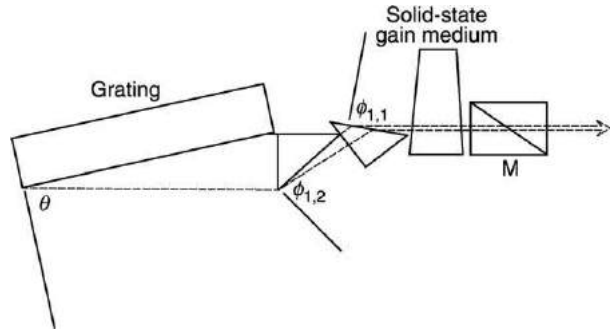

Fig. 8 MPL grating solid-state dye laser oscillator: optimized architecture. The physical dimensions of the optical components of this cavity are shown to scale. The length, of the solid-state dye gain medium, along the optical axis, is 10 mm. Reproduced with permission from Duarte FJ (1999) Multiple-prism grating solid-state dye laser oscillator: optimized architecture. *Applied Optics* 38: 6347–6349.

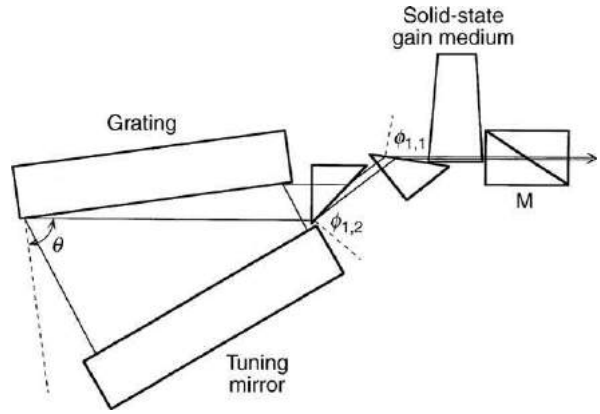

Fig. 9 HMPGI grating solid-state dye laser. Schematics to scale. The length of the solid state dye gain medium, along the optical axis, is 10 mm. Reproduced with permission from Duarte FJ (1997) Multiple-prism near-grazing-incidence grating solid-state dye laser oscillator. *Optics and Laser Technology* 29: 513–516.

The dispersive cavity architectures described here have been used with a variety of laser gain media in the gas, the liquid, and the solid state. Applications to tunable semiconductor lasers have also been reported by Zorabedian, in 1992, Duarte, in 1993, and Fox and colleagues, in 1997.

Concepts important to MPL and HMPGI grating tunable laser oscillators include the emission of a single-transverse-mode (TEM_{00}) laser beam in a compact cavity, the use of multiple-prism arrays, the expansion of the intracavity beam incident on the grating, the control of the intracavity dispersion, and the quantification of the overall dispersion of the multiple-prism grating assembly via generalized dispersion equations. Sufficiently high intracavity dispersion leads to the achievement of return-pass dispersive linewidths close to the free-spectral range of the cavity. Under these circumstances single-longitudinal-mode lasing is readily achieved as a result of multipass effects.

A Note on the Cavity Linewidth Equation

So far we have described how the cavity linewidth equation

$$\Delta\lambda \approx \Delta\theta(\nabla_{\lambda}\theta)^{-1}$$

and the dispersion equations can be applied to achieve highly coherent, or very narrow-linewidth, emission. It should be noted that the same physics can be applied to achieve ultrashort, or femtosecond, laser pulses. It turns out that intracavity prisms can be configured to yield negative dispersion as described by Duarte and Piper, in 1982, Dietel and colleagues, in 1983, and Fork and colleagues in 1984. This negative dispersion can reduce significantly the overall dispersion of the cavity. Under those circumstances, Eq. (1) predicts broadband emission which according to the uncertainty principle, in the form of,

$$\Delta\nu\Delta t \approx 1 \tag{12}$$

can lead to very short pulse emission since $\Delta\nu$ and $\Delta\lambda$ are related by the identities

$$\Delta\lambda \approx \lambda^2/\Delta x \tag{13}$$

and

$$\Delta\nu \approx c/\Delta x \tag{14}$$

Distributed-Feedback Solid-State Dye Lasers

Recently, narrow-linewidth laser emission from solid-state dye lasers has also been obtained using a distributed-feedback (DFB) configuration by Wadsworth and colleagues, in 1999, and Zhu and colleagues, in 2000. As is well known, in a DFB laser no external cavity is required, the feedback being provided by Bragg reflection from a permanently or dynamically written grating structure within the gain medium as reported by Kogelnik and Shank, in 1971. By using a DFB laser configuration where the interference of two pump beams induces the required periodic modulation in the gain medium, laser emission with linewidth in the 0.01–0.06 nm range have been reported. Specifically, Wadsworth and colleagues reported a laser linewidth of 12 GHz (0.016 nm at 616 nm) using Perylene Red doped PMMA at a conversion efficiency of 20%.

It should also be mentioned that the DFB laser is also a dye laser development that has found wide and extensive applicability in semiconductor lasers employed in the telecommunications industry.

Laser Dye Survey

Cyanines

These are red and near-infrared dyes which present long conjugated methine chains (!CH=CH!) and are useful in the spectral range longer than 800 nm, where no other dyes compete with them. An important laser dye belonging to this class is 4-dicyanomethylene-2-methyl-6-(*p*-dimethylaminostyryl)-4H-pyran (DCM, Fig. 10(a)), with laser emission in the 600–700 nm range depending on the pump and solvent, liquid or solid, used.

Xanthenes

These dyes have a xanthene ring as the chromophore and are classified into rhodamines, incorporating amino radicals substituents, and fluoresceins, with hydroxyl (OH) radical substituents. They are generally very efficient and chemically stable and their emission covers the wavelength region from 500 to 700 nm.

Rhodamine dyes are the most important group of all laser materials, with rhodamine 6G (Rh6G, Fig. 10(b)), also called rhodamine 590 chloride, being probably the best known of all laser dyes. Rh6G exhibits an absorption peak, in ethanol, at 530 nm and a fluorescence peak at 556 nm, with laser emission typically in the 550–620 nm region. It has been demonstrated to lase efficiently both in liquid and solid solutions.

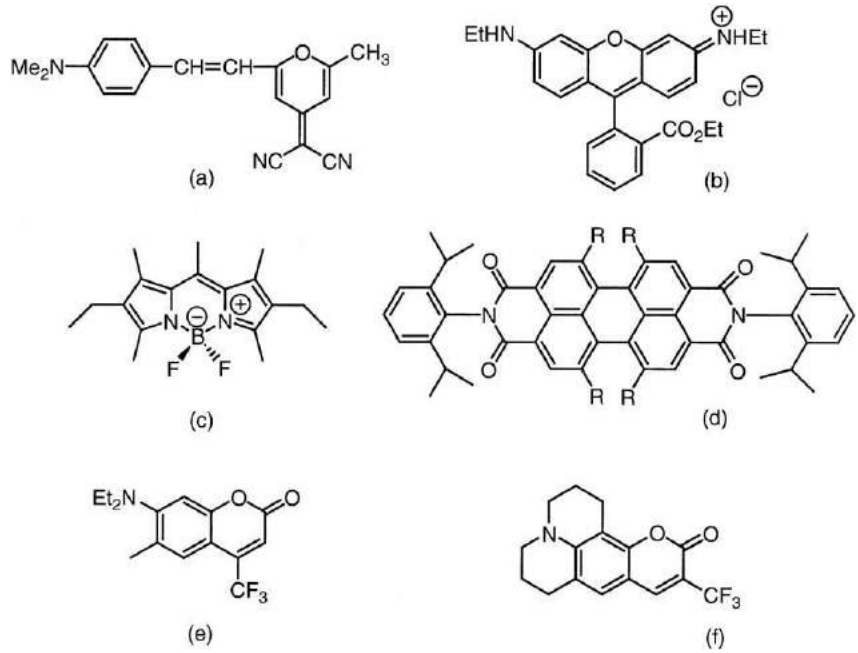

Fig. 10 Molecular structures of some common laser dyes: (a) DCM; (b) Rh6G; (c) PM567; (d) Perylene Orange (R = H) and Perylene Red (R = C_4H_6O); (e) Coumarin 307 (also known as Coumarin 503); (f) Coumarin 153 (also known as Coumarin 540A).

Pyrromethenes

Pyrromethene.BF_2 complexes are a new class of laser dyes synthesized and characterized more recently as reported by Shah and colleagues, in 1990, Pavlopoulos and colleagues, in 1990, and Boyer and colleagues, in 1993. These laser dyes exhibit reduced triplet–triplet absorption over their fluorescence and lasing spectral region while retaining a high quantum fluorescence yield. Depending on the substituents on the chromophore, these dyes present laser emission over the spectral region from the green/yellow to the red, competing with the rhodamine dyes and have been demonstrated to lase with good performance when incorporated into solid hosts. Unfortunately, they are relatively unstable because of the aromatic amine groups in their structure, which render them vulnerable to photochemical reactions with oxygen as indicated by Rahn and colleagues, in 1997. The molecular structure of a representative and well-known dye of this class, 1,3,5,7,8-pentamethyl-2,6-diethylpyrromethene-difluoroborate complex (pyrromethene 567, PM567) is shown in Fig. 10(c). Recently, analogs of dye PM567 substituted at position 8 with an acetoxypolymethylene linear chain or a polymerizable methacryloyloxy polymethylene chain have been developed. These new dipyrromethene.BF_2 complexes have demonstrated improved efficiency and better photostability than the parent compound both in liquid and solid state.

Conjugated Hydrocarbons

This class of organic compounds includes perylenes, stilbenes, and *p*-terphenyl. Perylene and perylimide dyes are large, nonionic, nonpolar molecules (Fig. 10(d)) characterized by their extreme photostability and negligible singlet–triplet transfer, as discussed by Seybold and Wagenblast and colleagues, in 1989, which have high quantum efficiency because of the absence of nonradiative relaxation. The perylene dyes exhibit limited solubility in conventional solvents but dissolve well in acetone, ethyl acetate, and methyl methacrylate, with emission wavelengths in the orange and red spectral regions.

Stilbene dyes are derivatives of uncyclic unsaturated hydrocarbons such as ethylene and butadiene, with emission wavelengths in the 400–500 nm range. Most of them end in phenyl radicals. Although they are chemically stable, their laser performance is inferior to that of coumarins. *p*-Terphenyl is a valuable and efficient *p*-oligophenylene dye with emission in the ultraviolet.

Coumarins

Coumarin derivatives are a popular family of laser dyes with emission in the blue-green region of the spectrum. Their structure is based on the coumarin ring (Fig. 10(e) and (f)) with different substituents that strongly affect its chemical characteristics and allow covering an emission spectral range between 420 and 580 nm. Some members of this class rank among the most efficient laser dyes known, but their chemical photobleaching is rapid compared with xanthene and pyrromethene dyes. An additional class of blue-green dyes is the tetramethyl derivatives of coumarin dyes introduced by Chen and colleagues, in 1988. The emission from

these dyes spans the spectrum in the 453–588 nm region. These coumarin analogs exhibit improved efficiency, higher solubility, and better lifetime characteristics than the parent compounds.

Azaquinolone Derivatives

The quinolone and azaquinolone derivatives have a structure similar to that of coumarins but with their laser emission range extended toward the blue by 20–30 nm. The quinolone dyes are used when laser emission in the 400–430 nm region is required. Azaquinolone derivatives, such as 7-dimethylamino-1-methyl-4-metoxy-8-azaquinolone-2 (LD 390), exhibit laser action below 400 nm.

Oxadiazole Derivatives

These are compounds which incorporate an axodiazole ring with aryl radical substituents and which lase in the 330–450 nm spectral region. Dye 2-(4-biphenyl)-5-(4-*t*-butylphenyl)-1,3,4-oxadiazole (PBD), with laser emission in the 355–390 nm region, belongs to this family.

Solid-State Laser Dye Matrices

Although some first attempts to incorporate dye molecules into solid matrices were already made in the early days of development of dye lasers, it was not until the late 1980s that host materials with the required properties of optical quality and high damage threshold to laser radiation began to be developed. In subsequent years, the synthesis of new high-performance dyes and the implementation of new ways of incorporating the organic molecules into the solid matrix resulted in significant advances towards the development of practical tunable solid-state dye lasers. A recent detailed review of the work done in this field has been given by Costela *et al.*

Inorganic glasses, transparent polymers, and inorganic–organic hybrid matrices have been successfully used as host matrices for laser dyes. Inorganic glasses offer good thermal and optical properties but present the difficulty that the high melting temperature employed in the traditional process of glass making would destroy the organic dye molecules. It was not until the development of the low-temperature sol-gel technique for the synthesis of glasses that this limitation was overcome. The sol-gel process, based on inorganic polymerization reactions performed at about room temperature and starting with metallo-organic compounds such as alkoxides or salts, as reported by Brinker and Scherrer, in 1989, provides a safe route for the preparation of rigid, transparent, inorganic matrix materials incorporating laser dyes. Disadvantages of these materials are the possible presence of impurities embedded in the matrix or the occurrence of interactions between the dispersed dye molecules and the inorganic structure (hydrogen bonds, van der Waals forces) which, in certain cases, can have a deleterious effect on the lasing characteristics of the material.

The use of matrices based on polymeric materials to incorporate organic dyes offers some technical and economical advantages. Transparent polymers exhibit high optical homogeneity, which is extremely important for narrow-linewidth oscillators, as explained by Duarte, in 1994, good chemical compatibility with organic dyes, and allow control over structure and chemical composition making possible the modification, in a controlled way, of relevant properties of these materials such as polarity, free volume, molecular weight, or viscoelasticity. Furthermore, the polymeric materials are amenable to inexpensive fabrication techniques, which facilitate miniaturization and design of integrated optical systems. The dye molecules can be either dissolved in the polymer or linked covalently to the polymeric chains.

In the early investigations on the use of solid polymer matrices for laser dyes, the main problem to be solved was that of the low resistance to laser radiation exhibited by the then existing materials. In the late 1980s and early 1990s new modified polymeric organic materials began to be developed with a laser-radiation-damage threshold comparable to that of inorganic glasses and crystals as reported by Gromov and colleagues, in 1985, and Dyumaev and colleagues, in 1992. This, together with the above-mentioned advantages, made the use of polymers in solid-state dye lasers both attractive and competitive.

An approach that intends to bring about materials in which the advantages of both inorganic glasses and polymers are preserved but the difficulties are avoided is that of using inorganic–organic hybrid matrices. These are silicate-based materials, with an inorganic Si–O–Si backbone, prepared from organosilane precursors by sol-gel processing in combination with organic cross-linking of polymerizable monomers as reported by Novak, in 1993, Sanchez and Ribot, in 1994, and Schubert, in 1995. In one procedure, laser dyes are mixed with organic monomers, which are then incorporated into the porous structure of a sol-gel inorganic matrix by immersing the bulk in the solution containing monomer and catalyst or photoinitiator as indicated by Reisfeld and Jorgensen, in 1991, and Bosch and colleagues, in 1996. Alternatively, hybrids can be obtained from organically modified silicon alkoxides, producing the so-called ORMOCERS (organically modified ceramics) or ORMOSILS (organically modified silanes) as reported by Reisfeld and Jorgensen, in 1991. An inconvenient aspect of these materials is the appearance of optical inhomogeneities in the medium due to the difference in the refractive index between the organic and inorganic phases as well as to the difference in density between monomer and polymer which causes stresses and the formation of optical defects. As a result, spatial inhomogeneities appear in the laser beam, thereby decreasing its quality as discussed by Duarte, in 1994.

Organic Hosts

The first polymeric materials with improved resistance to damage by laser radiation were obtained by Russian workers, who incorporated rhodamine dyes into modified poly(methyl methacrylate) (MPMMA), obtained by doping PMMA with low-molecular weight additives. Some years later, in 1995, Maslyukov and colleagues demonstrated lasing efficiencies in the range 40–60% with matrices of MPMMA doped with rhodamine dyes pumped longitudinally at 532 nm, with a useful lifetime (measured as a 50% efficiency drop) of 15 000 pulses at a prf of 3.33 Hz.

In 1995, Costela and colleagues used an approach based on adjusting the viscoelastic properties of the material by modifying the internal plasticization of the polymeric medium by copolymerization with appropriate monomers. Using dye Rh6G dissolved in a copolymer of 2-hydroxyethyl methacrylate (HEMA) and methyl methacrylate (MMA) and transversal pumping at 337 nm, they demonstrated laser action with an efficiency of 21% and useful lifetime of 4500 pulses (20 GJ/mol in terms of total input energy per mole of dye molecule when the output energy is down to 50% of its initial value). The useful lifetime increased to 12 000 pulses when the Rh6G chromophore was linked covalently to the polymeric chains as reported by Costela and colleagues, in 1996.

Comparative studies on the laser performance of Rh6G incorporated either in copolymers of HEMA and MMA or in MPMMA were carried out by Giffin and colleagues, in 1999, demonstrating higher efficiency of the MPMMA materials but superior normalized photostability (up to 240 GJ/mol) of the copolymer formulation.

Laser conversion efficiencies in the range 60–70% for longitudinal pumping at 532 nm were reported by Ahmad and colleagues, in 1999, with PM567 dissolved in PMMA modified with 1,4-diazobicyclo[2,2,2] octane (DABCO) or perylene additives. When using DABCO as an additive, a useful lifetime of up to 550 000 pulses, corresponding to a normalized photostability of 270 GJ/mol, was demonstrated at a 2 Hz prf.

Much lower efficiencies and photostabilities have been obtained with dyes emitting in the blue-green spectral region. Costela and colleagues, in 1996, performed some detailed studies on dyes coumarin 540A and coumarin 503 incorporated into methacrylate homopolymers and copolymers, and demonstrated efficiencies of at most 19% with useful lifetimes of up to 1200 pulses, at a 2 Hz prf, for transversal pumping at 337 nm. In 2000, Somasundaram and Ramalingam obtained useful lifetimes of 5240 and 1120 pulses, respectively, for coumarin 1 and coumarin 490 dyes incorporated into PMMA rods modified with ethyl alcohol, under transversal pumping at 337 nm and a prf of 1 Hz.

A detailed study of photo-physical parameters of R6G-doped HEMA-MMA gain media was performed by Holzer and colleagues, in 2000. Among the parameters determined by these authors are quantum yields, lifetimes, absorption cross-sections, and emission cross-sections. A similar study on the photophysical parameters of PM567 and two PM567 analogs incorporated into solid matrices of different acrylic copolymers and in corresponding mimetic liquid solutions was performed by Bergmann and colleagues, in 2001.

Preliminary experiments with rhodamine 6G-doped MPMMA at high prfs were conducted using a frequency-doubled Nd:YAG laser at 5 kHz by Duarte and colleagues, in 1996. In those experiments, involving longitudinal excitation, the pump laser was allowed to fuse a region at the incidence window of the gain medium so that a lens was formed. This lens and the exit window of the gain medium comprised a short unstable resonator that yielded broadband emission. In 2001, Costela and colleagues demonstrated lasing in rhodamine 6G- and pyrromethene-doped polymer matrices under copper-vapor-laser excitation at prfs in the 1.0–6.2 kHz range. In these experiments, the dye-doped solid-state matrix was rotated at 1200 rpm. Average powers of 290 mW were obtained at a prf of 1kHz and a conversion efficiency of 37%. For a short period of time average powers of up to 1 W were recorded at a prf of 6.2 kHz. In subsequent experiments by Abedin and colleagues the rpf was increased to 10 kHz by using as pump laser a diode-pumped, *Q*-switched, frequency doubled, solid state Nd:YLF laser. Initial average powers of 560 mW and 430 mW were obtained for R6G-doped and PM567-doped polymer matrices, respectively. Lasing efficiency was 16% for R6G and the useful lifetime was 6.6 min (or about 4.0 million shots).

Inorganic and Hybrid Hosts

In 1990, Knobbe and colleagues and McKiernan and colleagues, from the University of California at Los Angeles, incorporated different organic dyes, via the sol-gel techniques, into silicate (SiO_2), aluminosilicate (Al_2O_3–SiO_2), and ORMOSIL host matrices, and investigated the lasing performance of the resulting materials under transversal pumping. Lasing efficiencies of 25% and useful lifetimes of 2700 pulses at 1 Hz repetition rate were obtained from Rh6G incorporated into aluminosilicate gel and ORMOSIL matrices, respectively. When the dye in the ORMOSIL matrices was coumarin 153, the useful lifetime was still 2250 pulses. Further studies by Altman and colleagues, in 1991, demonstrated useful lifetimes of 11 000 pulses at a 30 Hz prf when rhodamine 6G perchlorate was incorporated into ORMOSIL matrices. The useful lifetime increased to 14 500 pulses when the dye used was rhodamine B.

When incorporated into xerogel matrices, pyrromethene dyes lase with efficiencies as high as the highest obtained in organic hosts but with much higher photostability: a useful lifetime of 500 000 pulses at 20 Hz repetition rate was obtained by Faloss and colleagues, in 1997, with PM597 using a pump energy of 1 mJ. The efficiency of PM597 dropped to 40% but its useful lifetime increased to over 1 000 000 pulses when oxygen-free samples were prepared, in agreement with previous results that documented the strong dependence of laser parameters of pyrromethene dyes on the presence of oxygen. A similar effect was observed with perylene dyes: deoxygenated xerogel samples of Perylene Red exhibited a useful lifetime of 250 000 pulses at a prf of 5 Hz and 1 mJ pump energy, to be compared with a useful lifetime of only 10 000 pulses obtained with samples prepared in normal conditions. Perylene Orange incorporated into polycom glass and pumped longitudinally at 532 nm gave an efficiency of 72% and a useful lifetime of 40 000 pulses at a prf of 1 Hz as reported by Rhan and King, in 1998.

A direct comparison study of laser performance of Rh6G, PM567, Perylene Red, and Perylene Orange in organic, inorganic, and hybrid hosts was carried out by Rahn and King in 1995. They found that the nonpolar perylene dyes had better performance in partially organic hosts, whereas the ionic rhodamine and pyrromethene dyes performed best in the inorganic sol-gel glass host. The most promising combinations of dye and host for efficiency and photostability were found to be Perylene Orange in polycom glass and Rh6G in sol-gel glass.

An all-solid-state optical configuration, where the pump was a laser diode array side-pumped, Q-switched, frequency-doubled, Nd:YAG slab laser, was demonstrated by Hu and colleagues, in 1998, with dye DCM incorporated into ORMOSIL matrices. A lasing efficiency of 18% at the wavelength of 621 nm was obtained, and 27 000 pulses were needed for the output energy to decrease by 10% of its initial value at 30 Hz repetition rate. After 50 000 pulses the output energy decreased by 90% but it could be recovered after waiting for a few minutes without pumping.

Violet and ultraviolet laser dyes have also been incorporated into sol-gel silica matrices and their lasing properties evaluated under transversal pumping by a number of authors. Although reasonable efficiencies have been obtained with some of these laser dyes, the useful lifetimes are well below 1000 pulses.

More recently, Costela and colleagues have demonstrated improved conversion efficiencies in dye-doped inorganic–organic matrices using pyrromethene 567 dye. The solid matrix in this case is poly-trimethylsilyl-methacrylate cross-linked with ethylene glycol and copolymerized with methyl methacrylate. Also, Duarte and James have reported on dye-doped polymer-nanoparticle laser media where the polymer is PMMA and the nanoparticles are made of silica. These authors report TEM_{00} laser beam emission and improved dn/dT coefficients of the gain media that results in reduced $\Delta\theta$.

Dye Laser Applications

Dye lasers generated a renaissance in diverse applied fields such as isotope separation, medicine, photochemistry, remote sensing, and spectroscopy. Dye lasers have also been used in many experiments that have advanced the frontiers of fundamental physics.

Central to most applications of dye lasers is their capability of providing coherent narrow-linewidth radiation that can be tuned continuously from the near ultraviolet, throughout the visible, to the near infrared. These unique coherent and spectral properties have made dye lasers particularly useful in the field of high-resolution atomic and molecular spectroscopy. Dye lasers have been successfully and extensively used in absorption and fluorescence spectroscopy, Raman spectroscopy, selective excitations, and nonlinear spectroscopic techniques. Well documented is their use in photochemistry, that is, the chemistry of optically excited atoms and molecules, and in biology, studying biochemical reaction kinetics of biological molecules. By using nonlinear optical techniques, such as harmonic and sum frequency and difference frequency generation, the properties of dye lasers can be extended to the ultraviolet.

Medical applications of dye lasers include cancer photodynamic therapy, treatment of vascular lesions, and lithotripsy as described by Goldman in 1990.

The ability of dye lasers to provide tunable sub-GHz linewidths in the orange-red portion of the spectrum, at kW average powers, made them particularly suited to atomic vapor laser isotope separation of uranium as reported by Bass and colleagues, in 1992, Singh and colleagues, in 1994, and Nemoto and colleagues, in 1995. In this particular case the isotopic shift between the two uranium isotopes is a few GHz. Using dye lasers yielding $\Delta\nu \leq 1$ GHz the ^{235}U can be selectively excited in a multistep process by tuning the dye laser(s) to specific wavelengths, in the orange-red spectral region, compatible with a transition sequence leading to photoionization. Also, high-prf narrow-linewidth dye lasers, as described by Duarte and Piper, in 1984, are ideally suited for guide star applications in astronomy. This particular application requires a high-average power diffraction-limited laser beam at $\lambda \approx 589$ nm to propagate for some 95 km, above the surface of the Earth, and illuminate a layer of sodium atoms. Ground detection of the fluorescence from the sodium atoms allows measurements on the turbulence of the atmosphere. These measurements are then used to control the adaptive optics of the telescope thus compensating for the atmospheric distortions.

A range of industrial applications is also amenable to the wavelength agility and high-average power output characteristics of dye lasers. Furthermore, tunability coupled with narrow-linewidth emission, at large pulsed energies, make dye lasers useful in military applications such as directed energy and damage of optical sensors.

The Future of Dye Lasers

Predicting the future is rather risky. In the early 1990s, articles and numerous advertisements in commercial laser magazines predicted the demise and even the oblivion of the dye laser in a few years. It did not turn out this way. The high power and wavelength agility available from dye lasers have ensured their continued use as scientific and research tools in physics, chemistry, and medicine. In addition, the advent of narrow-linewidth solid-state dye lasers has sustained a healthy level of research and development in the field. Today, some 20 laboratories around the world are engaged in this activity.

The future of solid-state dye lasers depends largely on synthetic and manufacturing improvements. There is a need for strict control of the conditions of fabrication and a careful purification of all the compounds involved. In the case of polymers, in particular, stringent control of thermal conditions during the polymerization step is obligatory to ensure that adequate optical uniformity of the polymer matrix is achieved, and that intrinsic anisotropy developed during polymerization is minimized. From

an engineering perspective what are needed are dye-doped solid-state media with improved photostability characteristics and better thermal properties. By better thermal properties is meant better dn/dT factors. To a certain degree progress in this area has already been reported. As in the case of liquid dye lasers, the lifetime of the material can be improved by using pump lasers at compatible wavelengths. In this regard, direct diode laser pumping of dye-doped solid-state matrices should lead to very compact, long-lifetime, tunable lasers emitting throughout the visible spectrum. An example of such a device would be a green laser-diode-pumped rhodamine-6G doped MPMMA tunable laser configured in a multiple-prism grating architecture.

As far as liquid lasers are concerned, there is a need for efficient water-soluble dyes. Successful developments in this area were published in the literature with coumarin-analog dyes by Chen and colleagues, in 1988, and some encouraging results with rhodamine dyes have been reported by Ray and colleagues, in 2002. Efficient water-soluble dyes could lead to significant improvements and simplifications in high-power tunable lasers.

Further Reading

Costela, A., García-Moreno, I., Sastre, R., 2001. In: Nalwa, H.S. (Ed.), Handbook of Advanced Electronic and Photonic Materials and Devices, vol. 7. San Diego: Academic Press, pp. 161–208.
Diels, J.-C., Rudolph, W., 1996. Ultrashort Laser Pulse Phenomena. New York: Academic.
Demtröder, W., 1995. Laser Spectroscopy, 2nd edn Berlin: Springer-Verlag.
Duarte, F.J. (Ed.), 1991. High Power Dye Lasers. Berlin: Springer-Verlag.
Duarte, F.J. (Ed.), 1995. Tunable Lasers Handbook. New York: Academic.
Duarte, F.J. (Ed.), 1995. Tunable Laser Applications. New York: Marcel Dekker.
Duarte, F.J., 2003. Tunable Laser Optics. New York: Elsevier Academic.
Duarte, F.J., Hillman, L.W. (Eds.), 1990. Dye Laser Principles. New York: Academic Press.
Maeda, M., 1984. Laser Dyes. New York: Academic Press.
Schäfer, F.P. (Ed.), 1990. Dye Lasers, 3rd edn Berlin: Springer-Verlag.
Radziemski, L.J., Solarz, R.W., Paisner, J.A. (Eds.), 1987. Laser Spectroscopy and its Applications. New York: Marcel Dekker.
Weber, M.J., 2001. Handbook of Lasers. New York: CRC.

Edge Emitters

JJ Coleman, University of Illinois, Urbana, IL, USA

Introduction

The semiconductor edge emitting diode laser is a critical component in a wide variety of applications, including fiber optics telecommunications, optical data storage, and optical remote sensing. In this section, we describe the basic structures of edge emitting diode lasers and the physical mechanisms for converting electrical current into light. Laser waveguides and cavity resonators are also outlined. The power, efficiency, gain, loss, and threshold characteristics of laser diodes are presented along with the effects of temperature and modulation. Quantum well lasers are outlined and a description of grating coupled lasers is provided.

Like all forms of laser, the edge emitting diode laser is an oscillator and has three principal components; a mechanism for converting energy into light, a medium that has positive optical gain, and a mechanism for obtaining optical feedback. The basic edge emitting laser diode is shown schematically in **Fig. 1**. Current flow is vertical through a pn junction and light is emitted from the ends. The dimensions of the laser in **Fig. 1** are distorted, in proportion to typical real laser diodes, to reveal more detail. In practical laser diodes, the width of the stripe contact is actually much smaller than shown and the thickness is smaller still. Typical dimensions are (stripe width × thickness × length) $2 \times 100 \times 1000$ μm. Light emission is generated only within a few micrometers of the surface, which means that 99.98% of the volume of this small device is inactive.

Energy conversion in diode lasers is provided by current flowing through a forward-biased pn junction. At the electrical junction, there is a high density of injected electrons and holes which can recombine in a direct energy gap semiconductor material to give an emitted photon. Charge neutrality requires equal numbers of injected electrons and holes. **Fig. 2** shows a simplified energy versus momentum (E–k) diagrams for (a) direct, and (b) indirect semiconductors. In a direct energy gap material, such as GaAs, the energy minima have the same momentum vector, so recombination of an electron in the conduction band and a hole in the valence band takes place directly. In an indirect material, such as silicon, momentum conservation requires the participation of a third particle – a phonon. This third-order process is far less efficient than direct recombination and, thus far, too inefficient to support laser action.

Optical gain in direct energy gap materials is obtained when population inversion is reached. Population inversion is the condition where the probability for stimulated emission exceeds that of absorption. The density of excited electrons n, in the conduction band is given by

$$n = \int_{E_c}^{\infty} \rho_n(E) f_n(E) \mathrm{d}E \ \mathrm{cm}^{-3} \tag{1}$$

where $\rho_n(E)$ is the conduction band density of states function and $f_n(E)$ is the Fermi occupancy function. A similar equation can be

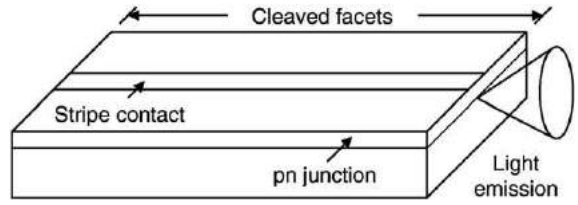

Fig. 1 Schematic drawing of an edge emitting laser diode.

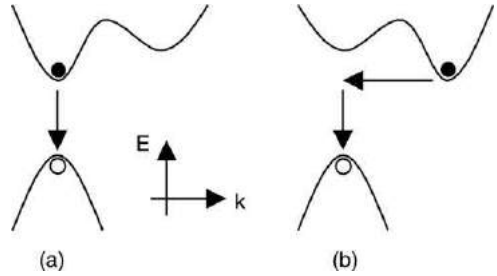

Fig. 2 Energy versus momentum for (a) direct, and (b) indirect semiconductors.

 doi:10.1016/B978-0-12-809283-5.00852-1

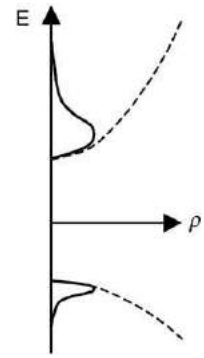

Fig. 3 Density of states versus energy for conduction and valence bands.

written for holes in the valence band. Fig. 3 shows the density of states ρ versus energy as a dashed line for conduction and valence bands. The solid lines in Fig. 3 are the $\rho(E)f(E)$ products, and the areas under the solid curves are the carrier densities n and p.

The Fermi functions are occupation probabilities based on the quasi-Fermi levels E_{Fn} and E_{Fp}, which are, in turn, based on the injected carrier densities, δn and δp. Population inversion implies that the difference between the quasi-Fermi levels must exceed the emission energy which, for semiconductors, must be greater than the bandgap energy:

$$E_{Fn} - E_{Fp} \geq \hbar\omega \sim E_g \tag{2}$$

Charge neutrality requires that $\delta n = \delta p$. Transparency is the point where these two conditions are just met and any additional injected current results in optical gain. The transparency current density J_o, is given by

$$J_o = \frac{qn_oL_z}{\tau_{spon}} \tag{3}$$

where n_o is the transparency carrier density ($\delta n = \delta p = n_o$ at transparency), L_z is the thickness of the optically active layer, and τ_{spon} is the spontaneous carrier lifetime in the material (typically 5 nsec).

A resonant cavity for optical feedback is obtained simply by taking advantage of the natural crystal formation in single crystal semiconductors. Most III–V compound semiconductors form naturally into a zincblende lattice structure. With the appropriate choice of crystal planes and surfaces, this lattice structure can be cleaved such that nearly perfect plane–parallel facets can be formed at arbitrary distances from each other. Since the refractive index of these materials is much larger (~ 3.5) than that of air, there is internal optical reflection for normal incidence of approximately 30%. This type of optical cavity is called a Fabry–Perot resonator.

Effective lasers of all forms make use of optical waveguides to optimize the overlap of the optical field with material gain and cavity resonance. The optical waveguide in an edge emitting laser is best described by considering it in terms of a transverse waveguide component, defined by the epitaxial growth of multiple thin heterostructure layers, and a lateral waveguide component, defined by conventional semiconductor device processing methods. One of the simplest, and yet most common, heterostructure designs is shown in Fig. 4. This five-layer separate confinement heterostructure (SCH) offers excellent carrier confinement in the active layer as well as a suitable transverse waveguide.

Clever choice of different materials for each layer results in the energy band structure, index of refraction profile, and optical field mode profile shown in Fig. 5. The outer confining layers are the thickest (~ 1 μm) and have the largest energy bandgap and lowest refractive index. The inner barrier layers are thinner (~ 0.1 μm), have intermediate bandgap energy and index of refraction, and serve both an optical role in defining the optical waveguide and an electronic role in confining carriers to the active layer. The active layer is the narrowest bandgap, highest index material, and is the only layer in the structure designed to provide optical gain. It may be a single layer or, as in the case of a multiple quantum well laser, a combination of layers. The bandgap energy, E, of this active layer plays a key role in determining the emission wavelength, λ, of the laser:

$$E = \frac{hc}{\lambda} \tag{4}$$

where h is Planck's constant and c is the speed of light.

The bandgap energy of the SCH structure is shown in Fig. 5. The lowest energy active layer collects injected electrons and holes and the carrier confinement provided by the inner barrier layers allows the high density of carriers necessary for population inversion and gain. The layers in the structure also have the refractive index profile shown in Fig. 5 which forms a five-layer slab dielectric waveguide. With appropriate values for thicknesses and indices of refraction, this structure can be a very strong fundamental mode waveguide with the field profile shown. A key parameter of edge emitting diode lasers is Γ, the optical confinement factor. This parameter is the areal overlap of the optical field with the active layer, shown as dashed lines in the figure and can be as little as a few percent, even in very high performance lasers.

Lateral waveguiding in a diode laser can be accomplished in a variety of ways. A common example of an index guided laser is the ridge waveguide laser shown in Fig. 6. After the appropriate transverse waveguide heterostructure sample is grown,

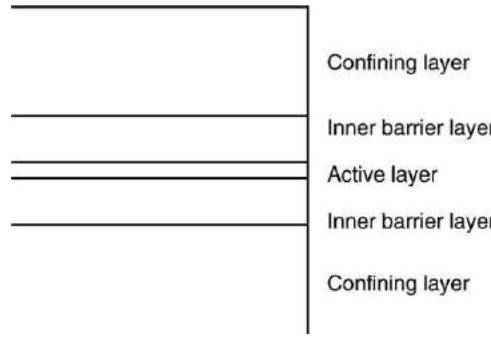

Fig. 4 Schematic cross-section of a typical five-layer separate confinement heterostructure (SCH) laser.

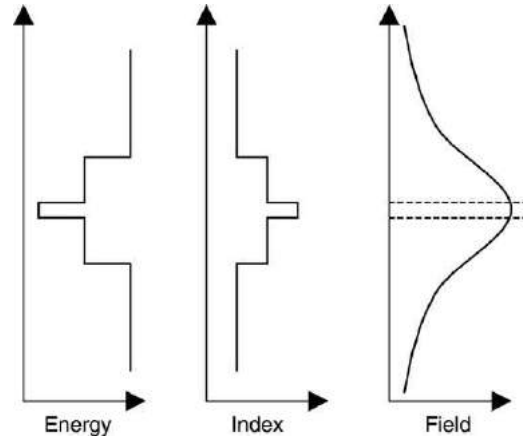

Fig. 5 Energy band structure, index of refraction profile, and optical field mode profile of the SCH laser of Fig. 4.

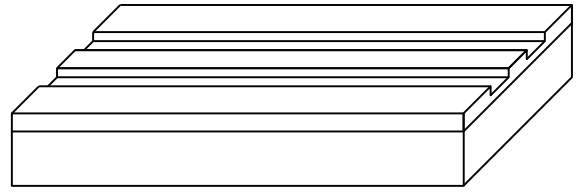

Fig. 6 Schematic diagram of a real index guided ridge waveguide diode laser.

conventional semiconductor processing methods are used to form parallel etched stripes from the surface near, but not through, the active layer. An oxide mask is patterned such that electrical contact is formed only to the center (core) region between the etched stripes, the core stripe being only a few microns wide. The lateral waveguide that results arises from the average index of refraction (effective index) in the etched regions outside the core being smaller than that of the core. Fig. 7 shows the asymmetric near field emission profile of the ridge waveguide laser and cross-sectional profiles of the laser emission and effective index of refraction.

The optical spectrum of an edge emitting diode laser below and above threshold is the superposition of the material gain of the active layer and the resonances provided by the Fabry–Perot resonator. The spectral variation of material gain with injected carrier density (drive current) is shown in Fig. 8. As the drive is increased, the intensity and energy of peak gain both increase.

The Fabry–Perot resonator has resonances every half wavelength, given by

$$L = \frac{m}{2}\frac{\lambda}{n} \tag{5}$$

where L is the cavity length, λ/n is the wavelength in the medium (n is the refractive index), and m is an integer. For normal length edge emitter cavities, these cavity modes are spaced only 1 or 2 Å apart. The result is optical spectra, near laser threshold and above, that look like the spectra of Fig. 9. Many cavity modes resonate up to threshold, while above threshold, the superlinear increase in emission intensity with drive current tends to greatly favor one or two of the Fabry–Perot longitudinal cavity modes.

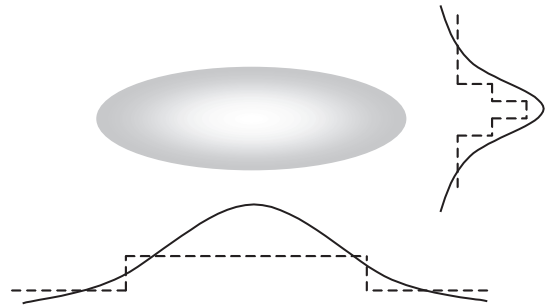

Fig. 7 Near field emission pattern, cross-sectional emission profiles (solid lines), and refractive index profiles (dashed lines) from an index guided diode laser.

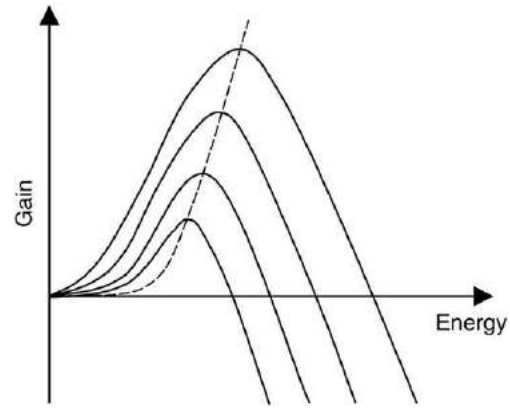

Fig. 8 Gain versus emission energy for four increasing values of carrier (current) density. Peak gain is shown as a dashed line.

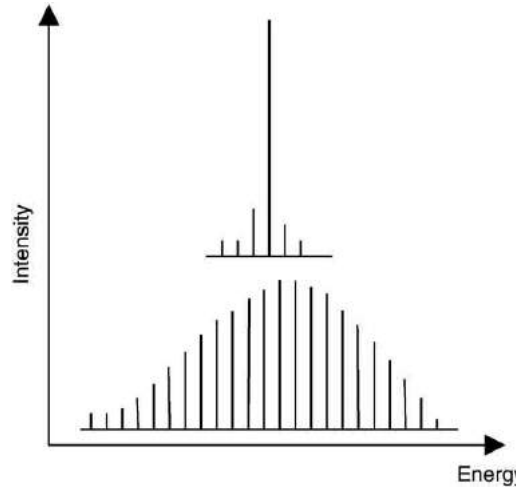

Fig. 9 Emission spectra from a Fabry–Perot edge emitting laser structure at, and above, laser threshold.

Laser threshold can be determined from a relatively simple analysis of the round trip gain and losses in the Fabry–Perot resonator. If the initial intensity is I_o, after one round trip the intensity will be given by

$$I = I_o R_1 R_2 e^{(g-\alpha_i)2L} \tag{6}$$

where R_1 and R_2 are the reflectivities of the two facets, g is the optical gain, α_i are losses associated with optical absorption (typically 3–15 cm^{-1}), and L is the cavity length. At laser threshold, $g=g_{th}$ and $I=I_o$, i.e., the round trip gain exactly equals the losses. The result is

$$g_{th} = \alpha_i + \frac{1}{2L}\ln\frac{1}{R_1 R_2} \tag{7}$$

where the second term represents the mirror losses α_m. Of course, mirror losses are desirable in the sense that they represent useable output power. Actually, the equation above does not take into account the incomplete overlap of the optical mode with the gain in the active layer. The overlap is defined by the confinement factor Γ, described above, and the equation must be modified to become

$$\Gamma g_{th} = \alpha_i + \frac{1}{2L}\ln\frac{1}{R_1 R_2} \tag{8}$$

For quantum well lasers, the material peak gain, shown in **Fig. 8**, as a function of drive current is given approximately by

$$g_t = \beta J_o \ln\frac{J}{J_o} \tag{9}$$

where J_o is the transparency current density. The threshold current density J_{th}, which is the current density where the gain exactly equals all losses, is given by

$$J_{th} = \frac{J_o}{\eta_i}\exp\frac{\alpha + \frac{1}{2L}\ln\frac{1}{R_1 R_2}}{\Gamma\beta J_o} \tag{10}$$

Note that an additional parameter η_i has appeared in this equation. Not all of the carriers injected by the drive current participate in optical processes; some are lost to other nonradiative processes. This is accounted for by including the internal quantum efficiency term η_i, which is typically at least 90%. The actual drive current, of course, must include the geometry of the device and is given by

$$I_{th} = wLJ_{th} \tag{11}$$

where L is the cavity length, and w is the effective width of the active volume. The effective width is basically the stripe width of the core but also includes current spreading and carrier diffusion under the stripe.

The power–current characteristic (P–I) for a typical edge emitting laser diode is shown in **Fig. 10**. As the drive current is increased from zero, the injected carrier densities increase and spontaneous recombination is observed. At some point the gain equals the internal absorption losses, and the material is transparent. Then, as the drive current is further increased, additional gain eventually equals the total losses, including mirror losses, and laser threshold is reached. Above threshold, nearly all additional injected carriers contribute to stimulated emission (laser output). The power generated internally is given by

$$P = wL(J - J_{th})\eta_i \frac{h\nu}{q} \tag{12}$$

where wL is the area and $h\nu$ is the photon energy. Only a fraction of the power generated internally is extracted from the device

$$P_o = wL(J - J_{th})\eta_i \frac{h\nu}{q}\frac{\alpha_m}{\alpha_i + \alpha_m} \tag{13}$$

Clearly a useful design goal is to minimize the internal optical loss α_i. The slope of the P–I curve of **Fig. 10** above threshold is described by an external differential quantum efficiency η_d, which is related to the internal quantum efficiency by

$$\eta_d = \eta_i\left[\frac{\alpha_m}{\alpha_i + \alpha_m}\right] \tag{14}$$

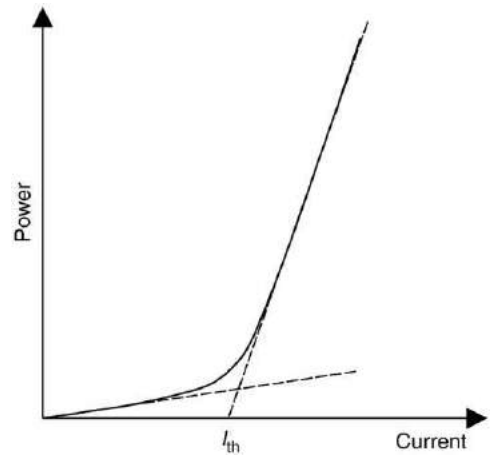

Fig. 10 Power versus current (P–I) curve for a typical edge emitting laser diode.

Other common parameters used to characterize the efficiency of laser diodes include the power conversion efficiency η_p

$$\eta_p = \frac{P_o}{V_F I} \tag{15}$$

and the wall plug efficiency η_w

$$\eta_w = \frac{P_o}{I} \tag{16}$$

The power conversion efficiency η_p, recognizes that the forward bias voltage V_F is larger than the photon energy by an amount associated with the series resistance of the laser diode. The wall plug efficiency η_w, has the unusual units of WA^{-1} and is simply a shorthand term that relates the common input unit of current to the common output unit of power. In high performance low power applications, the ideal device has minimum threshold current since the current used to reach threshold contributes little to light emission. In high power applications, the threshold current becomes insignificant and the critical parameter is external quantum efficiency.

Temperature effects may be important for edge emitting lasers, depending on a particular application. Usually, concerns related to temperature arise under conditions of high temperature operation (T > 50°C), where laser thresholds rise and efficiencies fall. It became common early in the development of laser diodes to define a characteristic temperature T_o for laser threshold, which is given by

$$I_{th} = I_{th_o} \exp\left(\frac{T}{T_o}\right) \tag{17}$$

where a high T_o is desirable. Unfortunately, this simple expression does not describe the temperature dependence very well over a wide range of temperatures, so the appropriate temperature range must also be specified. For applications where the emission wavelength of the laser is important, the change in wavelength with temperature $d\lambda/dT$ may also be specified. For conventional Fabry–Perot cavity lasers, $d\lambda/dT$ is typically 3–5 $ÅC^{-1}$.

In order for a diode laser to carry information, some form of modulation is required. One method for obtaining this is direct modulation of the drive current in a semiconductor laser. The transient and temporal behavior of lasers is governed by rate equations for carriers and photons which are necessarily coupled equations. The rate equation for carriers is given by

$$\frac{dn}{dt} = \frac{J}{qL_z} - \frac{n}{\tau_{sp}} - (c/n)\beta(n - n_o))\varphi(E) \tag{18}$$

where J/qL_z is the supply, n/τ_{sp} is the spontaneous emission rate, and the third term is the stimulated emission rate $(c/n)\beta(n - n_o)\varphi(E)$ where β is the gain coefficient and $)\varphi(E)$ is the photon density. The rate equation for photons is given by

$$\frac{d\varphi}{dt} = (c/n)\beta(n - n_o)\varphi + \frac{\theta n}{\tau_{sp}} - \frac{\varphi}{\tau_p} \tag{19}$$

where θ is the fraction of the spontaneous emission that couples into the mode (a small number), and τ_p is the photon lifetime ($\sim$1 psec) in the cavity. These rate equations can be solved for specific diode laser materials and structures to yield a modulation frequency response. A typical small signal frequency response, at two output power levels, is shown in Fig. 11. The response is typically flat to frequencies above 1 GHz, rises to a peak value at some characteristic resonant frequency, and quickly rolls off. The characteristic small signal resonant frequency ω_r, is given by

$$\omega_r^2 = \frac{(c/n)\beta\overline{\varphi}}{\tau_p} \tag{20}$$

where $\overline{\varphi}$ is the average photon density. Direct modulation of edge emitting laser diodes is limited by practicality to less than 10 GHz. This is in part because of the limits imposed by the resonant frequency and in part by chirp. Chirp is the frequency modulation that arises indirectly from direct current modulation. Modulation of the current modulates the carrier densities which,

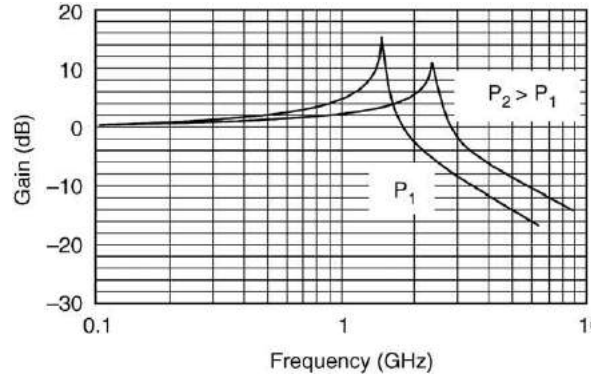

Fig. 11 Direct current modulation frequency response for an edge emitting diode laser at two output power levels.

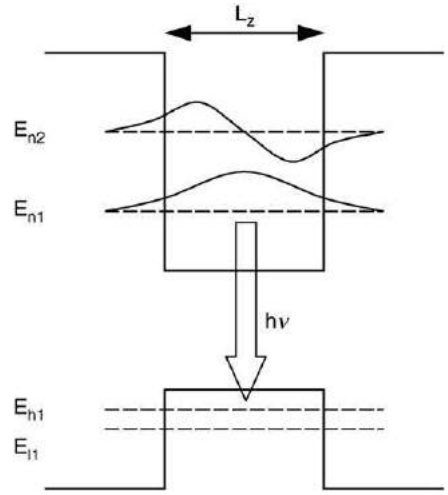

Fig. 12 Energy band diagram for a quantum well heterostructure.

in turn, results in modulation of the quasi-Fermi levels. The separation of the quasi-Fermi levels is the emission energy (wavelength). Most higher-speed lasers make use of external modulation schemes.

The discussion thus far has addressed aspects of semiconductor diode edge emitting lasers that are common to all types of diode lasers irrespective of the choice of materials or the details of the active layer structure. The materials of choice for efficient laser devices include a wide variety of III–V binary compounds and ternary or quaternary alloys. The particular choice is based on such issues as emission wavelength, a suitable heterostructure lattice match, and availability of high-quality substrates. For example, common low-power lasers for CD players ($\lambda \sim 820$ nm) are likely to be made on a GaAs substrate using a GaAs–$Al_xGa_{1-x}As$ heterostructure. Lasers for fiberoptic telecommunications systems ($\lambda \sim 1.3$–1.5 μm) are likely to be made on an InP substrate using an InP–$In_xGa_{1-x}As_yP_{1-y}$ heterostructure. Red laser diodes ($\lambda \sim 650$ nm) are likely to be made on a GaAs substrate using a GaAs–$In_xGa_{1-x}As_yP_{1-y}$ heterostructure. Blue and ultraviolet lasers, a relatively new technology compared to the others, make use of $Al_xGa_{1-x}N$–GaN– $In_yGa_{1-y}N$ heterostructures formed on sapphire or silicon carbide substrates.

An important part of many diode laser structures is a strained layer. If a layer is thin enough, the strain arising from a modest amount of lattice mismatch can be accommodated elastically. In thicker layers, lattice mismatch results in an unacceptable number of dislocations that affect quantum efficiency, optical absorption losses, and, ultimately, long-term failure rates. Strained layer GaAs–$In_xGa_{1-x}As$ lasers are a critical component in rare-earth doped fiber amplifiers.

The structure of the active layer in edge emitting laser diodes was originally a simple double heterostructure configuration with a single active layer of 500–1000 Å in thickness. In the 1970s however, advances in the art of growing semiconductor heterostructure materials led to growth of high- quality quantum well active layers. These structures, having thicknesses that are comparable to the electron wavelength in a semiconductor (<200 Å), revolutionized semiconductor lasers, resulting in much lower threshold current densities, higher efficiencies, and a broader range of available emission wavelengths. The energy band diagram for a quantum well laser active region is shown in **Fig. 12**.

This structure is a physical realization of the particle-in-a-box problem in elementary quantum mechanics. The quantum well yields quantum states in the conduction band at discrete energy levels, with odd and even electron wavefunctions, and breaks the degeneracy in the valence band, resulting in separate energy states for light holes and heavy holes. The primary transition for recombination is from the $n=1$ electron state to the $h=1$ heavy hole state and takes place at a higher energy than the bulk bandgap energy. In addition, the density of states for quantum wells becomes a step-like function and optical gain is enhanced. The advantages in practical edge emitting lasers that are the result of one-dimensional quantization are such that virtually all present commercial laser diodes are quantum well lasers. These kinds of advantages are driving development of laser diodes with additional degrees of quantization, including two dimensions (quantum wires) and three dimensions (quantum dots).

The Fabry–Perot cavity resonator described here is remarkably efficient and relatively easy to fabricate, which makes it the cavity of choice for many applications. There are many applications, however, where the requirements for linewidth of the laser emission or the temperature sensitivity of the emission wavelength make the Fabry–Perot resonator less desirable. An important laser technology that addresses both of these concerns involves the use of a wavelength selective grating as an integral part of the laser cavity. A notable example of a grating coupled laser is distributed feedback laser shown schematically in **Fig. 13(a)**. This is similar to the SCH cross-section of **Fig. 4** with the addition of a Bragg grating above the active layer.

The period Λ, of the grating is chosen to fulfill the Bragg condition for mth-order coupling between forward- and backward-propagating waves, which is

$$\Lambda = \frac{m\lambda}{2n} \tag{21}$$

where λ/n is the wavelength in the medium. The index step between the materials on either side of the grating is small and the

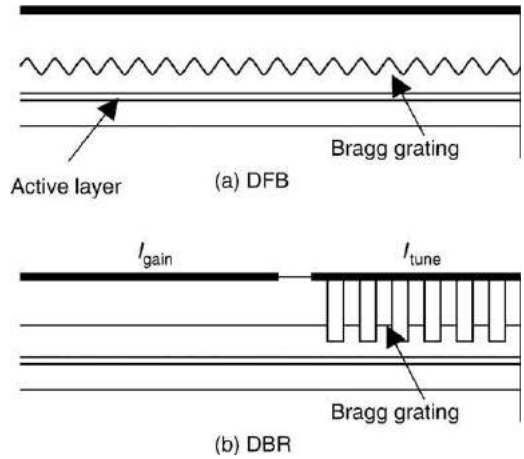

Fig. 13 Schematic diagram of distributed feedback (DFB) and distributed Bragg reflector (DBR) edge emitting laser diodes.

amount of reflection is also small. Since the grating extends throughout the length of the cavity, however, the overall effective reflectivity can be large. Another variant is the distributed Bragg reflector (DBR) laser resonator, shown in Fig. 13(b). This DBR laser has a higher index contrast, deeply etched surface grating located at one or both ends of the otherwise conventional SCH laser heterostructure. One of the advantages of this structure is the opportunity to add a contact for tuning the Bragg grating by current injection.

Both the DBR and DFB laser structures are particularly well suited for telecommunications or other applications where a very narrow single laser emission line is required. In addition, the emission wavelength temperature dependence for this type of laser is much smaller, typically 0.5 $\text{Å}C^{-1}$.

Further Reading

Agrawal, G.P. (Ed.), 1995. Semiconductor Lasers Past, Present, and Future. Woodbury, NY: American Institute of Physics.
Agrawal, G.P., Dutta, N.K., . Semiconductor Lasers. New York: Van Nostrand Reinhold.
Bhattacharya, P., 1994. Semiconductor Optoelectronic Devices. Upper Saddle River, NJ: Prentice-Hall.
Botez, D., Scifres, D.R. (Eds.), 1994. Diode Laser Arrays. Cambridge, UK: Cambridge University Press.
Chuang, S.L., 1995. Physics of Optoelectronic Devices. New York: John Wiley & Sons.
Coleman, J.J., 1992. Selected Papers on Semiconductor Diode Lasers. Bellingham, WA: SPIE Optical Engineering Press.
Einspruch, N.G., Frensley, W.R. (Eds.), 1994. Heterostructures and Quantum Devices. San Diego, CA: Academic Press.
Iga, K., 1994. Fundamentals of Laser Optics. Plenum Press.
Kapon, E., 1999. Semiconductor Lasers I and II. San Diego, CA: Academic Press.
Nakamura, S., Fasol, G., 1997. The Blue Laser Diode. Berlin: Springer.
Verdeyen, J.T., 1989. Laser Electronics, 3rd edn. Englewood Cliffs, NJ: Prentice-Hall.
Zory Jr, P.S. (Ed.), 1993. Quantum Well Lasers. San Diego, CA: Academic Press.

Excimer Lasers

JJ Ewing, Ewing Technology Associates, Inc., Bellevue, WA, USA

Introduction

We present a summary of the fundamental operating principles of ultraviolet, excimer lasers. After a brief discussion of the economics and application motivation, the underlying physics and technology of these devices is described. Key issues and limitations in the scaling of these lasers are presented.

Background: Why Excimer Lasers?

Excimer lasers have become the most widely used source of moderate power pulsed ultraviolet sources in laser applications. The range of manufacturing applications is also large. Production of computer chips, using excimer lasers as the illumination source for lithography, has by far the largest commercial manufacturing impact. In the medical arena, excimer lasers are used extensively in what is becoming one of the world's most common surgical procedures, the reshaping of the lens to correct vision problems. Taken together, the annual production rate for these lasers is a relatively modest number compared to the production of semiconductor lasers. Current markets are in the range of $0.4 billion per year (Fig. 1). More importantly, the unique wavelength, power, and pulse energy properties of the excimer laser enable a systems and medical procedure market, based on the excimer laser, to exceed $3 billion per year. The current average sales price for these lasers is in the range of $250 000 per unit, driven primarily by the production costs and reliability requirements of the lasers for lithography for chip manufacture. Relative to semiconductor diode lasers, excimer lasers have always been, and will continue to be, more expensive per unit device by a factor of order 10^4. UV power and pulse energy, however, are considerably higher.

The fundamental properties of these lasers enable the photochemical and photophysical processes used in manufacturing and medicine. Short pulses of UV light can photo-ablate materials, being of use in medicine and micromachining. Short wavelengths enable photochemical processes. The ability to provide a narrow linewidth using appropriate resonators enables precise optical processes, such as lithography. We trace out some of the core underlying properties of these lasers that lead to the core utility. We also discuss the technological problems and solutions that have evolved to make these lasers so useful.

The data in Fig. 1 provide a current answer to the question of 'why excimer lasers?' At the dawn of the excimer era, the answer to this question was quite different. In the early 1970s, there were no powerful short wavelength lasers, although IR lasers were being scaled up to significant power levels. However, visible or UV lasers could in principle provide better propagation and focus to a very small spot size at large distances. Solid state lasers of the time offered very limited pulse repetition frequency and low

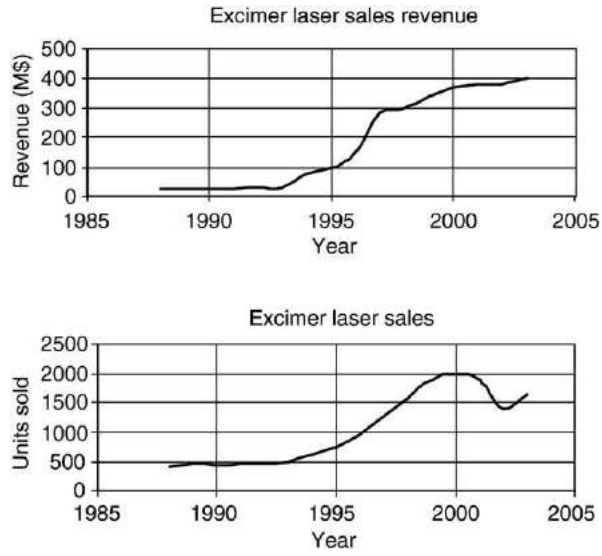

Fig. 1 The sales of excimer lasers have always been measured in small numbers relative to other, less-expensive lasers. For many years since their commercial release in 1976 the sales were to R&D workers in chemistry, physics and ultimately biology and medicine. The market for these specialty but most useful R&D lasers was set to fit within a typical researcher's annual capital budget, i.e., less than $100K. As applications and procedures were developed and certified or approved, sales and average unit sales price increased considerably. Reliance on cyclical markets like semiconductor fabrication led to significant variations in production rates and annual revenues as can be seen.

 doi:10.1016/B978-0-12-809283-5.00846-6

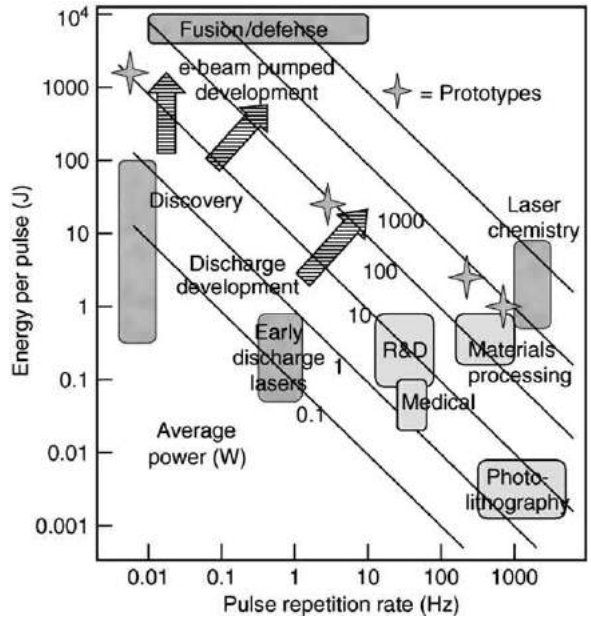

Fig. 2 The typical energy and pulse rate for certain applications and technologies using excimer lasers. For energy under a few J, a discharge laser is used. The earliest excimer discharge lasers were derivatives of CO_2 TEA (transversely excited atmospheric pressure) and produced pulse energy in the range of 100 mJ per pulse. Pulse rates of order 30 Hz were all that the early pulsed power technology could provide. Early research focused on generating high energy. Current markets are at modest UV pulse energy and high pulse rates for lithography and low pulse rates for medical applications.

efficiency. Diode pumping of solid state lasers, and diode arrays to excite such lasers, was a far removed development effort. As such, short-wavelength lasers were researched over a wide range of potential media and lasing wavelengths. The excimer concept was one of those candidates.

We provide, in **Fig. 2**, a 'positioning' map showing the range of laser parameters that can be obtained with commercial or developmental excimer lasers. The map has coordinates of pulse energy and pulse rate, with diagonal lines expressing average power. We show where both development goals and current markets lay in a map. Current applications are shown by the boxes in the lower right-hand corner. Excimer lasers have been built with clear apertures in the range of a fraction of 0.1 to 10^4 cm^2, with single pulse energies covering a range of 10^7. Lasers with large apertures require electron beam excitation. Discharge lasers correspond to the more modest pulse energy used for current applications. In general, for energies lower than ~ 5 J, the excitation method of choice is a self-sustained discharge, clearly providing growth potential for future uses, if required.

The applications envisaged in the early R&D days were in much higher energy uses, such as in laser weapons, laser fusion, and laser isotope separation. The perceived need for lasers for isotope separation, laser-induced chemistry, and blue-green laser-based communications from satellites, drove research in high pulse rate technology and reliability extension for discharge lasers. This research resulted in prototypes, with typical performance ranges as noted on the positioning map that well exceed current market requirements. However, the technology and underlying media physics developed in these early years made possible advances needed to serve the ultimate real market applications.

The very important application of UV lithography requires high pulse rates and high reliability. These two features differentiate the excimer laser used for lithography from the one used in the research laboratory. High pulse rates, over 2000 Hz in current practice, and the cost of stopping a computer chip production line for servicing, drive the reliability requirements toward 10^{10} shots per service interval. In contrast, the typical laboratory experiments are more often in the range of 100 Hz and, unlike a lithography production line, do not run 24 hours a day, 7 days a week. The other large market currently is in laser correction of myopia and other imperfections in the human eye. For this use the lasers are more akin to the commercial R&D style laser in terms of pulse rate and single pulse energy. Not that many shots are needed to ablate tissue to make the needed correction, and shot life is a straightforward requirement. However, the integration of these lasers into a certified and accepted medical procedure and an overall medical instrument drove the growth of this market.

The excimer concept, and the core technology used to excite these lasers, is applicable over a broad range of UV wavelengths, with utility at specific wavelength ranges from 351 nm in the near UV to 157 nm in the vacuum UV (VUV). There were many potential excimer candidates in the early R&D phase (**Table 1**). Some of these candidates can be highly efficient, $>50\%$ relative to deposited energy, at converting electrical power into UV or visible emission, and have found a parallel utility as sources for lamps, though with differing power deposition rates and configurations. The key point in the table is that these lasers are powerful and useful sources at very short wavelengths. For lithography, the wavelength of interest has consistently shifted to shorter and shorter wavelengths as the printed feature size decreased. Though numerous candidates were pursued, as shown in **Table 1**, the most useful

Table 1 Key excimer wavelengths

Excimer emitter	*λ (nm)*	*Comment*
Ar_2	126	Very short emission wavelength, deep in the Vacuum UV; inefficient as laser due to absorption, see Xe_2.
Kr_2	146	Broad band emitter and early excimer laser with low laser efficiency. Very efficient converter of electric power into Vacuum UV light.
F_2	157	Molecule itself is not an excimer, but uses lower level dissociation; candidate for next generation lithography.
Xe_2	172	The first demonstrated excimer laser. Very efficient formation and emission but excited state absorption limits laser efficiency. Highly efficient as a lamp.
ArF	193	Workhorse for corneal surgery and lithography.
KrCl	222	Too weak compared to KrF and not as short a wavelength as ArF.
KrF	248	Best intrinsic laser efficiency; numerous early applications but found major market in lithography. Significant use in other materials processing.
XeI	254	Never a laser, but high formation efficiency and excellent for a lamp. Historically the first rare gas halide excimer whose emission was studied at high pressure.
XeBr	282	First rare gas halide to be shown as a laser; inefficient laser due to excited state absorption, but excellent fluorescent emitter; a choice for lamps.
Br_2	292	Another halogen molecule with excimer like transitions that has lased but had no practical use.
XeCl	308	Optimum medium for laser discharge excitation.
Hg_2	335	Hg vapor is very efficiently excited in discharges and forms excimers at high pressure. Subject of much early research, but as in the case of the rare gas excimers such as Xe_2 suffer from excited state absorption. Despite the high formation efficiency, this excimer was never made to lase.
I_2	342	Iodine UV molecular emission and lasing was first of the pure halogen excimer-like lasers demonstrated. This species served as kinetic prototype for the F_2'honorary' excimer that has important use as a practical VUV source.
XeF	351, 353	This excimer laser transition terminates on a weakly bound lower level that dissociates rapidly, especially when heated. The focus for early defense-related laser development; as this wavelength propagates best of this class in the atmosphere.
XeF	480	This very broadband emission of XeF terminates on a different lower level than the UV band. Not as easily excited in a practical system, so has not had significant usage.
HgBr	502	A very efficient green laser that has many of the same kinetic features of the rare gas halide excimer lasers. But the need for moderately high temperature for the needed vapor pressure made this laser less than attractive for UV operation.
XeO	540	This is an excimer-like molecular transition on the side of the auroral lines of the O atom. The optical cross-section is quite low and as a result the laser never matured in practice.

lasers are those using rare gas halide molecules. These lasers are augmented by dye laser and Raman shifting to provide a wealth of useful wavelengths in the visible and UV.

All excimer lasers utilize molecular dissociation to remove the lower laser level population. This lower level dissociation is typically from an unbound or very weakly bound ground-state molecule. However, halogen molecules also provide laser action on excimer-like transitions, that terminate on a molecular excited state that dissociates quickly. The vacuum UV transition in F_2 is the most practical method for making very short wavelength laser light at significant power. Historically, the rare gas dimer molecules, such as Xe_2, were the first to show excimer laser action, although self absorption, low gain, and poor optics in the VUV limited their utility. Other excimer-like species, such as metal halides, metal rare gas continua on the edge of metal atom resonance lines, and rare gas oxides were also studied. The key differentiators of rare gas halides from other excimer-like candidates are strong binding in the excited state, lack of self absorption, and relatively high optical cross-section for the laser transition.

Excimer Laser Fundamentals

Excimer lasers utilize lower-level dissociation to create and sustain a population inversion. For example, in the first excimer laser demonstrated, Xe_2, the lasing species is one which is comprised of two atoms that do not form a stable molecule in the ground state, as xenon dimers in the ground state are not a stable molecule. In the lowest energy state, that with neither of the atoms electronically excited, the interaction between the atoms in a collision is primarily one of repulsion, save for a very weak 'van der Waals' attraction at larger internuclear separations. This is shown in the potential energy diagram of **Fig. 3**. If one of the atoms is excited, for example by an electric discharge, there can be a chemical binding in the electronically excited state of the molecule relative to the energy of one atom with an electron excited and one atom in the ground state. We use the shorthand notation * to denote an electronically excited atomic molecular species, e.g., Xe*. The excited dimer, excimer for short, will recombine rapidly at high pressure from atoms into the Xe_2* electronically excited molecule. Radiative lifetimes of the order 10 nsec to 1 μsec are typical

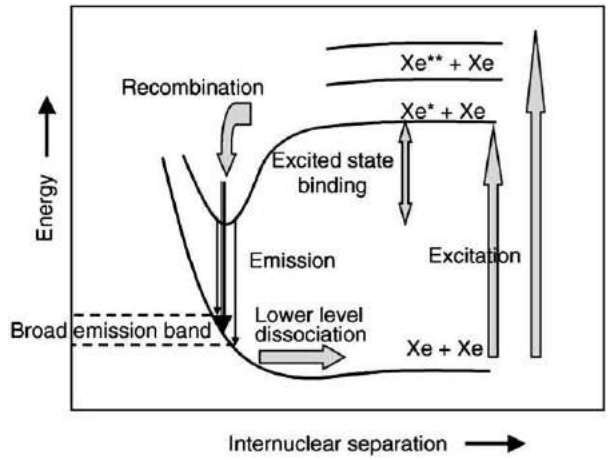

Fig. 3 A schematic of the potential energy curves for an excimer species such as Xe_2. The excited molecule is bound relative to an excited atom and can radiate to a lower level that rapidly dissociates on psec timescales since the ground state is not bound, save for weak binding due to van der Waals forces. The emission band is intrinsically broad.

before photon emission returns the molecule to the lower level. Collisional quenching often reduces the lifetime below the radiative rate. For Xe_2* excimers, the emission is in the vacuum ultraviolet (VUV).

There are a number of variations on this theme, as noted in **Table 1**. Species, other than rare gas pairs, can exhibit broadband emission. The excited state binding energy can vary significantly, changing the fraction of excited states that will form molecules. The shape of the lower potential energy curve can vary as well. From a semantic point of view heteronuclear, diatomic molecules, such as XeO, are not excimers, as they are not made of two of the same atoms. However, the more formal name 'exciplex' did not stick in the laser world; excimer laser being preferred. Large binding energy is more efficient for capturing excited atoms into excited excimer-type molecules in the limited time they have before emission. Early measurements of the efficiency of converting electrical energy deposited into fluorescent radiation showed that up to 50% of the energy input could result in VUV light emission. Lasers were not this efficient due to absorption.

The diagram shown in **Fig. 3** is simplified because it does not show all of the excited states that exist for the excited molecule. There can be closely lying excited states that share the population and radiate at a lower rate, or perhaps absorb to higher levels, also not shown in **Fig. 3**. Such excited state absorption may terminate on states derived from higher atomic excited states, noted as Xe** in the figure, or yield photo-ionization, producing the diatomic ion-molecule plus an electron, such as:

$$Xe_2^* + h\nu \rightarrow Xe_2^+ + e \tag{1}$$

Such excited state absorption limited the efficiency for the VUV rare gas excimers, even though they have very high formation and fluorescence efficiency.

Rare gas halide excimer lasers evolved from 'chemi-luminescence' chemical kinetics studies which were looking at the reactive quenching of rare gas metastable excited states in flowing after glow experiments. Broadband emission in reactions with halogen molecules was observed in these experiments, via reactions such as:

$$Xe^* + Cl_2 \rightarrow XeCl^* + Cl \tag{2}$$

At low pressure the emissions from molecules, such as XeCl*, are broad in the emissions bandwidth. Shortly after these initial observations, the laser community began examination of these species in high-pressure, electron-beam excited mixtures. The results differed remarkably from the low-pressure experiments and those for the earlier excimers such as Xe_2*. The spectrum was shifted well away from the VUV emission. Moreover, the emission was much sharper at high pressure, though still a continuum with a bandwidth of the order of 4 nm. A basis for understanding is sketched in the potential energy curves of **Fig. 4**. In this schematic we identify the rare gas atoms by Rg, excited rare gas atoms by Rg*, and halogen atoms by X. Ions play a very important role in the binding and reaction chemistry, leading to excited laser molecules.

The large shift in wavelength from the VUV of rare gas dimer excimers, $\sim$308 nm in XeCl* versus 172 nm in Xe_2*, is due to the fact that the binding in the excited state of the rare gas halide is considerably stronger. The sharp and intense continuum at high pressure ($> \sim$100 torr total pressure) is due to the fact that the 'sharp' transition terminates on a 'flat' portion of the lower-level potential energy curve. Indeed, in XeCl and XeF, the sharp band laser transition terminates on a slightly bound portion of the lower-level potential energy curve. Both the sharp emissions and broad bands are observed, due to the fact that some transitions from the excited levels terminate on a second, more repulsive potential energy curve.

An understanding of the spectroscopy and kinetic processes in rare gas plus halogen mixtures is based on the recognition that the excited state of a rare gas halide molecule is very similar to the ground state of the virtually isoelectronic alkali halide. The binding energy and mechanism of the excited state is very close to that of the ground state of the most similar alkali halide. Reactions of excited state rare gases are very similar to those of alkali atoms. For example, Xe* and Cs are remarkably similar, from

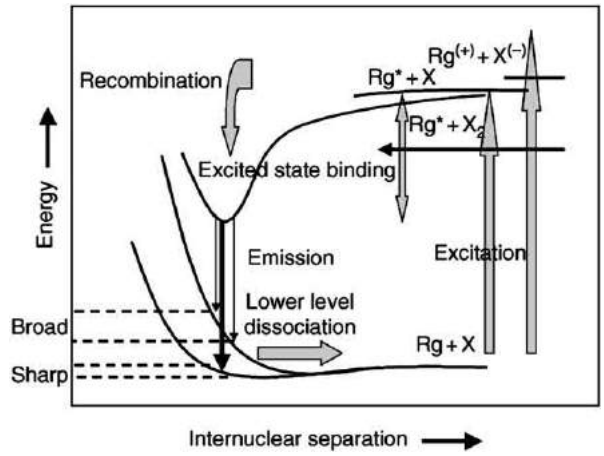

Fig. 4 The potential energy curves for rare gas halides are sketched here. Both relatively sharp continuum emission is observed along with broad bands corresponding to different lower levels. The excited ion pair states consist of 3 distinct levels, two that are very close in energy and one that is shifted up from the upper laser level (the B state) by the spin orbit splitting of the rare gas ion.

a chemistry and molecular physics point of view, as they have the same outer shell electron configuration and similar ionization potentials. The ionization potential of Xe* is similar to that of cesium (Cs). Cs, and all the other alkali atoms, react rapidly with halogen molecules. Xenon metastable excited atoms react rapidly with halogen molecules. The alkali atoms all form ionically bonded ground states with halogens. The rare gas halides are effectively strongly bonded ion pairs, $Xe^{(+)}X^{(-)}$, that are very strongly bonded relative to the excited states that yield them. The difference between, for example, XeCl* and CsCl, is that XeCl* radiates in a few nanoseconds while CsCl is a stable ground-state molecule. The binding energy for these ionic bonded molecules is much greater than the more covalent type of bonding that is found in the rare gas excimer excited states. Thus XeI*, which is only one electron different (in the I) than Xe_2*, emits at 254 nm instead of 172 nm.

The first observed and most obvious connection to alkali atoms is in formation chemistry. Mostly, rare gas excited states react with halogen molecules rapidly. The rare gas halide laser analog reaction sequence is shown below, where some form of electric discharge excitation kicks the rare gas atom up to an excited state, so that it can react like an alkali atom, **Eq. (3)**:

$$e_{fast} + Kr \rightarrow Kr^* + e_{slow} \tag{3}$$

$$Kr^* + F_2 \rightarrow KrF^*(v,J)_{high} + F \tag{4}$$

There are subtle and important cases where the neutral reaction, such as that shown in **Eqs. (2) or** (4), is not relevant as it is too slow compared to other processes. Note that the initial reaction does not yield the specific upper levels for the laser but states that are much higher up in the potential well of the excited rare gas halide excimer molecule. Further collisions with a buffer gas are needed to relax the states to the vibrational levels that are the upper levels for the laser transition. Such relaxation at high pressure can be fast, which leads the high-pressure spectrum to be much sharper than those first observed in flowing after-glow experiments.

The excited states of the rare gas halides are ion pairs bound together for their brief radiative lifetime. This opens up a totally different formation channel, indeed the one that often dominates the mechanism, ion–ion recombination. Long before excimer lasers were studied, it was well known that halogen molecules would react with 'cold' electrons (those with an effective temperature less than a few eV) in a process called attachment. The sequence leading to the upper laser level is shown below for the Kr/F_2 system, but the process is generically identical in other rare gas halide mixtures.

KrF* formation kinetics via ion channel:

$$\text{Ionization:}\quad e + Kr \rightarrow Kr^+ + 2e \tag{5}$$

$$\text{Attachment:}\quad e + F_2 \rightarrow F^{(-)} + F \tag{6}$$

$$\text{Ion} - \text{ion recombination:}\quad Kr^+ + F^{(-)} + M \rightarrow KrF * (v,J)_{hight} + M \tag{7}$$

$$\text{Relaxation:}\quad KrF^*(v,J)_{hight} + M \rightarrow KrF^*(v,J)_{low} + M \tag{8}$$

Finally, there is one other formation mechanism, displacement, that is unique to the rare gas halides. In this process, a heavier rare gas will displace a lighter rare gas ion in an excited state to form a lower-energy excited state:

$$ArF^* + Kr \rightarrow KrF^*(v,J)_{hight} + Ar \tag{9}$$

In general, the most important excimer lasers tend to have all of these channels contributing to excited state formation for optimized mixtures.

These kinetic formation sequences express some important points regarding rare gas halide lasers. The source of the halogen atoms for the upper laser level disappears via both attachment reactions with electrons and by the reaction with excited states. The halogen atoms eventually recombine to make molecules, but at a rate too slow for sufficient continuous wave (cw) laser gain. For excimer-based lamps, however, laser gain is not a criterion and very efficient excimer lamps can be made. Another point to note is that forming the inversion requires transient species (and in some cases halogen fuels) that absorb the laser light. Net gain and extraction efficiency are a trade-off between pumping rate, an increase in which often forms more absorbers and absorption itself. **Table 2** provides a listing of some of the key absorbers. Finally, the rare gas halide excited states can be rapidly quenched in a variety of processes. More often than not, the halogen molecule and the electrons in the discharge react with the excited states of the rare gas halide, removing the contributors to gain. We show, in **Table 3**, a sampling of some of the types of formation and quenching reactions with their kinetic rate constant parameters. Note that one category of reaction, three-body quenching, is unique to excimer lasers. Note also that the three-body recombination of ions has a rate coefficient that is effectively pressure

Table 2 Absorbers in rare gas halide lasers

Species	*XeF*	*XeCl*	*KrF*	*ArF*
Laser λ (nm)	351, 353	308	248	193
F_2 absorption σ (cm^2)	8×10^{-21}	NA	1.5×10^{-20}	
Cl_2 absorption σ (cm^2)	NA	1.7×10^{-19}	NA	NA
HCl absorption σ (cm^2)	NA	Nil	NA	NA
$F^{(-)}$ absorption σ (cm^2)	2×10^{-18}	NA	5×10^{-18}	5×10^{-18}
$Cl^{(-)}$ absorption σ (cm^2)	NA	2×10^{-17}	NA	NA
$Rg_2^{(+)}$ diatomic ion absorption σ (cm^2)[a]	Ne: $\sim 10^{-18}$ Ar: $\sim 3 \times 10^{-17}$ Xe: $\sim 3 \times 10^{-17}$	Ne: $\sim 10^{-17}$ Ar: $\sim 4 \times 10^{-17}$ Xe: $\sim 3 \times 10^{-18}$	Ne: $\sim 2.5 \times 10^{-17}$ Ar: $\sim 2 \times 10^{-18}$ Kr $< 10^{-18}$	Ne: $\sim 10^{-18}$ Ar: NIL
Other transient absorbers	Excited states of rare gas atoms and rare gas dimer molecules $\sigma \sim 10^{-19}$–10^{-17}	Excited states of rare gas atoms and rare gas dimer molecules $\sigma \sim 10^{-19}$–10^{-17}	Excited states of rare gas atoms and rare gas dimer molecules $\sigma \sim 10^{-19}$–10^{-17}	Excited states of rare gas atoms and rare gas dimer molecules $\sigma \sim 10^{-19}$–10^{-17}
Other	Lower laser level absorbs unless heated	Weak lower level absorption $\sigma \sim 4 \times 10^{-16}$	Windows and dust can be an issue	Windows and dust can be an issue

[a]Note that the triatomic rare gas halides of the form Rg_2X will exhibit similar absorption as the diatomic rare gas ions as the triatomic rare gas halides of this form are essentially ion pairs of an $Rg_2^{(+)}$ ion and the halide ion.

Table 3 Typical formation and quenching reactions and rate coefficients

Formation	
Rare gas plus halogen molecule	$Kr^* + F_2 \rightarrow KrF^*(v,J) + F$ $k \sim 7 \times 10^{-10}$ cm^3/sec; Branching ratio to $KrF^* \sim 1$
	$Ar^* + F_2 \rightarrow ArF^*(v,J) + F$ $k \sim 6 \times 10^{-10}$ cm^3/sec; Branching ratio to $ArF^* \sim 60\%$
	$Xe^* + F_2 \rightarrow XeF^*(v,J) + F$ $k \sim 7 \times 10^{-10}$ cm^3/sec; Branching ratio to $XeF^* \sim 1$
	$Xe^* + HCl(v=0) \rightarrow Xe + H + Cl$ $k \sim 6 \times 10^{-10}$ cm^3/sec; Branching ratio to $XeCl^* \sim 0$
	$Xe^* + HCl(v=1) \rightarrow XeCl^* + H$ $k \sim 2 \times 10^{-10}$ cm^3/sec; Branching ratio to $XeCl^* \sim 1$
Ion–ion recombination	$Ar^{(+)} + F^{(-)} + Ar \rightarrow ArF^*(v,J)$ $k \sim 1 \times 10^{-25}$ cm^6/sec at 1 atm and below $k \sim 7.5 \times 10^{-26}$ cm^6/sec at 2 atm; Note effective 3-body rate constant decreases further as pressure goes over ~ 2 atm
	$Kr^{(+)} + F^{(-)} + Kr \rightarrow KrF^*(v,J)$ $k \sim 7 \times 10^{-26}$ cm^6/sec at 1 atm and below; rolls over at higher pressure
Displacement	$ArF^*(v,J) + Kr \rightarrow KrF^*(v,J) + Ar$ $k \sim 2 \times 10^{-10}$ cm^3/sec; Branching ratio to $KrF^* \sim 1$
Quenching	
Halogen molecule	$XeCl^* + HCl \rightarrow Xe + Cl + HCl$ $k \sim 5.6 \times 10^{-10}$ cm^3/sec;
	$KrF^* + F_2 \rightarrow Kr + F + F$ $k \sim 6 \times 10^{-10}$ cm^3/sec;
	All halogen molecule quenching have very high reaction rate constants
3-body inert gas	$ArF^* + Ar + Ar \rightarrow Ar_2F^* + Ar$ $k \sim 4 \times 10^{-31}$ cm^6/sec
	$KrF^* + Kr + Ar \rightarrow Kr_2F^* + Ar$ $k \sim 6 \times 10^{-31}$ cm^6/sec
	$XeCl^* + Xe + Ne \rightarrow Xe_2Cl^* + Ne$ $k \sim 1 \times 10^{-33}$ cm^6/sec; 4×10^{-31} cm^6/sec with Xe as 3rd body
	These reactions yield a triatomic excimer at lower energy that can not be recycled back to the desired laser states
Electrons	$XeCl^* + e \rightarrow Xe + Cl + e$ $k \sim 2 \times 10^{-7}$ cm^3/sec
	In 2-body quenching by electrons, the charge of the electron interacts with the dipole of the rare gas halide ion pair at long range; above value is typical of others
2-body inert gas	$KrF^* + Ar \rightarrow Ar + Kr + F$ $k \sim 2 \times 10^{-12}$ cm^3/sec
	$XeCl^* + Ne \rightarrow Ne + Xe + Cl$ $k \sim 3 \times 10^{-13}$ cm^3/sec
	2-body quenching by rare gases is typically slow and less important than the reactions noted above

dependent. At pressures above ~3 atm, the diffusion of ions toward each other is slowed and the effective recombination rate constant decreases.

Though the rare gas halide excited states can be made with near-unity quantum yield from the primary excitation, the intrinsic laser efficiency is not this high. The quantum efficiency is only ~25% (recall rare gas halides all emit at much longer wavelengths than rare gas excimers); there is inefficient extraction due to losses in the gas and windows. In practice there are also losses in coupling the power into the laser, and wasted excitation during the finite start-up time. The effect of losses is shown in **Table 4** and **Fig. 5** where the pathways are charted and losses identified. In KrF, it takes ~20 eV to make a rare gas ion in an electron beam excited mixture, somewhat less in an electric discharge. The photon release is ~5 eV; the quantum efficiency is only 25%. Early laser efficiency measurements (using electron beam excitation) showed laser efficiency in the best cases slightly in excess of 10%, relative to the deposited energy. The discrepancy relative to the quantum efficiency is due to losses in the medium. For the sample estimate in **Table 4** and **Fig. 5** we show approximate density of key species in the absorption chain, estimated on a steady-state approximation of the losses in a KrF laser for an excitation rate of order 1 MW/cc, typical of fast discharge lasers. The net small signal gain is of the order of 5%/cm. The key absorbers are the parent halogen molecule, F_2, the fluoride ion $F^{(-)}$, and the rare gas excited state atoms. A detailed pulsed kinetics code will provide slightly different results due to transient kinetic effects and the intimate coupling of laser extraction to quenching and absorber dynamics. The ratio of gain to loss is usually in the range of 7 to 20, depending on the actual mixture used and the rare gas halide wavelength. The corresponding extraction efficiency is then limited to the range of ~50%. The small signal gain, g_o, is related to the power deposition rate per unit volume, $P_{\text{deposited}}$, by **Eq. (10)**.

Table 4 Typical ground and excited state densities shown for KrF at $P_{in} \sim 1$ MW/cm^3

Species	*Amount particles/cm^3*	*Absorption σ (cm)^2*	*Loss %/cm*
F_2	$\sim 1.2 \times 10^{17}$	1.5×10^{-20}	0.18
Kr	$\sim 3 \times 10^{18}$	–	–
Buffer	4×10^{19}	–	–
$F^{(-)}$	$\sim 10^{14}$	5×10^{-18}	0.05
Rg*, Rg**	1×10^{15}	$\langle 10^{-18} \rangle$ depends on ratio of Rg**/Rg*	0.1
KrF (before dissociating)	$\sim 10^{12}$	2×10^{-16}	0.02
Rg_2F^*	8×10^{14}, Note this species density decreases as flux grows	1×10^{-18}	0.08
KrF* (B)	2.5×10^{14}	2×10^{-16}	5 (gain)

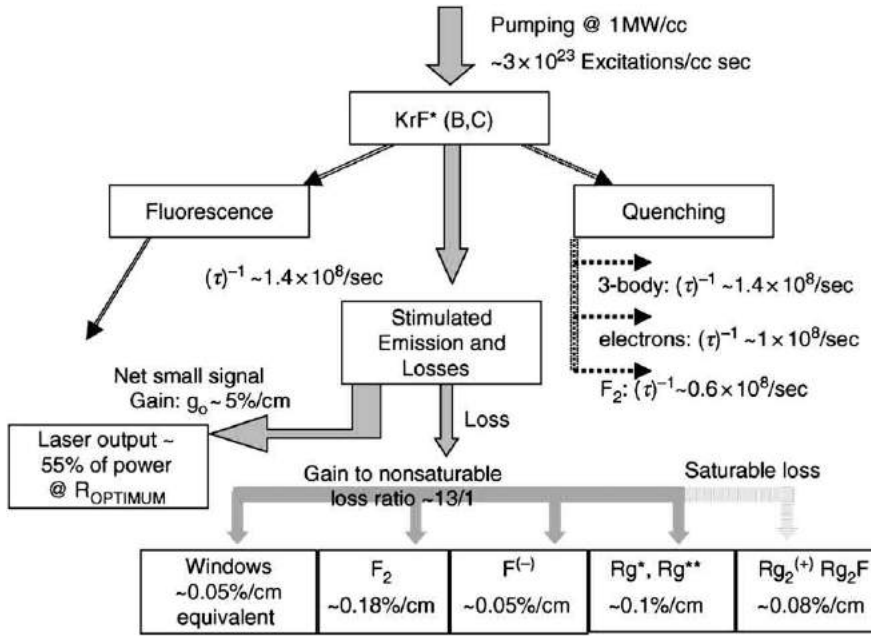

Fig. 5 The major decay paths and absorbers that impact the extraction of excited states of rare gas halide excimers. Formation is via the ion–ion recombination, excited atom reaction with halogens and displacement reactions, not shown. We note for a specific mixture of KrF the approximate values of the absorption by the transient and other species along with the decay rates for the major quenching processes. Extraction efficiency greater than 50% is quite rare.

$$P_{\text{deposited}} = g_0 E^* \left(\tau_{\text{upper}} \eta_{\text{pumping}} \eta_{\text{branching}} f \sigma\right)^{-1} \tag{10}$$

E^* is the energy per excitation, $\sim$20 eV or 3.2×10^{-18} J/excitation, f is the fraction of excited molecules in the upper laser level, $\sim$40% due to KrF* in higher vibrational states and the slowly radiating 'C' state. The laser cross-section is σ, and the η terms are efficiencies for getting from the primary excitation down to the upper laser level. The upper state lifetime, τ_{upper}, is the sum of the inverses of radiative and quenching lifetimes. When the laser cavity flux is sufficiently high, stimulated emission decreases the flow of power into the quenching processes.

The example shown here gives a ratio of net gain to the nonsaturating component of loss of 13/1. Of the excitations that become KrF*, only $\sim$55% or so will be extracted in the best of cases. For other less ideal rare gas halide lasers, the gain to loss ratio, the extraction efficiency, and the intrinsic efficiency are all lower. Practical discharge lasers have lower practical wall plug efficiency due to a variety of factors, discussed below.

Discharge Technology

Practical discharge lasers use fast-pulse discharge excitation. In this scheme, an outgrowth of the CO_2 TEA laser, the gas is subjected to a rapid high-voltage pulse from a low-inductance pulse forming line or network, PFL or PFN. The rapid pulse serves to break down the gas, rendering it conductive with an electron density of $> \sim 10^{14}$ electrons/cm^3. As the gas becomes conductive, the voltage drops and power is coupled into the medium for a short duration, typically less than 50 nsec. The discharge laser also requires some form of 'pre-ionization', usually provided by sparks or a corona-style discharge on the side of the discharge electrodes. The overall discharge process is intrinsically unstable, and pump pulse duration needs to be limited to avoid the discharge becoming an arc.

Moreover, problems with impedance matching of the PFN or PFL with the conductive gas lead to both inefficiency and extra energy that can go into post-discharge arcs, which both cause component failure and produce metal dust which can give rise to window losses. A typical discharge laser has an aperture of a few cm^2 and a gain length of order 80 cm and produces outputs in the range of 100 mJ to 1 J. The optical pulse duration tends to be $\sim$20 nsec, though the electrical pulse is longer due to the finite time needed for the gain to build up and the stimulated emission process to turn spontaneously emitted photons into photons in the laser cavity.

The pre-ionizer provides an initial electron density on the order of 10^8 electrons/cm^3. With spark pre-ionization or corona pre-ionization, it is difficult to make a uniform discharge over an aperture greater than a few cm^2. For larger apertures, one uses X-rays for pre-ionization. X-ray pre-ionized discharge excimer lasers have been scaled to energy well over 10 J per pulse and can yield pulse durations over 100 nsec. The complexity and expense of the X-ray approach, along with the lack of market for lasers in this size range, resulted in the X-ray approach remaining in the laboratory.

Aside from the kinetic issues of the finite time needed for excitation to channel into the excimer upper levels and the time needed for the laser to 'start up' from fluorescent photons, these lasers have reduced efficiency due to poor matching of the pump source to the discharge. Ideally, one would like to have a discharge such that when it is conductive the voltage needed to sustain the discharge is less than half of that needed to break the gas down. Regrettably for the rare gas halides (and in marked distinction to the CO_2 TEA laser) the voltage that the discharge sustains in a quasi-stable mode is $\sim$20% of the breakdown voltage, perhaps even less, depending on the specific gases and electrode shapes.

Novel circuits, which separate the breakdown phase from the fully conductive phase, can readily double the efficiency of an excimer discharge laser, though this approach is not necessary in many applications. The extra energy that is not dissipated in the useful laser discharge often results in arcing or 'hot' discharge areas after the main laser pulse, leading to electrode sputtering, dust that finds its way to windows, and the need to service the laser after $\sim 10^8$ pulses. When one considers the efficiency of the power conditioning system, along with the finite kinetics of the laser start-up process in short pulses, and the mismatch of the laser impedance to the PFN/PFL impedance, lasers that may have 10% intrinsic efficiency in the medium in e-beam excitation may only have 2% to 3% wall plug efficiency in a practical discharge laser.

A key aspect of excimer technology is the need for fast electrical circuits to drive the power into the discharge. This puts a significant strain on the switch that is used to start the process. While spark gaps, or multichannel spark gaps, can handle the required current and current rate of rise, they do not present a practical alternative for high-pulse rate operation. Thyratrons and other switching systems, such as solid state switches, are more desirable than a spark gap, but do not offer the desired needed current rate of rise. To provide the pumping rate of order 1 MW/cm^3 needed for enough gain to turn the laser on before the discharge becomes unstable, the current needs to rise to a value of the order of 20 kA in a time of $\sim$20 nsec; the current rate of rise is $\sim 10^{12}$ A/sec. This rate of rise will limit the lifetime of the switch that drives the power into the laser gas. As a response to this need, magnetic switching and magnetic compression circuits were developed so that the lifetime of the switch can be radically increased. In the simplest cases, one stage of magnetic compression is used to make the current rate of rise to be within the range of a conventional thyratron switch as used in radar systems. Improved thyratrons and a simple magnetic assist also work. For truly long operating lifetimes, multiple stages of magnetic switching and compression are used, and with a transformer one can use a solid state diode as the start switch for the discharge circuit. A schematic of an excimer pulsed power circuit, that uses a magnetic assist with a thyratron and one stage of magnetic compression, is shown in Fig. 6. The magnetic switching approach can also be used to provide two separate circuits to gain high efficiency by having one circuit to break down the gas (the spiker) and a second,

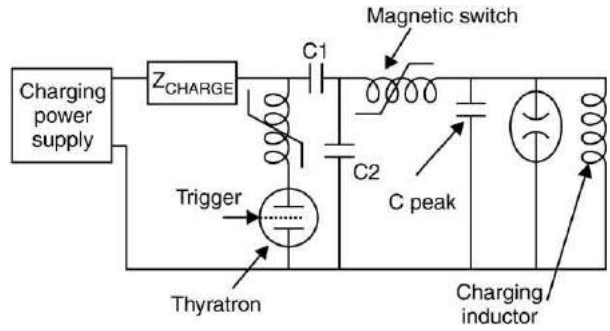

Fig. 6 One of the variants on magnetic switching circuits that have been used to enhance the lifetime of the primary start switch. A magnetic switch in the form of a simple 1 turn inductor with a saturable core magnetic material such as a ferrite allows the thyratron to turn on before major current flows. Other magnetic switches may be used in addition to peak up the voltage rate of rise on the laser head or to provide significant compression. For example one may have the major current flow in the thyratron taking place on the 500 nsec time scale while the voltage pulse on the discharge is in the 50 nsec duration. A small current leaks through and is accommodated by the charging inductor.

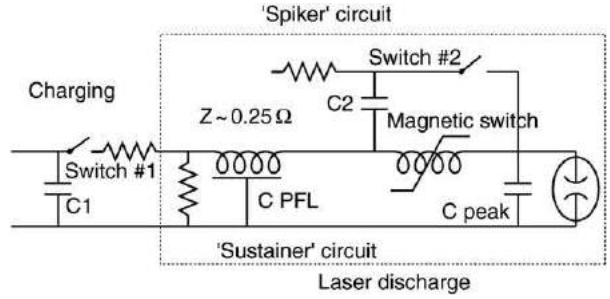

Fig. 7 For the ultimate in efficiency the circuit can be arranged to provide a leading edge spike that breaks the gas down using a high impedance 'spiker' and a separate circuit that is matched to the discharge impedance when it is fully conductive. High coupling efficiency reduces waste energy that can erode circuit and laser head components via late pulse arcs. The figure shows a magnetic switch as the isolation but the scheme can be implemented with a second laser head as the isolation switch, a rail gap (not appropriate for long life) or a diode at low voltage.

low-impedance 'sustainer' circuit, to provide the main power flow. In this approach, the laser itself can act as part of the circuit by holding off the sustaining voltage for a μ-sec or so during the charge cycle and before the spiker breaks down the gas. A schematic of such a circuit is shown in Fig. 7. Using this type of circuit, the efficiency of a XeCl laser can be over 4% relative to the wall plug. A wide variety of technological twists on this approach have been researched.

Excimer lasers present some clear differences in resonator design compared to other lasers. The apertures are relatively large, so one does not build a TEM_{oo} style cavity for good beam quality and expect to extract any major portion of the energy available. The typical output is divergent and multimode. When excimer lasers are excited by a short pulse discharge, $\sim$30 nsec, the resulting gain duration is also short, $\sim$20 nsec, limiting the number of passes a photon in a resonator can have in the gain. The gain duration of $\sim$20 nsec only provides $\sim$3.5 round trips of a photon in the cavity during the laser lifetime, resulting in the typical high-order multimode beam. Even with an unstable resonator, the gain duration is not long enough to collapse the beam into the lowest-order diffraction-limited output. If line narrowing is needed, and reasonable efficiency is required, an oscillator amplifier configuration is often used with the oscillator having an appropriate tuning resonator in the cavity. If both narrow spectral linewidth and excellent beam quality are needed one uses a seeded oscillator approach where the narrowband oscillator is used to seed an unstable resonator on the main amplifier. By injecting the seed into the unstable resonator cavity, a few nsec before the gain is turned on, the output of the seeded resonator can be locked in wavelength, bandwidth, and provide good beam quality. Long pulse excimer lasers, such as e-beam excited lasers or long pulse X-ray pre-ionized discharge devices can use conventional unstable resonator technology with good results.

Excimer lasers are gas lasers that run at both high pressure and high instantaneous power deposition rates. During a discharge pulse, much of the halogen bearing 'fuel' is burned out by the attachment and reactive processes. The large energy deposition per pulse means that pressure waves are generated. All of these combine with the characteristic device dimensions to require gas flow to recharge the laser mixture between pulses. For low pulse rates, less than $\sim$200 Hz, the flow system need not be very sophisticated. One simply needs to flush the gas from the discharge region with a flow having a velocity of the order of 5 m/sec for a typical discharge device. The flow is transverse to the optical and discharge directions. At high pulse rates, the laser designer needs to consider flow in much more detail to minimize the power required to push the gas through the laser head. Circulating pressure waves need to be damped out so that the density in the discharge region is controlled at the time of the next pulse. Devices called acoustic dampers are placed in the sides of the flow loop to remove pressure pulses. A final subtlety occurs in providing gas flow

over the windows. The discharge, post pulse arcs, and reactions of the halogen source with the metal walls and impurities creates dust. When the dust coats the windows, the losses in the cavity increase, lowering efficiency. By providing a simple gas flow over the window, the dust problem can be ameliorated. For truly long life, high-reliability lasers, one needs to take special care in the selection of materials for electrodes and insulators and avoid contamination, by assembling in clean environments.

Further Reading

Ballanti, S., Di Lazzaro, P., Flora, F., *et al*., 1998. Ianus the 3-electrode laser. Applied Physics B 66, 401–406.
Brau, C.A., 1978. Rare gas halogen lasers. In: Rhodes, C.K. (Ed.), Excimer Lasers, Topics of Applied Physics, vol. 30. Berlin: Springer Verlag.
Brau, C.A., Ewing, J.J., 1975. Emission spectra of XeBr, XeCl, XeF, and KrF. Journal of Chemical Physics 63, 4640–4647.
Ewing, J.J., 2000. Excimer laser technology development. IEEE Journal of Selected Topics in Quantum Electronics 6 (6), 1061–1071.
Levatter, J., Lin, S.C., 1980. Necessary conditions for the homogeneous formation of pulsed avalanche discharges at high pressure. Journal of Applied Physics 51, 210–222.
Long, W.H., Plummer, M.J., Stappaerts, E., 1983. Efficient discharge pumping of an XeCl laser using a high voltage prepulse. Applied Physics Letters 43, 735–737.
Rhodes, C.K. (Ed.), 1979. Excimer Lasers, Topics of Applied Physics, vol. 30. Berlin: Springer-Verlag.
Rokni, M., Mangano, J., Jacob, J., Hsia, J., 1978. Rare gas fluoride lasers. IEEE Journal of Quantum Electronics QE-14, 464–481.
Smilanskši, I., Byron, S., Burkes, T., 1982. Electrical excitation of an XeCl laser using magnetic pulse compression. Applied Physics Letters 40, 547–548.
Taylor, R.S., Leopold, K.E., 1994. Magnetic-spiker excitation of gas discharge lasers. Applied Physics B, Lasers and Optics 59, 479–509.

Metal Vapor Lasers

DW Coutts, University of Oxford, Oxford, UK

Nomenclature

k	Boltzmann constant [eV K^{-1}]
M^{+*}	Excited metal ion
N^*	Excited noble gas atom
?E	Energy difference [eV]
e^-	Electron
Gas flow	[mbar l min^{-1}]
R_1, R_2	Mirror curvatures [m]
M	Metal atom
N	Noble gas atom
N^+	Noble gas ion
Pulse duration	[ns]
Pulse repetition frequency	[kHz]
Q	Quality factor
T	Temperature [K]
D	Tube diameter [mm]
L	Tube length [m]
Wavelength	[nm], [μm]
M	Unstable resonator magnification

Introduction

Metal vapor lasers form a class of laser in which the active medium is a neutral or ionized metal vapor usually excited by an electric discharge. These lasers fall into two main subclasses, namely cyclic pulsed metal vapor lasers and continuous-wave metal ion lasers. Both types will be considered in this article, including basic design and construction, power supplies, operating characteristics (including principal wavelengths), and brief reference to their applications.

Self-Terminating Resonance-Metastable Pulsed Metal Vapor Lasers

The active medium in a self-terminating pulsed metal vapor laser consists of metal atoms or ions in the vapor phase usually as a minority species in an inert buffer gas such as neon or helium. Laser action occurs between a resonance upper laser level and a metastable lower laser level (Fig. 1). During a fast pulsed electric discharge (typically with a pulse duration of order 100 ns) the upper laser level is preferentially excited by electron impact excitation because it is strongly optically connected to the ground state (resonance transition) and hence has a large excitation cross-section. For a sufficiently large metal atom (or ion) density, the resonance radiation becomes optically trapped, thus greatly extending the lifetime of the upper laser level such that decay from the upper laser level is channelled through the emission of laser radiation to the metastable lower laser level. Lasing terminates when the electron temperature falls to a point such that preferential pumping to the upper laser level is no longer sustained, and the build-up of population in the metastable lower laser level destroys the population inversion. Therefore after each excitation pulse, the resulting excited species in the plasma (in particular the metastable lower laser levels which are quenched by collisions with cold electrons) must be allowed sufficient time to relax and the plasma must be allowed to partially recombine before applying the next excitation pulse. The relaxation times for self-terminating metal vapor lasers correspond to operating pulse repetition frequencies from 2 kHz to 200 kHz.

Many metal vapors can be made to lase in the resonance-metastable scheme and are listed together with their principal wavelengths and output powers in Table 1. The most important self-terminating pulsed metal vapor laser is the copper vapor laser and its variants which will be discussed in detail in the following sections. Of the other pulsed metal vapor lasers listed in Table 1,

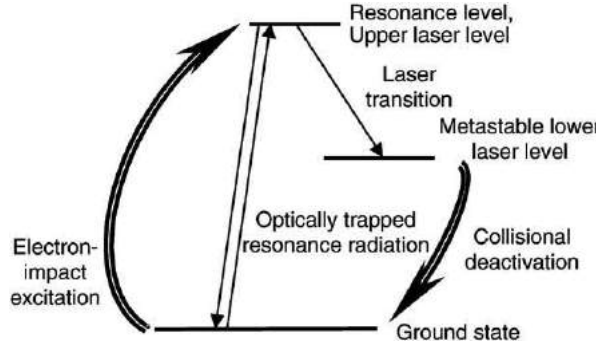

Fig. 1 Resonance-metastable energy levels for self-terminating metal vapor lasers.

 doi:10.1016/B978-0-12-809283-5.00847-8

Table 1 Principal self-terminating resonance-metastable metal vapor lasers

Metal	*Principal wavelengths (nm)*	*Powers*		*Total efficiencies*	*Pulse repetition frequency (kHz)*	*Technological development*
		Typical (W)	*Maximum (W)*			
Cu	510.55	2–70	2500 total	1%	4–40	Highly developed and commercially available
	578.2	1–50				
Au	627.8	1–8	20	0.13%	2–40	Commercially available
	312	0.1–0.2	1.2		1–8	
Ba	15002550	2–101	121.5	0.5%	5–15	Have been produced commercially, but largely experimental
	1130	0.5	1.0			
Pb	722.9		4.4	0.15%	10–30	Experimental
Mn	534.1 (>50%)		12 total	0.32%	~10	Experimental
	1290					

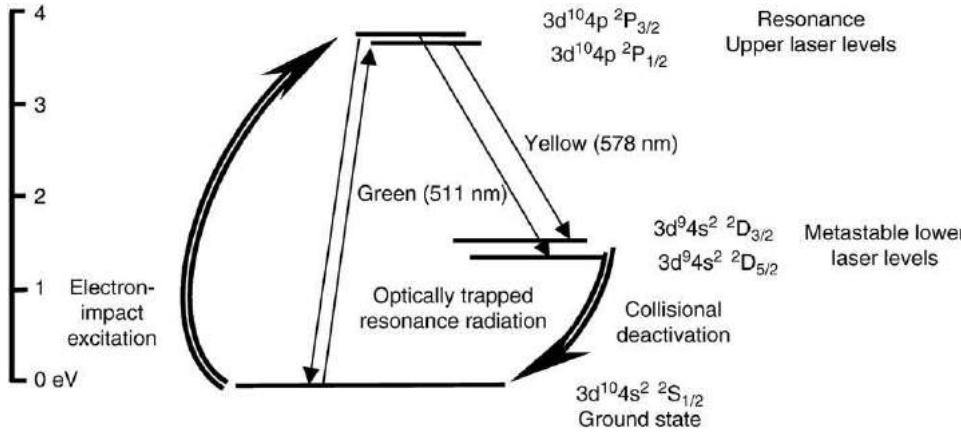

Fig. 2 Partial energy level scheme for the copper vapor laser.

only the gold vapor laser (principal wavelengths 627.8 nm and 312.3 nm), and the barium vapor laser which operates in the infrared (principal wavelengths 1.5 μm and 2.55 μm) have had any commercial success. All the self-terminating pulsed metal vapor lasers have essentially the same basic design and operating characteristics as exemplified by the copper vapor laser.

Copper Vapor Lasers

Copper vapor lasers (CVLs) are by far the most widespread of all the pulsed metal vapor lasers. Fig. 2 shows the energy level scheme for copper. Lasing occurs simultaneously from the $^2P_{3/2}$ level to the $^2D_{5/2}$ level (510.55 nm) and from the $^2P_{1/2}$ level to the $^2D_{3/2}$ level (578.2 nm). Commercial devices are available with combined outputs of over 100 W at 510.55 nm and 578.2 nm (typically with a green-to-yellow power ratio of 2:1).

A typical copper vapor laser tube is shown in Fig. 3. High-purity copper pieces are placed at intervals along an alumina ceramic tube which typically has dimensions of 1–4 cm diameter and 1–2 m long. The alumina tube is surrounded by a solid fibrous alumina thermal insulator, and a glass or quartz vacuum envelope. Cylindrical electrodes made of copper or tantalum are located at each end of the plasma tube to provide a longitudinal discharge arrangement. The cylindrical electrodes and silica laser end windows are supported by water-cooled end pieces. The laser windows are usually tilted by a few degrees to prevent back reflections into the active medium. The laser head is contained within a water cooled metal tube to provide a coaxial current return for minimum laser head inductance. Typically a slow flow ($\sim$5 mbar l min^{-1}) of neon at a pressure of 20–80 mbar is used as the buffer gas with an approximately 1% H_2 additive to improve the afterglow plasma relaxation. The buffer gas provides a medium to operate the discharge when the laser is cold and slows diffusion of copper vapor out of the ends of the hot plasma tube. Typical copper fill times are of order 200–2000 hours (for 20–200 g copper load). Sealed-off units with lifetimes of order 1000 hours have been in production in Russia for many years.

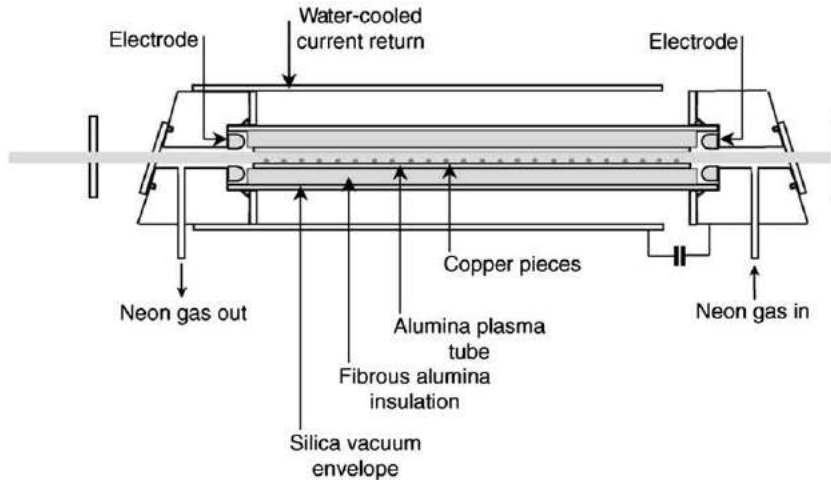

Fig. 3 Copper vapor laser tube construction.

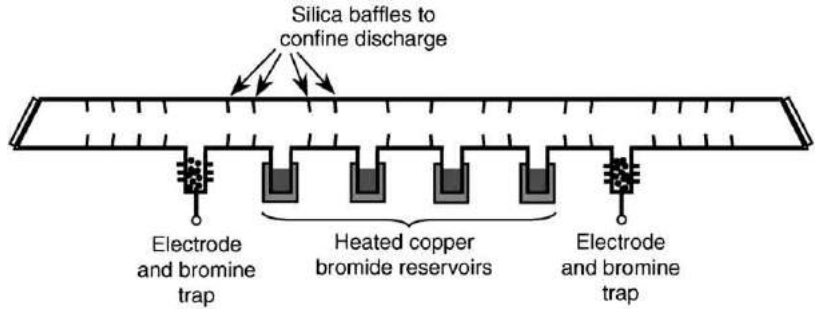

Fig. 4 Copper bromide laser tube construction.

During operation waste heat from the repetitively pulsed discharge heats the alumina tube up to approximately 1500°C at which point the vapor pressure of copper is of approximately 0.5 mbar which corresponds to the approximate density required for maximum laser output power. Typical warm-up times are therefore relatively long at around one hour to full power.

One way to circumvent the requirement for high temperatures required to produce sufficient copper density by evaporation of elemental copper (and hence also reduce warm-up times) is to use a copper salt with a low boiling point located in one or more side-arms of the laser tube (Fig. 4). Usually copper halides are used as the salt, with the copper bromide laser being the most successful. For the CuBr laser, a temperature of just 600°C is sufficient to produce the required Cu density by dissociation of CuBr vapor in the discharge. With the inclusion of 1–2% H_2 in the neon buffer gas, HBr is also formed in the CuBr laser, which has the additional benefit of improving recombination in the afterglow via dissociative attachment of free electrons: $HBr + e^- \rightarrow H + Br^-$, followed by ion neutralization: $Br^- + Cu^+ \rightarrow Br + Cu^*$. As a result of the lower operating temperature and kinetic advantages of HBr, CuBr lasers are typically twice as efficient (2–3%) as their elemental counterparts. Sealed-off CuBr systems with powers of order 10–20 W are commercially produced.

An alternative technique for reducing the operating temperature of elemental CVLs is to flow a buffer gas mixture consisting of ~5% HBr in neon at approximately 50 mbar l min^{-1} and allow this to react with solid copper metal placed within the plasma tube at about 600 °C to produce CuBr vapor *in situ*. The so-called Cu HyBrID (hydrogen bromide in discharge) laser has the same advantages as the CuBr laser (e.g., up to 3% efficiency) but at the cost of requiring flowing highly toxic HBr in the buffer gas, a requirement which has so far prevented commercialization of Cu HyBrID technology.

The kinetic advantages of the hydrogen halide in the CuBr laser discharge can also be applied to a conventional elemental CVL through the addition of small partial pressure of HCl to the buffer gas in addition to the 1–2% H_2 additive. HCl is preferred to HBr as it is less likely to dissociate (the dissociation energy of HCl at 0.043 eV is less than HBr at 0.722 eV). Such kinetic enhancement leads to a doubling in average output power, a dramatic increase in beam quality through improved gain characteristics, and shifts the optimum pulse repetition frequency for kinetically enhanced CVLs (KE-CVLs) from 4–10 kHz up to 30–40 kHz.

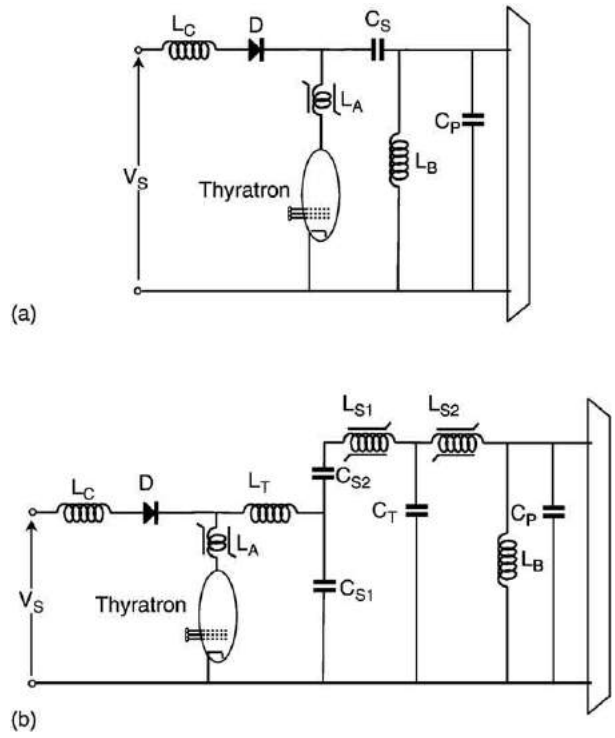

Fig. 5 Copper vapor laser excitation circuits. (a) Charge transfer circuit; (b) LC inversion circuit with magnetic pulse compression.

Preferential pumping of the upper laser levels requires an electron temperature in excess of the 2 eV range, hence high voltage (10–30 kV), high current (hundreds of A), short (75–150 ns) excitation pulses are required for efficient operation of copper vapor lasers. To generate such excitation pulses CVLs are typically operated with a power supply incorporating a high-voltage thyratron switch. In the most basic configuration, the charge-transfer circuit shown in **Fig. 5(a)**, a dc high-voltage power supply resonantly charges a storage capacitor (C_S, typically a few nF) through a charging inductor L_C, a high-voltage diode and bypass inductor L_B, up to twice the supply voltage V_S in a time of order 100 μs. When the thyratron is triggered, the storage capacitor discharges through the thyratron and the laser head on a time-scale of 100 ns. Note that during the fast discharge phase, the bypass inductor in parallel with the laser head can be considered to be an open circuit. A peaking capacitor C_P ($\sim 0.5\ C_S$) is provided to increase the rate of rise of the voltage pulse across the laser head. Given the relatively high cost of thyratrons, more advanced circuits are now often used to extend the service lifetime of the thyratron to several thousand hours. In the more advanced circuit (**Fig. 5(b)**) an LC inversion scheme is used in combination with magnetic pulse compression techniques and operates as follows. Storage capacitors C_S are resonantly charged in parallel to twice the dc supply voltage V_S as before. When the thyratron is switched, the charge on C_{S1} inverts through the thyratron and transfer inductor L_T. This LC inversion drags the voltage on the top of C_{S2} down to $-4V_S$. When the voltage across the first saturable inductor L_{S1} reaches a maximum ($-4V_S$) L_{S1} saturates, and allows current to flow from the storage capacitors (now charged in series) to the transfer capacitor ($C_T = 0.5C_{S2}$) thereby transferring the charge from C_{S1} and C_{S2} to C_T in a time much less than the initial LC inversion time. At the moment when the charge on transfer capacitor C_T reaches a maximum (also $-4V_S$), L_{S2} saturates, and the transfer capacitor is discharged through the laser tube, again with a peaking capacitor to increase the voltage rise time. By using magnetic pulse compression the thyratron switched voltage can be reduced by 4 and the peak current similarly reduced (at the expense of increased current pulse duration) thereby greatly extending the thyratron lifetime. Note that in both circuits a 'magnetic assist' L_A saturable inductor is provided in series with the thyratron to delay the current pulse through the thyratron until after the thyratron has reached high conductivity thereby reducing power deposition in the thyratron.

Copper vapor lasers produce high average powers (2–100 W available commercially, with laboratory devices producing average powers of over 750 W) and have wall-plug efficiencies of approximately 1%. Copper vapor lasers also make excellent amplifiers due to their high gains, and the amplifiers can be chained together to produce average powers of several kW. Typical pulse repetition frequencies range from 4 to 20 kHz, with a maximum reported pulse repetition frequency of 250 kHz. An approximate scaling law states that for an elemental device with tube diameter D (mm) and length L (m) the average output power in watts will be of order $D \times L$. For example, a typical 25 W copper vapor laser will have a 25 mm diameter by 1 m long laser tube, and operate at 10 kHz corresponding to 2.5 mJ pulse energy and 50 kW peak power (50 ns pulse duration). Copper vapor lasers have very high single-pass gains (greater than 1000 for a 1 m long tube), large gain volumes and short gain durations (20–80 ns, sufficient for the intracavity laser light to make only a few round trips within the optical resonator). Maximum output power is therefore usually obtained in a highly 'multimode' (spatially incoherent) beam by using either a fully stable or a plane–plane resonator with a low

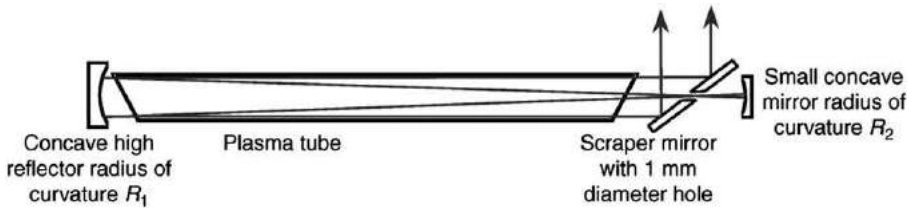

Fig. 6 Unstable resonator configuration often used to obtain high beam quality from copper vapor lasers.

reflectivity output coupler (usually Fresnel reflection from an uncoated optic is sufficient). To obtain higher beam quality a high magnification unstable resonator is required (Fig. 6). Fortunately, copper vapor lasers have sufficient gain to operate efficiently with unstable resonators with magnifications $(M = R_1/R_2)$ up to 100 and beyond. Resonators with such high magnifications impose very tight geometric constraints on propagation of radiation on repeated round-trips within the resonator, such that after two round-trips the divergence is typically diffraction-limited. Approximately half the stable-resonator output power can therefore be obtained with near diffraction-limited beam quality by using an unstable resonator. Often, a small-scale oscillator is used in conjunction with a single power amplifier to produce high output power with diffraction-limited beam quality. Hyperfine splitting combined with Doppler broadening lead to an inhomogeneous linewidth for the laser transitions of order 8–10 GHz corresponding to a coherence length of order 3 cm.

The high beam quality and moderate peak power of CVLs allows efficient nonlinear frequency conversion to the UV by second harmonic generation (510.55 nm → 255.3 nm, 578.2 nm → 289.1 nm) and sum frequency generation (510.55 nm + 578.2 nm → 271.3 nm) using β-barium borate β-BaB_2O_4 (BBO) as the nonlinear medium. Typically average powers in excess of 1 W can be obtained at any of the three wavelengths from a nominally 20 W CVL. Powers up to 15 W have been obtained at 255 nm from high-power CVL master-oscillator power-amplifier systems using cesium lithium borate; CsLi B_6O_{10} (CLBO) as the nonlinear crystal.

Key applications of CVLs include pumping of dye lasers (principally for laser isotope separation) and pumping Ti:sapphire lasers. Medically, the CVL yellow output is particularly useful for treatment of skin lesions such as port wine stain birth marks. CVLs are also excellent sources of short-pulse stroboscopic illumination for high-speed imaging of fast objects and fluid flows. The high beam quality, visible wavelength and high pulse repetition rate make CVLs very suited to precision laser micromachining of metals, ceramics and other hard materials. More recently, the second harmonic at 255 nm has proved to be an excellent source for writing Bragg gratings in optical fibers.

Afterglow Recombination Metal Vapor Lasers

Recombination of an ionized plasma in the afterglow of a discharge pulse provides a mechanism for achieving a population inversion and hence laser output. The two main afterglow recombination metal vapor lasers are the strontium ion vapor laser (430.5 nm and 416.2 nm) and calcium ion vapor laser (373.7 nm and 370.6 nm) whose output in the violet and UV spectral regions extends the spectral coverage of pulsed metal vapor lasers to shorter wavelengths.

A population inversion is produced by recombination pumping where doubly ionized Sr (or Ca) recombines to form singly ionized Sr (or Ca) in an excited state: $Sr^{++} + e^- + e^- \rightarrow Sr^{+*} + e^-$. Note that recombination rates for a doubly ionized species are much faster than for singly ionized species hence recombination lasers are usually metal ion lasers. Recombination (pumping) rates are also greatest in a cool dense plasma, hence helium is usually used as the buffer gas as it is a light atom hence promotes rapid collisional cooling of the electrons. Helium also has a much higher ionization potential than the alkaline-earth metals (including Sr and Ca) which ensures preferential ionization of the metal species (up to 90% may be doubly ionized).

Recombination lasers can operate where the energy level structure of an ion species can be considered to consist of two (or more) groups of closely spaced levels. In the afterglow of a pulsed discharge electron collisional mixing within each group of levels will maintain each group in thermodynamic equilibrium yielding a Boltzmann distribution of population within each group. If the difference in energy between the two groups of levels $(5\text{eV} \gg \text{kT})$ is sufficiently large then thermal equilibrium between the groups cannot be maintained via collisional processes. Given that recombination yields excited singly ionized species (usually with a flux from higher to lower ion levels), it is possible to achieve a population inversion between the lowest level of the upper group and the higher levels within the lower group. This is the mechanism for inversion in the Sr^+ (and analogous Ca^+) ion laser as indicated in the partial energy level scheme for Sr^+ shown in Fig. 7.

Strontium and calcium ion recombination lasers have similar construction to copper vapor lasers described above. The lower operating temperatures (500–800°C) mean that minimal or no thermal insulation is required for self-heated devices. For good tube lifetime BeO plasma tubes are required due to the reactivity of the metal vapors. Usually helium at a pressure of up to one atmosphere is used as the buffer gas. Typical output powers at 5 kHz pulse repetition frequency are of order 1 W ($\sim$0.1% wall plug efficiency) at 430.5 nm from the He Sr^+ laser and 0.7 W at 373.7 nm from the He Ca^+ laser. For the high specific input power densities (10–15 W cm^{-3}) required for efficient lasing, overheating of the laser gas limits aperture scaling beyond 10–15 mm diameter. Slab laser geometries have been used successfully to aperture-scale strontium ion lasers. Scaling of the laser tube length

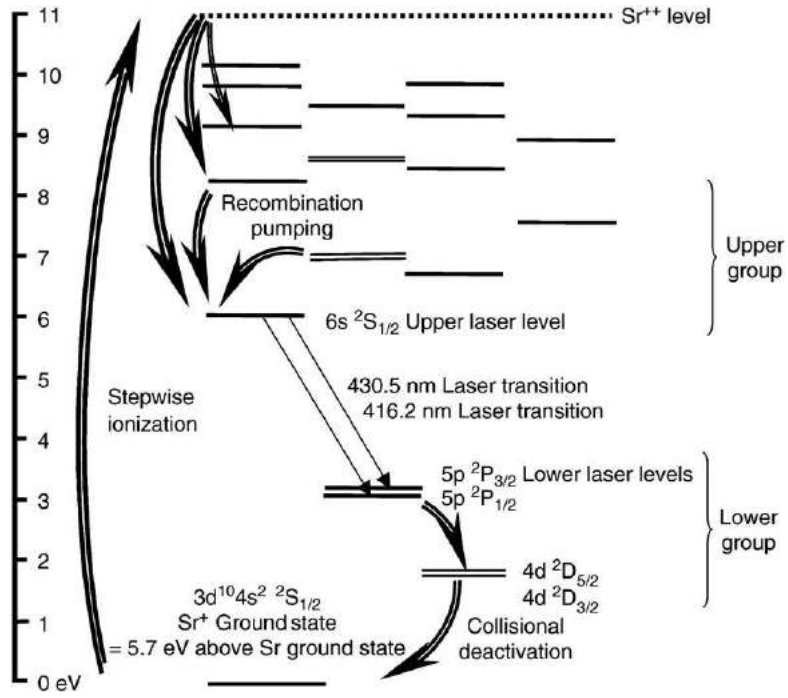

Fig. 7 Partial energy level scheme for the strontium ion laser.

beyond 0.5 m is not practical as achieving high enough excitation voltages for efficient lasing becomes problematic. Gain in strontium and calcium ion lasers is lower than in resonance-metastable metal vapor lasers such as the CVL, hence optimum output coupler reflectivity is approximately 70%. The pulse duration is also longer at around 200–500 ns resulting in moderate beam quality with plane–plane resonators.

For both strontium and calcium ion lasers the two principal transitions share upper laser levels and hence exhibit gain competition such that without wavelength-selective cavities usually only the longer wavelength of the pair is produced. With wavelength-selective cavities up to 60% (Sr^+) and 30% (Ca^+) of the normal power can be obtained at the shorter wavelength.

Many potential applications exist for strontium and calcium ion lasers, given their ultraviolet (UV)/violet wavelengths. Of particular importance is fluorescence spectroscopy in biology and forensics, treatment of neonatal jaundice, stereolithography, micromachining and exposing photoresists for integrated circuit manufacture. Despite these many potential applications, technical difficulties in power scaling mean that both strontium ion and calcium ion lasers have only limited commercial availability.

Continuous-Wave Metal Ion Lasers

In a discharge excited noble gas, there can be a large concentration of noble gas atoms in excited metastable states and noble gas ions in their ground states. These species can transfer their energy to a minority metal species (M) via two key processes: either charge transfer (Duffendack reactions) with the noble gas ions (N^+):

$$M + N^+ \rightarrow M^{+*} + N + \Delta E \text{ (where } M^{+*} \text{ is an excited metal ion state)}$$

or Penning ionization with a noble gas atom in an excited metastable state (N^*):

$$M + N^* \rightarrow M^{+*} + N + e^-$$

In a continuous discharge in a mixture of a noble gas and a metal vapor, steady generation of excited metal ions via energy transfer processes can lead to a steady state population inversion on one or more pairs of levels in the metal ion and hence produce cw lasing.

Several hundred metal ion laser transitions have been observed to lase in a host of different metals. The most important such laser is the helium cadmium laser, which has by far the largest market volume by number of unit sales of all the metal vapor lasers. In a helium cadmium laser, Cd vapor is present at a concentration of about 1–2% in a helium buffer gas which is excited by a dc discharge. Excitation to the upper laser levels of Cd (Fig. 8) is primarily via Penning ionization collisions with He^+ 2^3S_1 metastable ions produced in the discharge. Excitation via electron-impact excitation from the ion ground state may also play an important role in establishing a population inversion in Cd^+ lasers. Population inversion can be sustained continuously because the $^2P_{3/2}$ and $^2P_{1/2}$ lower laser levels decay via strong resonance transitions to the $^2S_{1/2}$ Cd^+ ground state, unlike in the

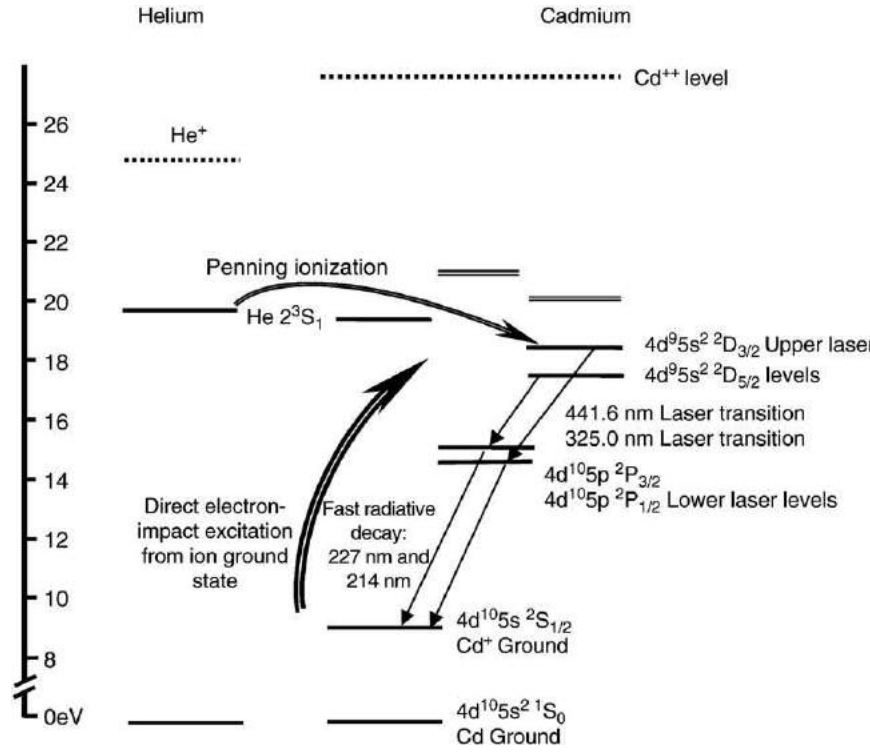

Fig. 8 Partial energy level diagram for helium and cadmium giving HeCd laser transitions.

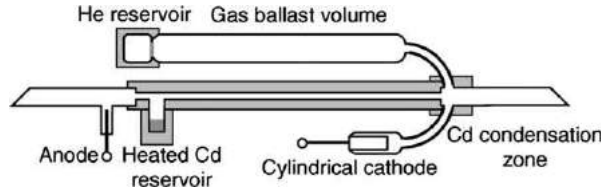

Fig. 9 Helium cadmium laser tube construction.

self-terminating metal vapor lasers. Two principal wavelengths can be produced, namely 441.6 nm (blue) and 325.0 nm (UV) with cw powers up to 200 mW and 50 mW available respectively from commercial devices.

Typical laser tube construction (Fig. 9) consists of a 1–3 mm diameter discharge channel typically 0.5 m long. A pin anode is used at one end of the laser tube together with a large-area cold cylindrical cathode located in a side-arm at the other end of the tube. Cadmium is transported to the main discharge tube from a heated Cd reservoir in a side-arm at 250–300 °C. With this source of atoms at the anode end of the laser a cataphoresis process (in which the positively charged Cd ions are propelled towards the cathode end of the tube by the longitudinal electric field) transports Cd into the discharge channel. Thermal insulation of the discharge tube ensures that it is kept hotter than the Cd reservoir to prevent condensation of Cd from blocking the tube bore. A large-diameter Cd condensation region is provided at the cathode end of the discharge channel. A large-volume side arm is also provided to act as a gas ballast for maintaining correct He pressure. As He is lost through sputtering and diffusion through the Pyrex glass tube walls, it is replenished from a high-pressure He reservoir by heating a permeable glass wall separating the reservoir from the ballast chamber which allows He to diffuse into the ballast chamber. Typical commercial sealed-off Cd lasers have operating lifetimes of several thousand hours. Overall laser construction is not that much more complex than a HeNe laser, hence unit costs are considerably lower than low-power argon ion lasers which also provide output in the blue. Usually Brewster angle windows are provided together with a high Q stable resonator (97–99% reflectivity output coupler) to provide polarized output.

With their blue and UV wavelengths and relatively low cost (compared to low-power argon ion lasers), HeCd lasers have found wide application in science, medicine and industry. Of particular relevance is their application for exposing photoresists where the blue wavelength provides a good match to the peak photosensitivity of photoresist materials. A further key application is in stereolithography where the UV wavelength is used to cure an epoxy resin. By scanning the UV beam in a raster pattern across the surface of a liquid epoxy, a solid three-dimensional object may be built up in successive layers.

Further Reading

Little, C.E., 1999. Metal Vapor Lasers. Chichester, UK: John Wiley.
Ivanov, I.G., Latush, E.L., Sem, M.F., 1996. Metal Vapor Ion Lasers, Kinetic Processes and Gas Discharges. Chichester, UK: John Wiley.
Little CE and Sabotinov NV (eds) (1996) *Pulsed Metal Vapor Lasers*. Dordrecht, The Netherlands: Kluwer.
Lyabin, N.A., Chursin, A.D., Ugol'nikov, S.A., Koroleva, M.E., Kazaryan, M.A., 2001. Development, production, and application of sealed-off copper and gold vapor lasers. Quantum Electronics 31, 191–202.
Petrash, G.G. (Ed.), 1989. Metal Vapor and Metal Halide Vapor Lasers. Commack, NY: Nova Science Publishers.
Silfvast, W.T., 1996. Laser Fundamentals. New York: Cambridge University Press.

Noble Gas Ion Lasers

WB Bridges, California Institute of Technology, Pasadena, CA, USA

History

The argon ion laser was discovered in early 1964, and is still commercially available in 2004, with about $70 million in annual sales 40 years after this discovery. The discovery was made independently and nearly simultaneously by four different groups; for three of the four, it was an accidental result of studying the excitation mechanisms in the mercury ion laser (historically, the first ion laser), which had been announced only months before. For more on the early years of ion laser research and development, see the articles listed in the Further Reading section at the end of this article.

The discovery was made with pulsed gas discharges, producing several wavelengths in the blue and green portions of the spectrum. Within months, continuous operation was demonstrated, as well as oscillation on many visible wavelengths in ionized krypton and xenon. Within a year, over 100 wavelengths were observed to oscillate in the ions of neon, argon, krypton, and xenon, spanning the spectrum from ultraviolet to infrared; oscillation was also obtained in the ions of other gases, for example, oxygen, nitrogen, and chlorine. A most complete listing of all wavelengths observed as gaseous ion lasers is given in the Laser Handbook cited in the Further Reading section. Despite the variety of materials and wavelengths demonstrated, however, it is the argon and krypton ion lasers that have received the most development and utilization.

Continuous ion lasers utilize high current density gas discharges, typically 50 A or more, and 2–5 mm in diameter. Gas pressures of 0.2 to 0.5 torr result in longitudinal electric fields of a few V/cm of discharge, so that the power dissipated in the discharge is typically 100 to 200 W/cm. Such high-power dissipation required major technology advances before long-lived practical lasers became available. Efficiencies have never been high, ranging from 0.01% to 0.2%. A typical modern ion laser may produce 10 W output at 20 kW input power from 440 V three-phase power lines and require 6–8 gallons/minute of cooling water. Smaller, air-cooled ion lasers, requiring 1 kW of input power from 110 V single-phase mains can produce 10–50 mW output power, albeit at even lower efficiency.

Theory of Operation

The strong blue and green lines of the argon ion laser originate from transitions between the 4p upper levels and 4s lower levels in singly ionized argon, as shown in **Fig. 1**. The 4s levels decay radiatively to the ion ground state. The strongest of these laser lines are listed in **Table 1**. The notation used for the energy levels is that of the L–S coupling model. The ion ground state electron configuration is $3s^2 3p^5 (^2P^o_{3/2})$. The inner ten electrons have the configuration $1s^2 2s^2 2p^6$, but this is usually omitted for brevity. The excited states shown in **Fig. 1** result from coupling a 4p or 4s electron to a $3s^2 3p^4 (^3P)$ core, with the resulting quantum numbers S (the net spin of the electrons), L (the net angular momentum of the electrons), and J (the angular momentum resulting from

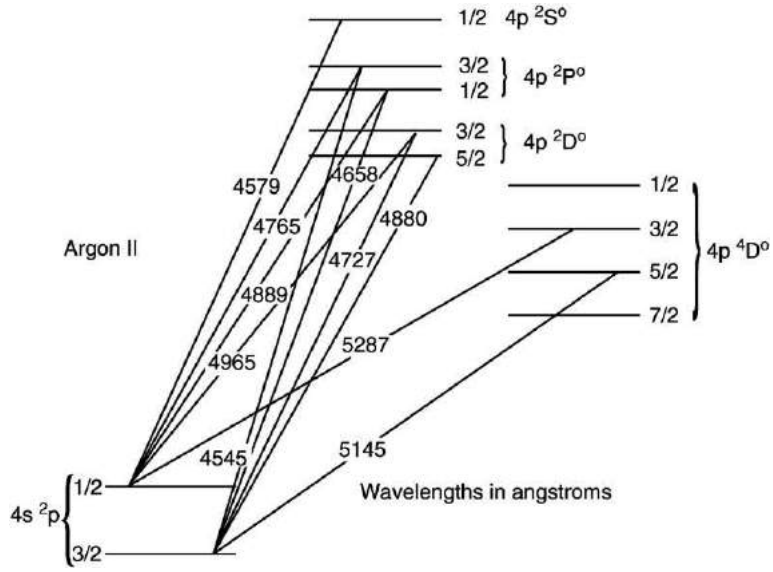

Fig. 1 4p and 4s doublet levels in singly ionized argon, showing the strongest blue and green laser transitions.

 doi:10.1016/B978-0-12-809283-5.01249-0

Table 1 Ar II laser blue-green wavelengths

Wavelength (nanometers)	*Transition*[a] *(upper level)→(lower level)*	*Relative strength*[b] *(Watt)*
454.505	$4p\ ^2P^o_{3/2} \rightarrow 4s\ ^2P_{3/2}$	0.8
457.935	$4p\ ^2S^o_{1/2} \rightarrow 4s\ ^2P_{1/2}$	1.5
460.956	$(^1D)4p\ ^2F^o_{7/2} \rightarrow (^1D)4s\ ^2D_{5/2}$	–
465.789	$4p\ ^2P^o_{1/2} \rightarrow 4s\ ^2P_{3/2}$	0.8
472.686	$4p\ ^2D^o_{3/2} \rightarrow 4s\ ^2P_{3/2}$	1.3
476.486	$4p\ ^2P^o_{3/2} \rightarrow 4s\ ^2P_{1/2}$	3.0
487.986	$4p\ ^2D^o_{5/2} \rightarrow 4s\ ^2P_{3/2}$	8.0
488.903	$4p\ ^2P^o_{1/2} \rightarrow 4s\ ^2P_{1/2}$	[c]
496.507	$4p\ ^2D^o_{3/2} \rightarrow 4s\ ^2P_{1/2}$	3.0
501.716	$(^1D)4p\ ^2D^o_{5/2} \rightarrow 3d\ ^2D_{3/2}$	1.8
514.179	$(^1D)4p\ ^2F^o_{7/2} \rightarrow 3d\ ^2D_{5/2}$	[c]
514.532	$4p\ ^4D^o_{5/2} \rightarrow 4s\ ^2P_{3/2}$	10
528.690	$4p\ ^4D^o_{3/2} \rightarrow 4s\ ^2P_{1/2}$	1.8

[a]All levels are denoted in *L–S* coupling with the (^{3}P) core unless otherwise indicated. Odd parity is denoted by the superscript 'o'.
[b]Relative strengths are given as power output from a commercial Spectra-Physics model 2080-25S argon ion laser.
[c]These lines may oscillate simultaneously with the nearby strong line, but are not resolved easily in the output beam.

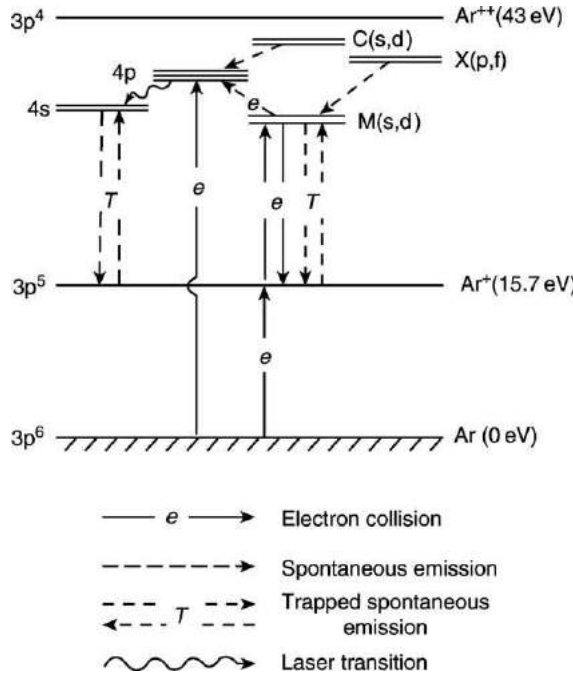

Fig. 2 Schematic representation of energy levels in neutral and singly ionized argon, indicating alternative pathways for excitation and de-excitation of the argon ion laser levels.

coupling S to L) are represented by $^{2S+1}L_J$, where $L=0,1,2,\ldots$ is denoted S, P, D, F,… The superscript 'o' denotes an odd level, while even levels omit a superscript. Note that some weaker transitions involving levels originating from the $3s^23p^4(^1D)$ core configuration also oscillate. Note also that the quantum mechanical selection rules for the *L–S* coupling model are not rigorously obeyed, although the stronger laser lines satisfy most or all of these rules. The selection rule $|\Delta J|=1$ or 0, but not $J=0\rightarrow J=0$ is always obeyed. All transitions shown in Fig. 1 and Table 1 belong to the second spectrum of argon, denoted Ar II. Lines originating from transitions in the neutral atom make up the first spectrum, Ar I; lines originating from transitions in doubly ionized argon are denoted Ar III, and so forth for even more highly ionized states.

Much work has been done to determine the mechanisms by which the inverted population is formed in the argon ion laser. Reviews of this extensive research are found in the Further Reading section. While a completely quantitative picture of argon ion laser operation is lacking to this day, the essential processes are known. Briefly, the 4p upper levels are populated by three pathways, as illustrated in Fig. 2:

(i) by electron collision with the $3p^6$ neutral ground state atoms. This 'sudden perturbation process' requires at least 37 eV electrons, and also singles out the $4p\ ^2P^o_{3/2}$ upper level, which implies that only the 476 and 455 nm lines would oscillate.

This is the behavior seen in pulsed discharges at very low pressure and very high axial electric field. This pathway probably contributes little to the 4p population under ordinary continuous wave (cw) operating conditions, however.

(ii) by electron collision from the lowest-lying s and d states in the ion, denoted M(s,d) in **Fig. 1**. This only requires 3–4 eV electrons. These states have parity-allowed transitions to the ion ground state, but are made effectively metastable by radiation trapping (that is, there is a high probability that an emitted photon is re-absorbed by another ground state ion before it escapes the discharge region), or by requiring $|\Delta J|$ to be 2 to make the transition (forbidden by quantum selection rules). Thus, these levels are both created and destroyed primarily by electron collision, causing the population of the M(s,d) states to follow the population of the singly ionized ground state, which, in turn, is approximately proportional to the discharge current. Since a second electron collision is required to get from the M(s,d) states to the 4p states, a quadratic variation with current for the laser output power would be expected, and that is what is observed over some reasonable range of currents between threshold and saturation. Note that radiative decay from higher-lying opposite-parity p and f states, denoted X(p,f), can also contribute to the population of M(s,d), but the linear variation in population of the M(s,d) with discharge current is assured by electron collision creation and destruction.

(iii) by radiative decay from higher-lying opposite-parity s and d states, denoted C(s,d). These states are populated by electron collision with the $3p^5$ ion ground states, and thus have populations that also vary quadratically with discharge current. The contribution of this cascade process has been measured to be 20% to 50% of the 4p upper laser level population.

Note that it is not possible to distinguish between processes (ii) and (iii) by the variation of output power with discharge current; both give the observed quadratic dependence.

The radiative lifetimes of the 4s ^{2}P levels are sufficiently short to depopulate the lower laser levels by radiative decay. However, radiation trapping greatly lengthens this decay time, and a bottleneck can occur. In pulsed ion lasers, this is exhibited by the laser pulse terminating before the excitation current pulse ends, which would seem to preclude continuous operation. However, the intense discharge used in continuous operation heats the ions to the order of 2300 K, thus greatly Doppler broadening the absorption linewidth and reducing the magnitude of the absorption. Additionally, the ions are attracted to the discharge tube walls, so that the absorption spectrum is further broadened by the Doppler shift due to their wall-directed velocities. The plasma wall sheath gives about 20 V drop in potential from the discharge axis to the tube wall, so most ions hit the wall with 20 eV of energy, or about ten times their thermal velocity. A typical cw argon ion laser operates at ten times the gas pressure that is optimum for a pulsed laser, and thus the radiation trapping of the $4s \rightarrow 3p^5$ transitions is so severe that it may take several milliseconds after discharge initiation for the laser oscillation to begin.

As the discharge current is increased, the intensities of the blue and green lines of Ar II eventually saturate, and then decrease with further current. At these high currents, there is a buildup in the population of doubly ionized atoms, and some lines of Ar III can be made to oscillate with the appropriate ultraviolet mirrors. **Table 2** lists the strongest of these lines, those that are available in the largest commercial lasers. Again, there is no quantitative model for the performance in terms of the discharge parameters, but the upper levels are assumed to be populated by processes analogous to those of the Ar II laser. At still higher currents, lines in Ar IV can be made to oscillate as well.

Table 2 Ultraviolet argon ion laser wavelengths

Wavelength (nanometers)	*Spectrum*	*Transition*[a] *(upper level)→(lower level)*	*Relative strength*[b] *(Watt)*
275.392	III	$(^2D^o)4p\ ^1D_2 \rightarrow (^2D^o)4s\ ^1D^o_2$	0.3
275.6	?	?	0.02
300.264	III	$(^2P^o)4p\ ^1P_1 \rightarrow (^2P^o)3d\ ^1D^o_2$	0.5
302.405	III	$(^2P^o)4p\ ^3D_3 \rightarrow (^2P^o)4s\ ^3P^o_2$	0.5
305.484	III	$(^2P^o)4p\ ^3D_2 \rightarrow (^2P^o)4s\ ^3P^o_1$	0.2
333.613	III	$(^2D^o)4p\ ^3F_4 \rightarrow (^2D^o)4s\ ^3D^o_3$	0.4
334.472	III	$(^2D^o)4p\ ^3F_3 \rightarrow (^2D^o)4s\ ^3D^o_2$	0.8
335.849	III	$(^2D^o)4p\ ^3F_2 \rightarrow (^2D^o)4s\ ^3D^o_1$	0.8
350.358	III	$(^2D^o)4p\ ^3D_2 \rightarrow (^2D^o)4s\ ^3D^o_2$	0.05
350.933	III	$(^4S^o)4p\ ^3P_0 \rightarrow (^4S^o)4s\ ^3S^o_1$	0.05
351.112	III	$(^4S^o)4p\ ^3P_2 \rightarrow (^4S^o)4s\ ^3S^o_1$	2.0
351.418	III	$(^4S^o)4p\ ^3P_1 \rightarrow (^4S^o)4s\ ^3S^o_1$	0.7
363.789	III	$(^2D^o)4p\ ^1F_3 \rightarrow (^2D^o)4s\ ^1D^o_2$	2.5
379.532	III	$(^2P^o)4p\ ^3D_3 \rightarrow (^2P^o)3d\ ^3P^o_2$	0.4
385.829	III	$(^2P^o)4p\ ^3D_2 \rightarrow (^2P^o)3d\ ^3P^o_1$	0.15
390.784	III	$(^2P^o)4p\ ^3D_1 \rightarrow (^2P^o)3d\ ^3P^o_0$	0.02
408.904	IV?	?	0.04
414.671	III	$(^2D^o)4p\ ^3P_2 \rightarrow (^2P^o)4s\ ^3P^o_2$	0.02
418.298	III	$(^2D^o)4p\ ^1P_1 \rightarrow (^2D^o)4s\ ^1D^o_2$	0.08

[a]All levels are denoted in *L–S* coupling with the core shown in ( ). Odd parity is denoted by the superscript 'o'.

[b]Relative strengths are given as power output from a commercial Spectra-Physics model 2085-25S argon ion laser.

Table 3 Krypton ion laser wavelengths

Wavelength (nanometers)	*Spectrum*	*Transition*[a] *(upper level)→(lower level)*	*Relative strength*[b] *(Watt)*
337.496	III	$(^2P^o)5p\ ^3D_3 \rightarrow (^2P^o)5s\ ^3P^o_2$	–
350.742	III	$(^4S^o)5p\ ^3P_2 \rightarrow (^4S^o)5s\ ^3S^o_1$	1.5
356.432	III	$(^4S^o)5p\ ^3P_1 \rightarrow (^4S^o)5s\ ^3S^o_1$	0.5
406.737	III	$(^2D^o)5p\ ^1F_3 \rightarrow (^2D^o)5s\ ^1D^o_2$	0.9
413.133	III	$(^4S^o)5p\ ^5P_2 \rightarrow (^4S^o)5s\ ^3S^o_1$	1.8
415.444	III	$(^2D^o)5p\ ^3F_3 \rightarrow (^2D^o)5s\ ^1D^o_2$	0.3
422.658	III	$(^2D^o)5p\ ^3F_2 \rightarrow (^2D^o)4d\ ^3D^o_1$	–
468.041	II	$(^3P)5p\ ^2S^o_{1/2} \rightarrow (^3P)5s\ ^2P_{1/2}$	0.5
476.243	II	$(^3P)5p\ ^2D^o_{3/2} \rightarrow (^3P)5s\ ^2P_{1/2}$	0.4
482.518	II	$(^3P)5p\ ^4S^o_{3/2} \rightarrow (^3P)5s\ ^2P_{1/2}$	0.4
520.832	II	$(^3P)5p\ ^4P^o_{3/2} \rightarrow (^3P)5s\ ^4P_{3/2}$	–
530.865	II	$(^3P)5p\ ^4P^o_{5/2} \rightarrow (^3P)5s\ ^4P_{3/2}$	1.5
568.188	II	$(^3P)5p\ ^4D^o_{5/2} \rightarrow (^3P)5s\ ^2P_{3/2}$	0.6
631.024	III	$(^2D^o)5p\ ^3P_2 \rightarrow (^2P^o)4d\ ^3D_1$	0.2
647.088	II	$(^3P)5p\ ^4P^o_{5/2} \rightarrow (^3P)5s\ ^2P_{3/2}$	3.0
676.442	II	$(^3P)5p\ ^4P^o_{1/2} \rightarrow (^3P)5s\ ^2P_{1/2}$	0.9
752.546	II	$(^3P)5p\ ^4P^o_{3/2} \rightarrow (^3P)5s\ ^2P_{1/2}$	1.2
793.141	II	$(^1D)5p\ ^4F^o_{7/2} \rightarrow (^3P)4d\ ^2F_{5/2}$	0.3
799.322	II	$(^3P)5p\ ^4P^o_{3/2} \rightarrow (^3P)4d\ ^4D_{1/2}$	

[a]All levels are denoted in *L–S* coupling with the core shown in ( ). Odd parity is denoted by the superscript 'o'.
[b]Relative strengths are given as power output from a commercial Spectra-Physics model 2080RS ion laser.

Much less research has been done on neon, krypton, and xenon ion lasers, but it is a good assumption that the population and depopulation processes are the same in these lasers. **Table 3** lists both the Kr II and Kr III lines that are available from the largest commercial ion lasers. Oscillation on lines in still-higher ionization states in both krypton and xenon have been observed.

Operating Characteristics

A typical variation of output power with discharge current for an argon ion laser is shown in **Fig. 3**. This particular laser had a 4 mm diameter discharge in a water-cooled silica tube, 71 cm in length, with a 1 kG axial magnetic field. The parameter is the argon pressure in the tube before the discharge was struck. Note that no one curve is exactly quadratic, but that the envelope of the curves at different filling pressures is approximately quadratic. At such high discharge current densities (50 A in the 4 mm tube is approximately 400 A/cm^2) there is substantial pumping of gas out of the small-bore discharge region. Indeed, a return path for this pumped gas must be provided from anode to cathode ends of the discharge to keep the discharge from self-extinguishing. The axial electric field in this discharge was 3–5 V/cm, so the input power was of the order of 10 to 20 kW, yielding an efficiency of less than 0.1%, an unfortunate characteristic of all ion lasers.

Technology

With such high input powers required in a small volume to produce several watts output, ion laser performance has improved from 1964 to the present only as new discharge technologies were introduced. The earliest laboratory argon ion lasers used thin-walled ($\approx$1 mm wall thickness) fused silica discharge tubes, cooled by flowing water over the outside wall of the tube. The maximum input power per unit length of discharge was limited by thermal stresses in the silica walls caused by the temperature differential from inside to outside. Typically, ring-shaped cracks would cause the tube to fail catastrophically. Attempts to make metal–ceramic structures with alumina (Al_2O_3) discharge tubes to contain the plasma were made early on (1965) but were not successful. Such tubes invariably failed from fracture by thermal shock as the discharge was turned on. Later, successful metal–ceramic tubes were made with beryllia (BeO), which has a much higher thermal conductivity than silica or alumina and is much more resistant to thermal shock. Today, all of the lower power (less than 100 mW output) are made with BeO discharge tubes. Some ion lasers in the 0.5 to 1 W range are also made with water-cooled BeO discharge tubes.

A typical low-power air cooled argon ion laser is shown in **Fig. 4**. The large metal can (2) on the right end of the tube contains an impregnated-oxide hot cathode, heated directly by current through ceramic feed-through insulators (3). The small ($\approx$1 mm diameter) discharge bore runs down the center of the BeO ceramic rod (1), and several smaller diameter gas return path holes run off-axis parallel to the discharge bore to provide the needed gas equalization between cathode can and anode region. A copper honeycomb cooler is brazed to the cathode can, two more to the outer wall of the BeO cylinder, and one to the anode (4) at the left end of the tube. The laser mirrors (5) are glass-fritted to the ends of the tube, forming a good vacuum seal. Note that in this small laser, the active discharge bore length is less than half of the overall length.

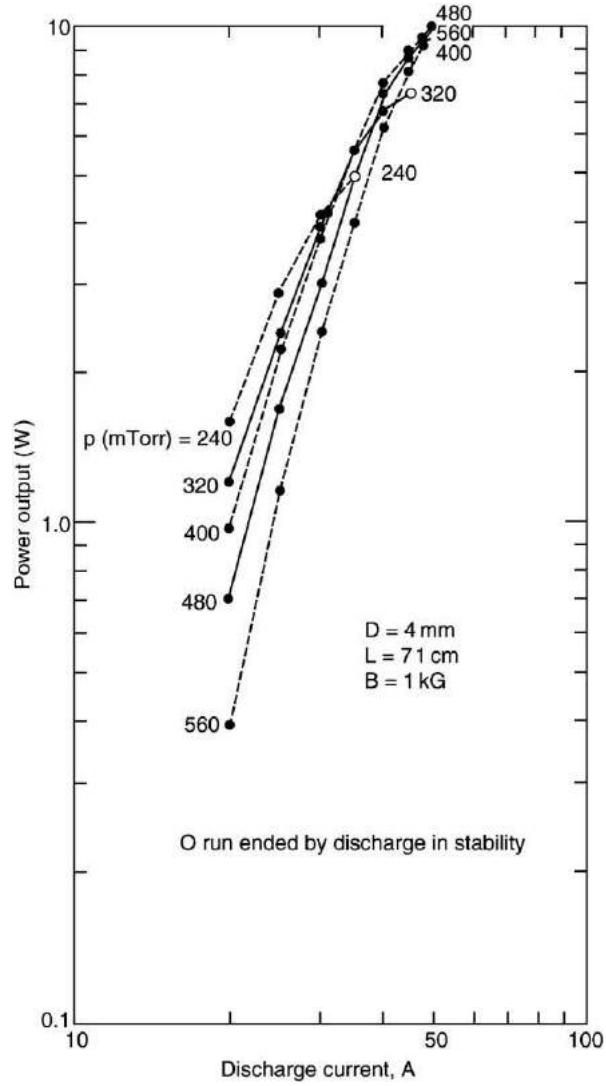

Fig. 3 Laser output power (summed over all the blue and green laser lines) versus discharge current for a laser discharge 4 mm in diameter and 71 cm long, with a 1 kilogauss longitudinal magnetic field. The parameter shown is the argon fill pressure in mTorr prior to striking the discharge. The envelope of the curves exhibits the quadratic variation of output power with discharge current.

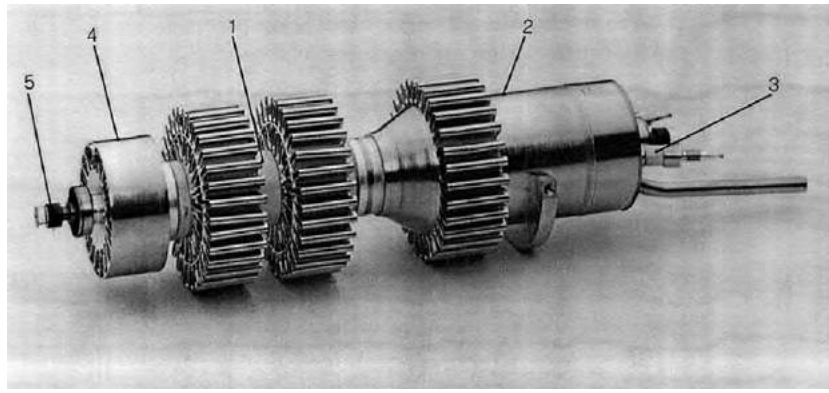

Fig. 4 A typical commercial low-power, air-cooled argon ion laser. The cathode can (2) is at the right and the beryllia bore (1) and anode (4) is at the left. The cathode current is supplied through ceramic vacuum feed-throughs (3). Mirrors (5) are glass fritted onto the ends of vacuum envelope. (Photo courtesy of JDS Uniphase Corp.)

The very early argon ion lasers were made with simple smooth dielectric tubes, allowing the continuous variation in voltage along the length of the tube required by the discharge longitudinal electric field. However, this variation can be step-wise at the discharge walls, and still be more or less smooth along the axis. Thus, the idea arose of using metal tube segments, insulated one from another, to form the discharge tube walls. The first version of this idea (1965) used short (≈ 1 cm) metal cylinders supported by metal disks and stacked inside a large diameter silica envelope, with each metal cylinder electrically isolated from the others. The tubes and disks were made of molybdenum, and were allowed to heat to incandescence, thus radiating several kilowatts of heat through the silica vacuum envelope to a water-cooled collector outside. While this eliminated the problems of thermal shock and poor thermal conductivity inherent in dielectric wall discharges, it made another problem painfully evident. The intense ion bombardment of the metal tube walls sputtered the wall material, eventually eroding the shape of the metal cylinders and depositing metal films on the insulating wall material, thus shorting one segment to another.

Many different combinations of materials and configurations were investigated in the 1960s and 1970s to find a structure that would offer good laser performance and long life. It was found that simple thin metal disks with a central hole would effectively confine the discharge to a small diameter (on the order of the hole size) if a longitudinal d-c magnetic field of the order of 1 kiloGauss were used. The spacing between disks can be as large as 2–4 discharge diameters. Of the metals, tungsten has the lowest sputtering yield for argon ions in the 20 eV range (which is approximately the energy they gain in falling to the wall across the discharge sheath potential difference). An even lower sputtering yield is exhibited by carbon, and graphite cylinders contained within a larger diameter silica or alumina tube were popular for a while for ion laser discharges. Unfortunately, graphite has a tendency to flake or powder, so such laser discharge tubes became contaminated with 'dust' which could eventually find its way to the optical windows of the tube. Beryllia and silica also sputter under argon ion bombardment, but with still lower yields than metals or carbon; however, their smaller thermal conductivities limit them to lower-power applications.

The material/configuration combination that has evolved for higher-power argon ion lasers today is to use a stack of thin tungsten disks with 2–3 mm diameter holes for the discharge. These disks, typically 1 cm in diameter, are brazed coaxially to a larger copper annulus (actually, a drawn cup with a 1 cm hole on its axis). A stack of these copper/tungsten structures is, in turn, brazed to the inside wall of a large diameter alumina vacuum envelope. The tungsten disks are exposed to and confine the discharge, while the copper cups conduct heat radially outward to the alumina tube wall, which, in turn, is cooled by fluid flow over its exterior. Thus, the discharge is in contact only with a low-sputtering material (tungsten) while the heat is removed by a high thermal conductivity material (copper). Details differ among manufacturers, but this 'cool disk' technology seems to have won out in the end. A photo of a half-sectioned disk/cup stacked assembly from a Coherent Innova™ ion laser is shown in **Fig. 5**. The coiled impregnated-tungsten cathode is also shown.

Sputtering of the discharge tube walls is not uniform along the length of the gas discharge. The small diameter region where the laser gain occurs is always joined to larger diameter regions containing cathode and anode electrodes (as shown in **Fig. 5**, for example). A plasma double sheath (that is, a localized increase in potential) forms across the discharge in the transition region

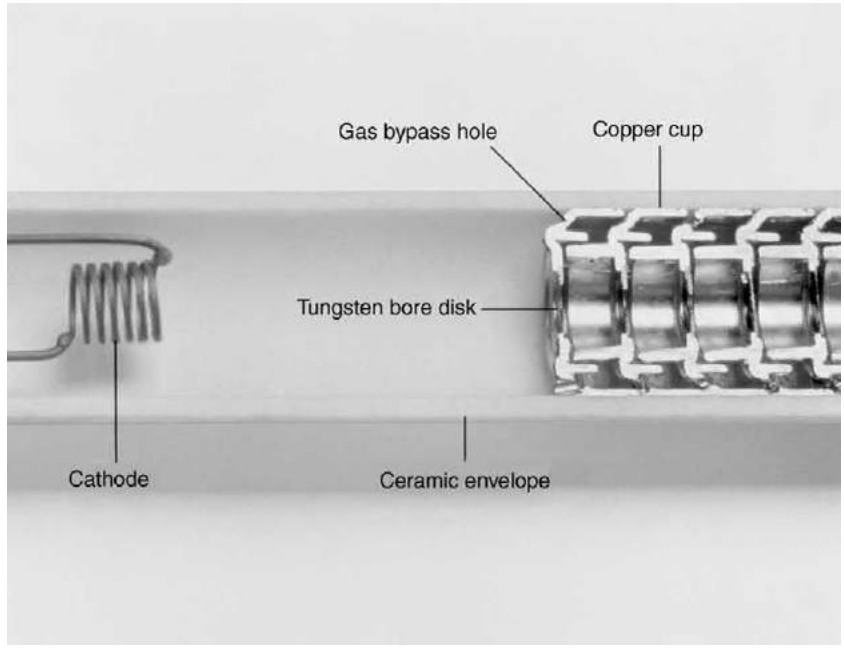

Fig. 5 Discharge bore structure of a typical commercial high-power argon ion laser, the Coherent Innova™. Copper cups are brazed to the inner wall of a ceramic envelope, which is cooled by liquid flow over its outer surface. A thin tungsten disk with a small hole defining the discharge path is brazed over a larger hole in the bottom of the copper cup. Additional small holes in the copper cup near the ceramic envelope provide a gas return path from cathode to anode. Also shown is the hot oxide-impregnated tungsten cathode. (Photo courtesy of Coherent, Inc.)

between large and small diameter regions (the discharge 'throats,' which may be abrupt or tapered). This sheath is required to satisfy the boundary conditions between the plasmas of different temperatures in the different regions. Such a double sheath imparts additional energy to the ions as they cross the sheath, perhaps an additional 20 eV. When these ions eventually hit the discharge walls near the location of the sheath, they have 40 eV of energy rather than the 20 eV from the normal wall sheath elsewhere in the small diameter discharge. Sputtering yield (number of sputtered wall atoms per incident ion) is exponentially dependent on ion energy in this low ion energy region, so the damage done to the wall in the vicinity of the 'throat' where the double sheath forms may be more than ten times that elsewhere in the small diameter bore region. This was the downfall of high-power operation of dielectric discharge bores, even BeO; while sputtering was acceptable elsewhere in the discharge, the amount of material removed in a small region near the discharge throat would cause catastrophic bore failure at that point. This localized increase in sputtering in the discharge throat is common to all ion lasers, including modern cooled tungsten disk tubes. Eventually, disks near the throat are eroded to larger diameters, and more material is deposited on the walls nearby. Attempts to minimize localized sputtering by tapering the throat walls or the confining magnetic field have proven unsuccessful; a localized double sheath always forms somewhere in the throat. Because of the asymmetry caused by ion flow, the double sheath is larger in amplitude in the cathode throat than the anode throat, so that the localized wall damage is larger at the cathode end of the discharge than at the anode end.

Another undesirable feature of sputtering is that the sputtered material 'buries' some argon atoms when it is deposited on a wall. This is the basis of the well-known Vac-Ion® vacuum pump. Thus, the operating pressure in a sealed laser discharge tube will drop during the course of operation. In low-power ion lasers, this problem is usually solved by making the gas reservoir volume large enough to satisfy the desired operating life (for example, the large cathode can in **Fig. 4**). In high-power ion lasers, the gas loss would result in unacceptable operating life even with a large reservoir at the fill pressure. Thus, most high-power ion lasers have a gas pressure measurement system and dual-valve arrangement connected to a small high-pressure reservoir to 'burp' gas into the active discharge periodically to keep the pressure within the operating range. Unfortunately, if a well-used ion laser is left inoperative for months, some of the 'buried' argon tends to leak back into the tube, with the gas pressure becoming higher than optimum. The pressure will gradually decrease to its optimum value with further operation of the discharge (in perhaps tens of hours). In the extreme case, the gas pressure can rise far enough so that the gas discharge will not strike, even at the maximum power supply voltage. Such a situation requires an external vacuum pump to remove the excess gas, usually a factory repair.

In addition to simple dc discharges, various other techniques have been used to excite ion lasers, primarily in a search for higher power, improved efficiency and longer operating life. Articles in the Further Reading section give references to these attempts. Radio-frequency excitation at 41 MHz was used in an inductively coupled discharge, with the laser bore and its gas return path forming a rectangular single turn of an air-core transformer. A commercial product using this technique was sold for a few years in the late 1960s. Since this was an 'electrode-less' discharge, ion laser lines in reactive gases such as chlorine could be made to oscillate, as well as the noble gases, without 'poisoning' the hot cathode used in conventional dc discharge lasers. A similar electrode-less discharge was demonstrated as a quasi-cw laser by using iron transformer cores and exciting the discharge with a 2.5 kHz square wave. Various microwave excitation configurations at 2.45 GHz and 9 GHz also resulted in ion laser oscillation. Techniques common to plasma fusion research were also studied. Argon ion laser oscillation was produced in Z-pinch and Θ-pinch discharges and also by high-energy (10–45 keV) electron beams. However, none of these latter techniques resulted in a commercial product, and all had efficiencies worse than the simple dc discharge lasers.

It is interesting that the highest-power output demonstrations were made in less than ten years after the discovery. Before 1970, 100 W output on the argon blue-green lines was demonstrated with dc discharges two meters in length. In 1970, a group in the Soviet Union reported 500 W blue-green output from a two-meter discharge with 250 kW of dc power input, an efficiency of 0.2%. Today, the highest output power argon ion laser offered for sale is 50 W output.

Manufacturers

More than 40 companies have manufactured ion lasers for sale over the past four decades. This field has now (2004), narrowed to the following:

Coherent, Inc.	http://www.coherentinc.com
INVERSion Ltd.	http://inversion.iae.nsk.su
JDS Uniphase	http://www.jdsu.com
Laser Physics, Inc.	http://www.laserphysics.com
Laser Technologies GmbH	http://www.lg-lasertechnologies.com
LASOS Lasertechnik GmbH	http://www.LASOS.com
Lexel Laser, Inc.	http://www.lexellaser.com
Melles Griot	http://lasers.mellesgriot.com
Spectra-Physics, Inc.	http://www.spectraphysics.com

Further Reading

Bridges, W.B., 1979. Atomic and ionic gas lasers. In: Marton, L., Tang, C.L. (Eds.), Methods of Experimental Physics, Vol 15: Quantum Electronics, Part A. New York: Academic Press, pp. 31–166.

Bridges, W.B., 1982. Ionized gas lasers. In: Weber, M.J. (Ed.), Handbook of Laser Science and Technology, Vol. II, Gas Lasers. Boca Raton, FL: CRC Press, pp. 171–269.
Bridges, W.B., 2000. Ion lasers – the early years. IEEE Journal of Selected Topics in Quantum Electronics 6, 885–898.
Davis, C.C., King, T.A., 1975. Gaseous ion lasers. In: Goodwin, D.W. (Ed.), Advances in Quantum Electronics, vol. 3. New York: Academic Press, pp. 169–469.
Dunn, M.H., Ross, J.N., 1976. The argon ion laser. In: Sanders, J.H., Stenholm, S. (Eds.), Progress in Quantum Electronics, vol. 4. New York: Pergamon, pp. 233–269.
Weber, M.J., 2000. Handbook of Laser Wavelengths. Boca Raton, FL: CRC Press.

Planar Waveguide Lasers

S Bhandarkar, Alfred University, Alfred, NY, USA

Nomenclature

a integer denoting number of laser modes
E electric field
k_z propagation constant for travel in z direction
L_c length of laser cavity
L_G length of grating
m order of the grating, an integer
n refractive index of the medium
$\overline{n}$ modal effective refractive index
R reflectance
V normalized frequency or the V parameter of the waveguide
Δn_{eff} modulation of the effective index of the grating
η confinement factor or fraction of the light confined in the region of interest
λ wavelength of the lightwave
λ_B resonant wavelength of a Bragg grating or Bragg wavelength
Λ grating pitch
Φ electric field distribution

Introduction

Lasers have evolved to become true marvels of technology. From their invention in the early 1960s, laser light has changed the way society operates, its application stretching to virtually every aspect of human activity. The versatility of lasers is derived from the fact that there are several different types, each with a unique set of advantages. Solid-state lasers, made from specific semiconducting materials, are commonly used in the optical networks that underlie the greatest machine on Earth – the Internet. High power gas lasers, such as CO_2 lasers, are used in industrial machining, whereas rare-earth doped and dye lasers have found applications ranging from eye surgery to tattoo removal.

Planar waveguide lasers are a relatively new class of lasers made possible by advances in the field of optical waveguides. Over the past decade, with the aim to miniaturize optical components, planar waveguide technology has been developed to a high degree of sophistication. Likewise, Er and other rare-earth doping was studied extensively owing to its importance in all-optical amplification. These elements turned out to be the building blocks for the development of fiber lasers and subsequently planar waveguide lasers.

This chapter will discuss the basic concepts underlying this category of lasers, the variations therein, and the main characteristics. It is important to understand the basic features of waveguides and lasers, which is included as a prelude. We will then discuss planar waveguide lasers with emphasis on the essential elements: the laser materials, the various designs, their performance, and the resultant advantages as well as the applications.

Waveguides

A waveguide is a general conduit for efficient propagation of an electromagnetic field. It is usually defined by making a region of higher refractive index within a dielectric medium. This behaves as an attractive potential well that confines light in the form of bound modes, much like a potential well that confines electrons to a bound state. The propagating field is referred to as the lightwave and is governed by Maxwell's equation. For a weakly guiding approximation, as is the case when the refractive index contrast (Δn) is small, Maxwell's equation can be reduced to a scalar wave equation for the transverse electric field. For a given wavelength (λ) and polarization, the scalar wave equation in a nonmagnetic medium is

$$\nabla^2 E(x) + \frac{4\pi^2}{\lambda^2} n(x)^2 E(x) = 0 \tag{1}$$

where E is the electric field. For a waveguide aligned in the z direction, the field is the sum of the two propagating transverse modes, each with the form $E(x) = \Phi(x,y)e^{ik_z z}$ where $\Phi(x,y)$ is the field distribution and k_z is the propagation constant. The wave equation for the transverse mode reduces to

$$\frac{\lambda^2}{4\pi^2}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\Phi(x,y) + n(x,y)^2\Phi(x,y) = (\overline{n})^2\Phi(x,y) \tag{2}$$

Here, $\overline{n}$ is referred to as the effective refractive index and is given by $\overline{n} = k_z\lambda/2\pi$. The effective index dictates the occurrence of bound modes – these states occur when $\overline{n}$ is between n_{core} and $n_{cladding}$. For smaller values of $\overline{n}$ there is a continuum of unbound and radiation modes. These are leaky modes that are dispersed quickly over a short distance. Only bound modes are guided in the core of the waveguide structure.

 doi:10.1016/B978-0-12-809283-5.00851-X

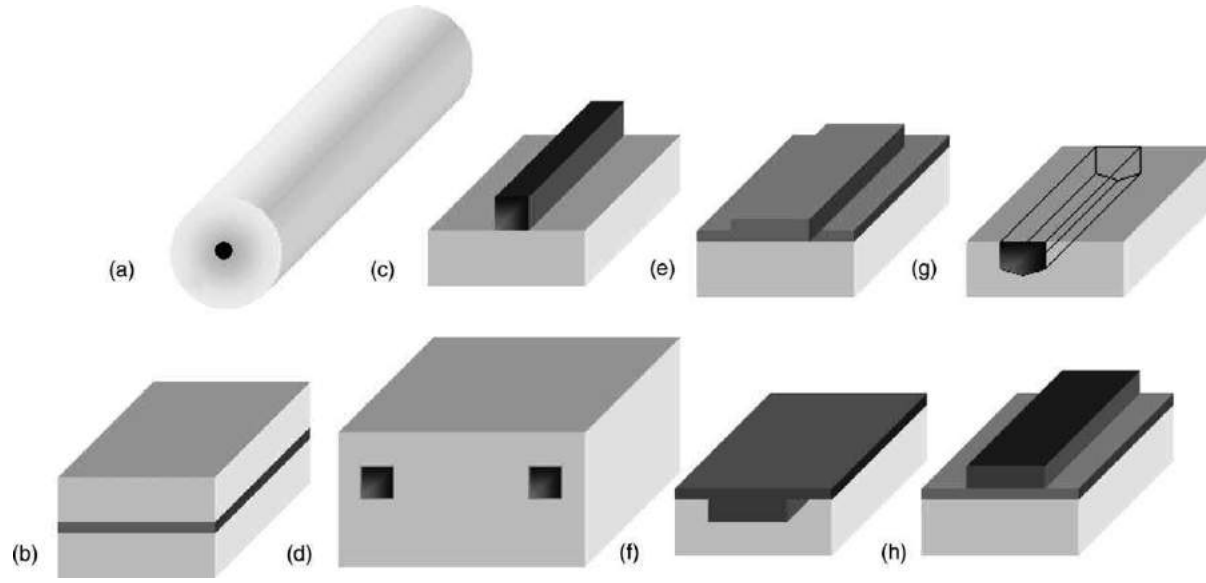

Fig. 1 Range of different waveguide designs, with the core represented by the darker regions: (a) Optical fiber; (b) slab waveguide; (c) channel waveguide; (d) embedded channel waveguides with a layer of the overcladding; (e) rib waveguide; (f) inverted rib; (g) ion diffused; and (h)strip loaded, comprising of a strip of a different higher index material (that serves to confine light) on a core layer.

Planar Waveguides

Waveguides formed on a flat substrate are called planar waveguides. These are typically made by stepwise deposition of films of dielectric materials (typically glass). The waveguide core is defined by one of several methods, the most common of which uses lithography and etching. In this case, a film of an appropriately higher index material is deposited on an 'under-cladding' film and then selectively etched to define a core. This is generally covered with a suitable 'upper-cladding' layer so that the index contrast is controlled in all directions. Other methods of forming the core include patterned ion exchange or ion implantation or increasing the index of the medium using a femtosecond laser, etc. In all these cases, the end result is the formation of an embedded higher-index region whose locus defines the optical path for the lightwave. A circuit made from these waveguides is known as planar lightwave circuits (PLC).

Waveguide cores may be designed in several different configurations, as shown in Fig. 1. These designs allow for different fabrication processes as well as the control over the modal properties of the lightwave. For a given index contrast, the number of modes traveling in the waveguide is primarily determined by the dimensions of the waveguide. The aspect ratio of the waveguide affects the propagation of the transverse electric (TE) and the transverse magnetic (TM) modes, due to different propagation constant k_z for each mode.

Lasers

The generation of laser light requires four basic conditions to be satisfied:

- a source capable of light emission;
- an energy supply source – typically electrical, optical, or chemical;
- population inversion to the excited state in the lasing medium; and
- absence of parasitic effects that would absorb the emitted light.

Laser light is often a result of either electronic or vibrational excitation. This is because these excitations involve transfer between energy states that coincide with the radiation of UV, visible, or infrared light. Population inversion is an important requirement in lasers since the emission is stimulated. This involves having a greater fraction of the atoms or molecules in the excited state so that a stimulating photon can cause a radiative decay to the lower energy level. The stimulated light from most lasing media is weak, which requires unacceptably long lengths to generate sufficient power. Instead, as shown in Fig. 2, lasers are made with two parallel mirrors with an axial light-generating source in between. This configuration is known as a resonant cavity, a resonator, or an oscillator. The back mirror is usually fully reflective but a partial front mirror allows a portion of the light in the cavity to pass through. The two mirrors cause the light to bounce to and fro, forcing it into making multiple passes through the cavity. Light emanating from this structure is thus amplified and is also coherent (that is the lightwaves are in phase with each other spatially and temporally). This has come to represent the signature output of a laser.

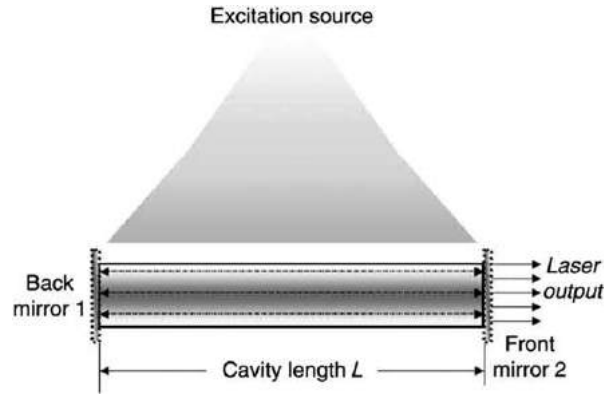

Fig. 2 Conceptual representation of a simple Fabry–Perot laser.

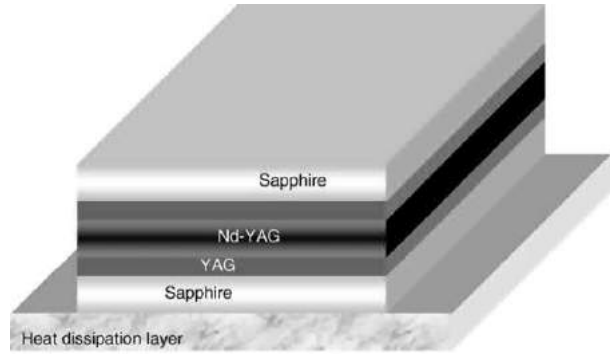

Fig. 3 End view of a planar waveguide form of a Nd-YAG laser made using sapphire (single crystal Al_2O_3) for cladding.

Planar Waveguide Lasers (PWL)

By way of an example, let us consider a solid state, rare-earth doped laser, such as an Nd-doped YAG laser. All rare-earth ions are able to fluoresce, some (e.g., Nd) more efficiently than others, and hence can be light emitters. These ions have to be dissolved (doped) into a solid host, say, a glass or a single crystal such as YAG (**Y**ttrium **A**luminum **G**arnet). The excitation of these ions is done optically with light of a suitably higher energy (or lower wavelength). When this 'pump' light is made incident on a rod of the rare earth doped material, it is absorbed by the rare earth ions and used to reach an excited electronic state. Amongst all the possible higher energy states, only certain ones are sustainable because the electron resides in that state for an acceptably long period of time (called the lifetime). Hence, population inversion can be more readily achieved in these longer lifetime states and stimulated processes such as amplification and lasing are readily possible. The laser light generated in the doped rod is emitted across its entire cross-section.

Planar waveguide lasers are designed to miniaturize the kinds of stand-alone solid state lasers describe above. In principle, a PWL is a laser where the light emitting resonant cavity is a planar waveguide. This topic of our interest is a relatively recent development that renders some unique advantages. These can be broadly classified into two categories: one relating to making efficient high-power lasers and the other, to the integration of lasers with other optical components.

An example of the first type is a Nd-YAG laser configured into a planar geometry so that the laser light is now contained within the waveguide. One design that has evolved to achieve this well is shown in Fig. 3. The light is guided by the waveguiding structure setup by the sapphire planes. Since both the excitation (or pump) and the lasing modes are well confined, high modal overlap is achieved. These PWLs can be edge-pumped or face-pumped and high powers (measured in the order of Watts) can be achieved.

In the above design, the cavity is defined by two parallel reflectors bracketing the light-generating medium and is known as a Fabry–Perot etalon. The distance between the two mirrors sets up a wavelength filtering mechanism whereby only resonant wavelengths are sustained. These are related to the intra-mirror distance by the simple equation shown below:

$$L_c = \frac{a\lambda}{2n} \tag{3}$$

where L_c is the length of the cavity, a is an integer, n is the refractive index of the medium, and λ is the wavelength in vacuum. As the cavity length increases, the number of possible resonant wavelengths increases and the spacing between wavelengths is

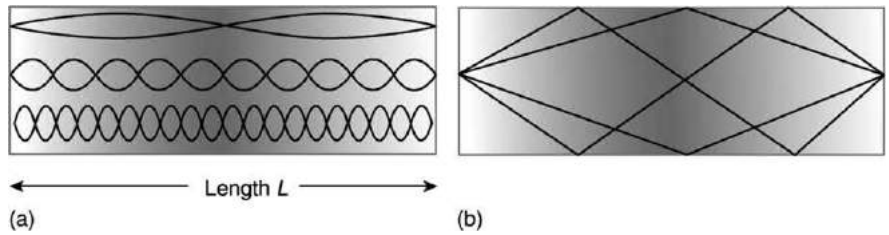

Fig. 4 Generation of longitudinal (a) and lateral modes (b) in a Fabry–Perot laser.

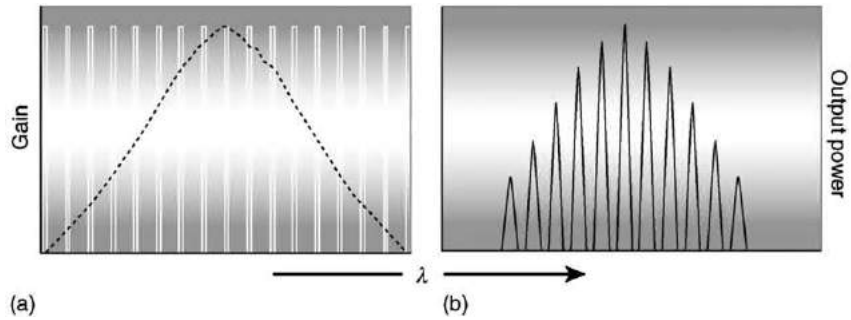

Fig. 5 The output of a Fabry–Perot laser: (b) is governed by the number of modes within the gain spectrum of the lasing medium shown by the dotted lines (a). The spacing between the modes can be altered changing the length of the cavity. The smallest mode in output spectrum (b) needs to have enough gain to overcome the lasing threshold.

reduced. These are called the longitudinal modes (Fig. 4(a)). Likewise, lateral modes (Fig. 4(b)) can be generated in the cavity as well, tending to make the operation of the laser multimoded, as shown in Fig. 5. This is characteristic of a Fabry–Perot laser, which can be overcome by embellishing the basic etalon design.

As stated above, another big advantage here is that the platform for fabricating PWLs can be chosen to be the same as to make PLCs. Waveguide comprizing the laser can be continued further and configured into other desirable components, such as splitters, couplers, etc. Thus the laser does not have to be separately coupled to the waveguide. This is a major advantage as this connecting process is time-consuming and incurs loss of light. In the field of photonics, the ability to integrate active components with passive ones onto the same chip is considered to be the first step toward photonic chips, analogous to the integrated circuit chip that has revolutionalized our world. This is the driver behind continued research and development of active elements such as PWLs that can be coupled with other actives, such as thin film electro-optic modulators or acousto-optic filters, etc.

Waveguide Laser Materials

The waveguiding structure is comprised of materials that are either dielectrics or semiconductors, both inorganic and organic. For their homogeneity and broad transmittance, high silica glasses are ideal materials for waveguiding. The chief advantages of high silica waveguides are easy index and mode matching to standard optical fibers, low propagation loss, and good durability. When crystalline materials are used, they have to be single-crystalline to avoid grain boundary scattering (for instance, Ti-diffused waveguide in single-crystal $LiNbO_3$). If semiconductors are used, the bandgap of the material needs to be sufficiently high so as to avoid absorption of the photons. Silicon waveguide (bandgap = 1.1 eV) on a silica cladding (SOI) is commonly used system in PLCs used for telecommunication applications.

The PWLs described above are fabricated in rare-earth doped glass waveguides. Waveguides have to be much shorter than fiber, so the challenge of incorporating the rare-earth ions at high concentration in the host material has to be overcome. Planar waveguide lasers are a compressed version fiber lasers that have the unique advantage that they can be directly integrated with other components on the same chip using very large-scale integration (VLSI)-style processing. The concentration of Er (or the lasing ion in general) in a PWL needs to be many times that in fiber amplifiers ($\sim$100 ppm). The need for high gain in a short length requires high pump powers and high lasing ion concentration, both of which lead to nonideal behavior. Since rare earth ions tend to have multiple excited electronic states, there can be absorption at the excited state level as well. When a higher energy state is separated from the upper emission level by the energy in one laser photon, the ions can undergo excited state absorption or ESA (Fig. 6). Rare-earth ions in close proximity tend to undergo cooperative upconversion processes, where two excited ions transfer energy between each other so that neither ion may fluoresce at the desired wavelength. These processes can negatively affect the useful gain of the system. Ensuring high spatial dispersion of the Er atoms is critical to minimize this effect. Special glass compositions, such as phosphate glasses, are used here to ensure that the Er is not phase separated or clustered. These glasses also

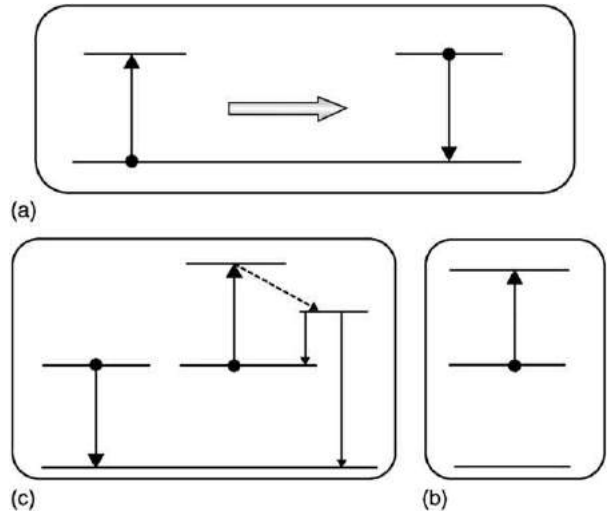

Fig. 6 The energy level diagram for a representative two level system. The inset (a) shows the excitation and emission process. Excited state absorption is shown in (b) and cooperative upconversion is shown in (c).

provide a high induced cross-section for gain. Glasses with low phonon energy are also desirable to minimize phonon-mediated nonradiative decays in the rare earths. Light emitting semiconductors and polymers containing fluorescent dyes have also been used to make PWLs.

Now that we have gained an overall perspective of PWLs, we can discuss different manifestations. The following discussion uses a Er doped system as a model since there is a substantial body of work on this system.

Distributed Bragg Reflector (DBR) PWL

In planar waveguide systems, it is difficult to construct efficient mirrors or reflective facets perpendicular to the waveguide. However, mirrors can be replaced by Bragg gratings, which be fabricated relatively easily. In-line Bragg gratings result in wavelength specific reflection as per **Eq. (4)**:

$$\Lambda = \frac{m\lambda}{2\bar{n}} \tag{4}$$

where Λ is the grating pitch, m is the order of the grating, and λ is the central Bragg wavelength. Gratings have another important benefit in that they ensure that unwanted frequencies outside the limited reflection spectrum are dispersed before they reach the lasing threshold.

A DBR laser is similar in principle to a Fabry–Perot design but uses a Bragg grating on either end of planar Er-doped waveguide. This is shown schematically in **Fig. 7**. Gratings can be made in a number of different ways. One popular way is to exploit the photosensitivity of glasses using 193 nm or 244 nm UV light to periodically alter the refractive index of a glass waveguide. Photosensitivity occurs by a combination of molecular changes (such as the formation of oxygen deficiency sites or color centers) as well as changes in the material strain in the region exposed to the UV light (**Fig. 7(c)**). Alternatively, gratings can also be formed using etching to make periodic grooves in the waveguide. These are called corrugated or surface relief gratings (**Fig. 7(a)** and **6(b)**).

The back Grating 1 has a high reflectance of almost 100%, whereas Grating 2 is a weaker grating. The reflectance strength of Grating 2 is determined by taking into account the gain coefficient of the material and the output power required from the laser. The reflection intensity can be altered according to the following equation:

$$R = \tanh^2\left[\frac{\pi L_G \Delta n_{eff}\eta(V)}{\lambda_B}\right] \tag{5}$$

where L_G is the grating length, Δn_{eff} is the index modulation of the grating, λ_B is the Bragg wavelength, $\eta(V)$ is the confinement factor, a function of the waveguide parameter V that represents the fraction of the integrated mode intensity contained in the core. Increasing the reflectance strength also widens the linewidth of reflected light. The overlap of the spectral width between the back (Grating 1) and the front grating (Grating 2) becomes the effective output window within the gain spectrum of the lasing material (**Fig. 8**).

Not unexpectedly, a DBR structure generates longitudinal and lateral modes. These modes are, however, limited to the overlapping section of the reflection spectrum of the two gratings, as opposed to the entire gain spectrum as in Fabry–Perot lasers (**Fig. 9**). Yet, this can lead to problems such as spectral hole burning and mode hopping, whereby the dominant mode hops from one wavelength to another. The separation between the modes needs to be made sufficiently high compared to the spectral overlap

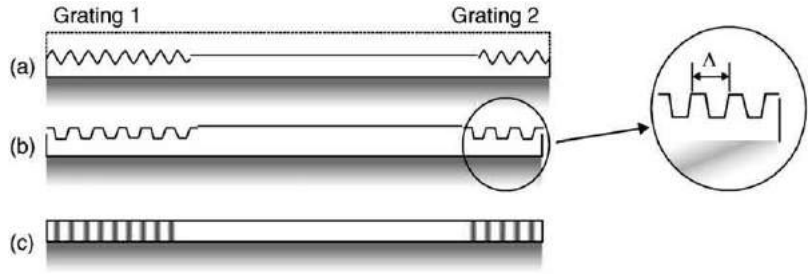

Fig. 7 DBR planar waveguide laser. The PWLs shown in (a) and (b) have sinusoidal and trapezoidal surface-relief gratings respectively etched in their cores. An in-line grating of the kind seen in (c) could be fabricated using the photosensitivity of the glass. The top cladding is shown only in (a).

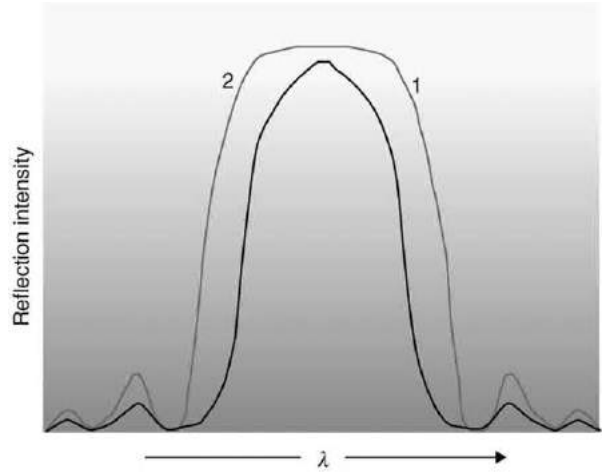

Fig. 8 The reflection spectrum of the back (1) and the front (2) gratings. As the intensity of the reflected light is made to increase, so does the linewidth. Hence, the narrower spectrum of partially reflecting grating 2 tends to dictate the operating wavelength range of a DBR.

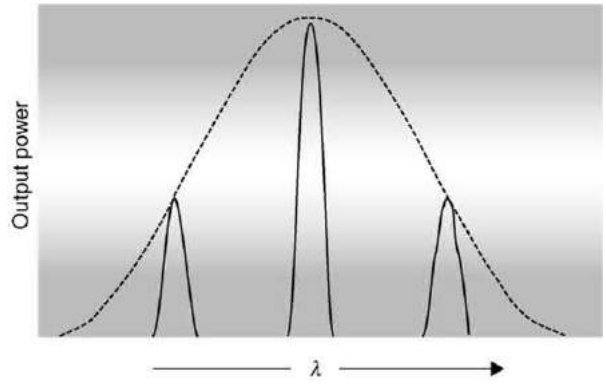

Fig. 9 A sample output from a DBR laser, shown in this case with 3 modes. The dotted line shows the effective overlap between the two gratings. The existence of the modes depends on several factors, in particular the pump intensity.

so that single-mode operation can be achieved. But in practice the length of the cavity (L_c in **Eq. (3)**) is determined by factors such as power and space and may not be independently altered to achieve this. However, the reflection spectrum of the back and the front gratings could be slightly shifted in opposite directions to realize a smaller overlap within which only one or a few modes may exist (**Fig. 10**). The occurrence of a single mode also depends on the gain profile of the lasing material and the optical pump power. In Er-doped PWLs, modes other than the primary one can be suppressed by using a pump power below a threshold limit and stable operation can be thus achieved.

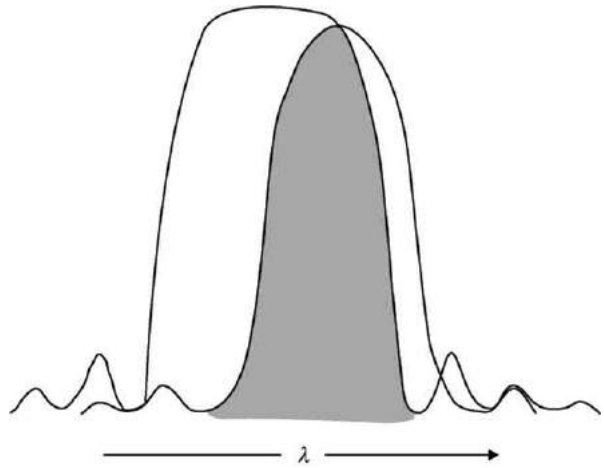

Fig. 10 Shifting the front and back gratings is one means of controlling the width of the overlap region, shown in gray. This way a single mode operation can be achieved even for a longer cavity with many modes.

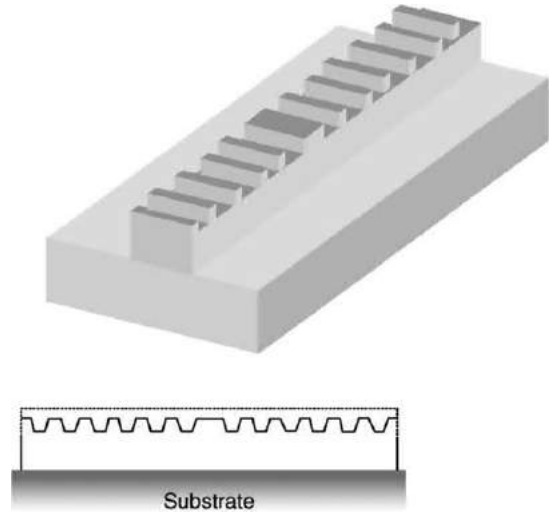

Fig. 11 A DFB planar waveguide laser.

Distributed Feedback (DFB) PWL

The modal problems listed above can be addressed by having a single, distributed grating that is placed along the entire length of the lasing waveguide itself (Fig. 11). Alternatively, the gratings may be also placed in the upper or the lower claddings to minimize loss of optical power due to scattering because of the longer grating. In this case, it is the evanescent portion of the light that interacts with the grating. In all cases, the periodicity of the grating leads to a condition for a single lasing wavelength based on constructive interference. The strongest mode occurs when the period Λ of the Bragg grating satisfies the primary Bragg condition (for first order grating or $m=1$) in Eq. (4). The device will also function if the grating pitch is equal to small integer multiples of $(\lambda/2\overline{n})$.

The grating needs to be made strong enough to generate sufficient feedback (reflections), which also widens the linewidth of the spectrum. To ensure a single narrow linewidth output, a quarter wave phase shift is introduced in the grating, as seen in Fig. 11. This phase shift creates a transmission fringe in the spectrum (Fig. 12) that significantly narrows the linewidth of the output signal. This design is considered to be standard for the DFB lasers of today. Fig. 13 shows a typical output of a DFB laser.

Any of the geometries seen in Fig. 1 can be used to make the waveguide for the laser. Due to processing constraints, rib (c) and ion diffused (d) configurations are commonly used. Waveguide losses, characteristically due to scattering from sidewall roughness, have to be low. Otherwise, this adds a significant overhead onto the pump power and lowers the output of the PWL. Birefringence caused by thermal expansion coefficient differences and excessive sidewall roughness also need to be minimized as they can substantially alter the loss of the TE mode compared to TM.

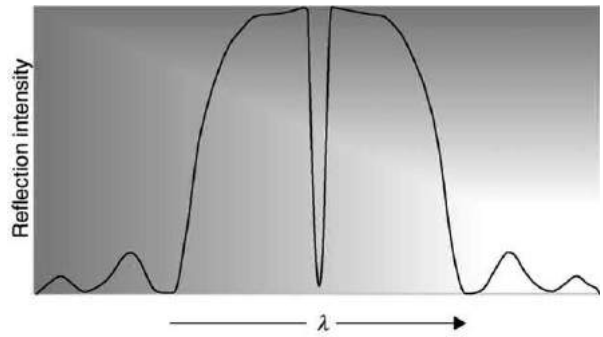

Fig. 12 The reflection spectrum of a grating with a central $\lambda/4$ shift in the pitch, showing the transmission fringe.

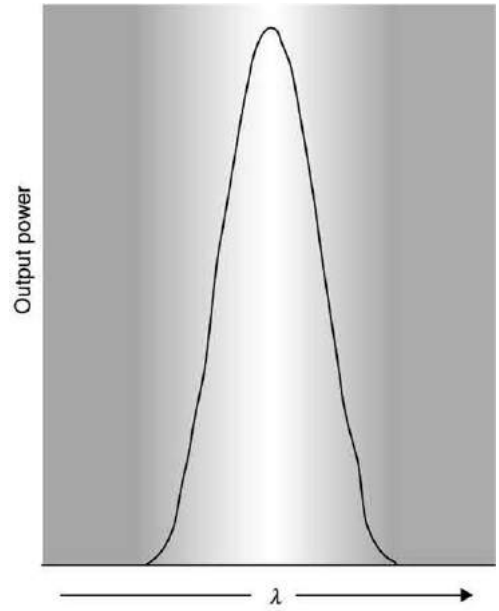

Fig. 13 The typical output of a DFB PWL. The linewidth can be made to be as narrow as a few hundred kHz.

Though their spectral properties are ideal as sources for long haul transmission, DFB lasers tend to be less powerful than the corresponding DBR designs. But the DBR PWLs are very sensitive to reflections that can cause a broadening of the laser linewidth or can act as a source of noise. In both cases, single mode lasers with very narrow linewidths in the kHz range with low radiance and noise, have been demonstrated. But neither is known to have been commercialized yet.

PWL Arrays

One of the unique advantages of the PWLs is that they are often processed using VLSI techniques. One important implication of this is the ability to make a PWL array where multiple lasers are formed simultaneously on the same chip, each with a unique output (Fig. 14). Such arrays are likely to be very useful in future dense wave division multiplexing (DWDM) optical network systems where as many as 128 channels (and more) are planned. Since packaging of individual lasers can be up to 75% of the total cost, making chips with multiple lasers is very attractive. The lasers may be DBRs or DFBs whose output wavelength is controlled by changing the grating parameters. To vary the individual wavelengths either the effective index and/or the pitch of the grating can be changed. Thus, any output spectrum can be designed and executed merely by making the appropriate lithography mask.

The gratings are often written holographically, where the wafer is exposed to an interference pattern caused by two coherent light beams. The wafer underneath can be sensitized with a photosensitive glass layer or a film a photoresist. It is not efficient to do multiple holographic exposures to individually change each of the gratings in the array. One way of doing this with a single holographic exposure is by constructing the waveguides at different relative angles. Then the component of the index modulation, resolved along the axis of the waveguide, would vary for each waveguide depending on the relative waveguide angle (Fig. 15); this would lead to differences in $\overline{n}$. Alternatively, the waveguide array can be made of the same height but of different widths. This is

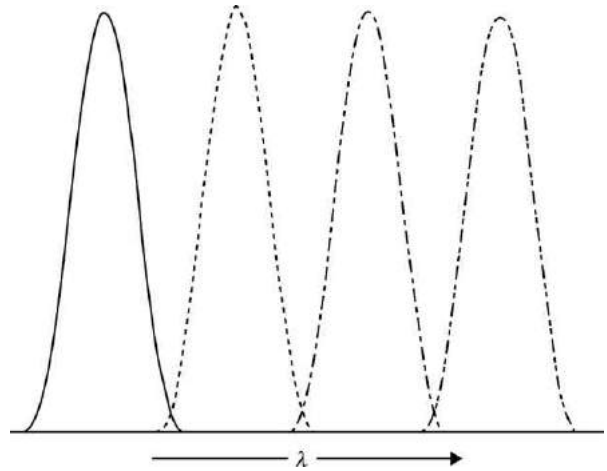

Fig. 14 Idealized output from a 4-PWL array.

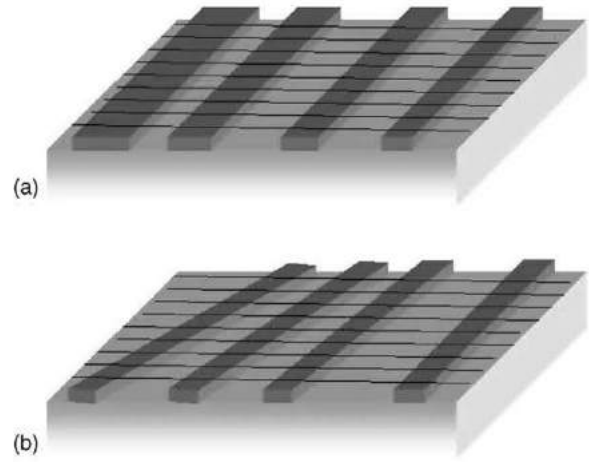

Fig. 15 Two different means of making planar waveguide laser array with a single holographic exposure to write the gratings. The darker rectangular blocks represent the core channels or ribs. In (a) the width and hence the effective index of the waveguide is varied whereas the relative angle between the waveguides is altered in (b). The lines represent the direction of the gratings, which is orthogonal only to the left most waveguide in (b).

easily done since waveguide height is usually deposited uniformly but the width is a function of the mask layout. Here the waveguide and the grating are orthogonal to each other and the variation in wavelengths is achieved by the fact that the effective propagation constant for each of the waveguides is different.

Arrays of PWLs have been made on a silicon substrate using deposition of an Er containing glass that is subsequently patterned into waveguide, and ultimately laser, arrays. Another approach has been to start with a phosphate glass as the host and the substrate. A phosphate glass is a superior host for rare-earth ions such as Er since the sensitization efficiency is nearly unity and much high doping levels are possible before concentration quenching effects set in. Waveguides are created by ion-exchange by immersion in a hot bath containing suitable ions (Na for Li, for instance). When starting with a uniformly doped glass, one issue in the assimilation of the laser array with other components is the fact that Er-doped region does not start and end with the waveguide laser section. Recently, the Er has been deposited selectively in the laser waveguide in phosphate glass using methods such as ion implantation, bringing the concept of integration closer to reality.

Pumping

One of the key issues here is the optical pumping of the rare-earth PWLs to generate acceptably high output power. Individual diode lasers or an array of diode lasers all on a single chip (called a diode bar) have been utilized as pumps. Both single and multimode diode lasers are available and have been used. There are pros and cons for either choice: whereas single mode lasers are less powerful (in the range of few 100 mW), it is difficult to convert the multimoded output of their counterparts to useful power.

The choice is determined by the waveguide design, which needs to be optimized for high overlap of the lasing mode with the pump. Carefully tapered waveguides can be used distribute the pump power from a diode bar to the various lasers in an array without significant loss.

PWL Stability

Stable operation of a PWL is one of the big challenges. In DBR PWLs, it is seen that there is a threshold for pump power only below which single moded operation can be achieved. Additionally, any PWL die needs to be made immune to ambient temperature changes. These changes can affect many of the critical output parameters of the PWL: wavelength, linewidth, power and in the case of an array, the wavelength separation. This occurs primarily due to the thermal shift in the grating pitch as well as the thermo-optic effect. It is overcome by supplanting the laser die onto a heat spreader chip made from conducting material such as Cu or BeO, followed by a thermoelectric cooler. Ironically, the laser's output wavelength and linewidth are fine-tuned and 'fixed' using temperature to overcome any drifts due to processing errors or other reasons.

Reliability is another important issue that needs to be addressed, especially for telecommunication applications where these types of components have to pass high quality standards such as the Telecordia qualification in the USA.

Other PWL Systems

Many of the PWLs to date have been developed using Er doping, which emits light in the range of interest for telecommunications (the C-band and even the L-band). Here, Yb can be added as a co-dopant to improve power conversion efficiency. Other fluorescing ions have also been used: Nd, Tm, Ho, and Cr. Several research teams have successfully demonstrated Nd-YAG PWLs, in particular. Ridge waveguide lasers have been made using semiconductor materials InGaAs/InP heterojunctions in both Fabry–Perot and DFB configurations. These systems are electrically pumped. DFB PWLs have also been made using dye doped polymers.

It must be mentioned that there are other ways to use waveguides to make lasers. The primary example is a configuration where the waveguide resonator is external to the active region. Designs, such as ring resonator lasers and an external cavity lasers, fall in this category.

Future research and development of PWLs would undoubtedly utilize photonic crystal structures. These structures have a periodic fluctuation in the index of refraction in two or three dimensions. Owing to the resultant Bragg reflections, specific allowed and forbidden states are created which can be used to control the confinement of light. Hitherto, these structures have been used to form mirrors or hexagonal ring resonators. These more sophisticated laser designs are likely to enable new functions such as tunability of the output spectrum.

Summary

Planar waveguide lasers are a special class of laser where light is confined to a waveguide. They have distinctive advantages that include high optical gains, low laser thresholds, narrow linewidths in the kHz range, and optimal thermal management. Significantly, they afford a platform that can be readily expanded to include an array of lasers, each with an unique output, and can ultimately be combined with multiple optical components on the same chip. Such compact integrated devices are bound to take photonics to a new high level of capability.

Further Reading

Carroll, J.E., Whiteaway, J.E.A., Plumb, R.G.S., Plumb, D., 1998. Distributed Feedback Semiconductor Lasers. London: IEE Publishing.
Ghafouri-Shiraz, H., 2003. Principles of Semiconductor Laser Diodes and Amplifiers: Analysis and Transmission Line Laser. London: Imperial College.
Hecht, J., 2001. Understanding Lasers: An Entry Level Guide. New York: Wiley Interscience.
Lee, J.R., Baker, H.J., Friel, J.G., Hilton, G.J., Hall, D.R., 2002. High-average-power Nd:YAG planar waveguide laser, face-pumped by 10 laser diode bars. Optical Letters 27, 524–526.
Milonni PW and Eberly JH (1988) *Lasers*. New York: John Wiley and Sons.
Saleh, B.E.A., Teich, M.C., 1991. Fundamentals of Photonics. New York: John Wiley & Sons.
Shepherd, D.P., Hettrick, S.J., Li, C., *et al*., 2001. High-power planar dielectric waveguide lasers. Journal of Physics D: Applied Physics 34, 2420–2432.
Siegman, A.E., 1986. Lasers. Harendon, Virginia: University Science Books.
Svelto, O. (Ed.), 1998. The Principles of Lasers. Dordrecht, The Netherlands: Kluwer Academic Publishers.
Veasey, D.L., Funk, D.S., Sanford, A., 1999. Arrays of distributed-Bragg-reflector waveguide lasers at 1536 nm in Yb/Er codoped phosphate glass. Applied Physics Letters 74 (6), 789–791.
Young, M., 2001. Optics and Lasers: Including Fibers and Optical Waveguides. NewYork: Springer-Verlag New York, Inc.

Up-Conversion Lasers

A Brenier, University of Lyon, Villeurbanne, France

Introduction

Under photoluminescence excitation, a luminescent center located inside a transparent material usually emits at a longer wavelength than the excitation. This is because of the existence of de-excitation losses inside the material. Thus, efficient down-conversion lasers can be built, for example, the commercial Nd^{3+}:$Y_3Al_5O_{12}$ laser which works at 1064 nm, under usual pumping near 800 nm. More surprisingly, it has been demonstrated that fluorescence can also be emitted at shorter wavelength than the excitation. This is possible if the losses are reduced and overcome by so-called 'up conversion mechanisms'. Generally speaking, a laser emitting at a shorter wavelength than the excitation is classified as an up-conversion laser.

These lasers are able to generate coherent light in the red, green, blue, and uv spectral ranges. A point of practical importance is that some of them can be pumped in the near-infrared range (around 800 nm or 980 nm) by commercially available laser diodes. They can be based on all-solid-state technology and share its advantages: compactness of devices, low maintenance, and efficiency. They are needed for a variety of applications: color displays, high density optical data storage, laser printing, biotechnology, submarine communications, gas-laser replacement, etc.

Because the emitted photons have higher energy than the excitation ones, more than one pump photon is needed to generate a single laser photon, so the physical processes involved in the up-conversion laser are essentially nonlinear. Two kinds of laser devices can be considered. In the first one, the excitation of the luminescent center up-migrates through the electronic levels and reaches the initial laser level. This step is due to some luminescence mechanisms. Then, in a second step, lasing at a short wavelength occurs through a transition towards the final laser level. In the second kind of device, the pump excitation activates a down-conversion lasing, but the laser wave is maintained inside the cavity, cannot escape, and is up-converted in short wavelengths by the second-order nonlinear susceptibility of the crystal host. This requires bifunctional nonlinear optical and laser crystals. The device is thus a self-frequency doubling laser or a self-sum frequency mixing laser. If the nonlinearity is not an intrinsic property of the laser crystal but is due to a second nonlinear optical crystal located inside or outside the cavity, the device operates intracavity or extracavity up conversion.

Up-Conversion from Luminescence Mechanisms

The numerous $^{2S+1}L_J$ energy levels ($4f^n$ configuration) of the trivalent rare earth ions, inserted in crystals or glasses, offer many possibilities for up conversion, particularly from long-lived intermediate levels that can be populated with infrared pumping. Crystal field induced transitions between them are comprised of sharp lines, because the $4f^n$ electrons are protected from external electric fields by the $5s^2p^6$ electrons. Their fluorescence spectra can exhibit cross-sections high enough (10^{-19} cm^2) to lead to laser operation in the visible range. For up-conversion purposes, the most popular ions are Er^{3+}, Pr^{3+}, Tm^{3+}, Ho^{3+}, and Nd^{3+}. Up conversion in Er^{3+} was used for the first time in 1959, to detect infrared radiation.

Energy Transfers

Two ions in close proximity interact electrostatically (Coulomb interaction) and can exchange energy. The ion which gives energy is called the sensitizer S or donor, and the energy receiving ion is the activator A or acceptor. Two examples discussed below are shown in **Fig. 1**. The most typical case of such nonradiative energy transfer in trivalent rare earth ions in insulating materials is due to electrical dipole–dipole interaction. For the latter, the transition probability $W(s^{-1})$ takes the form:

$$W = \frac{\hbar^4 c^4}{4\pi n^4} \frac{1}{R^6} \frac{Q_A}{\tau_{rad}} \int \frac{f_A(E) f_S(E)}{E^4} dE \qquad (1)$$

where R is the S–A distance, E is the photon energy of the transition operated by each ion, n the refractive index, and Q_A is the integrated absorption cross-section of the A-transition:

$$Q_A = \int \sigma_A(E) dE, \quad f_A = \frac{\sigma_A}{Q_A} \qquad (2)$$

In **Eq. (1)**, τ_{rad} is the radiative emission probability of S for the involved transition calculated from the emission probability A_S:

$$\frac{1}{\tau_{rad}} = \int A_S(E) dE, \quad f_S = \tau_{rad} A_S \qquad (3)$$

The integral in **Eq. (1)** is called the overlap integral and shows that the transfer is efficient if the luminescence spectrum of S is resonant with the absorption spectrum of A (corresponding to the S and A transitions involved in the nonradiative transfer).

 doi:10.1016/B978-0-12-809283-5.00854-5

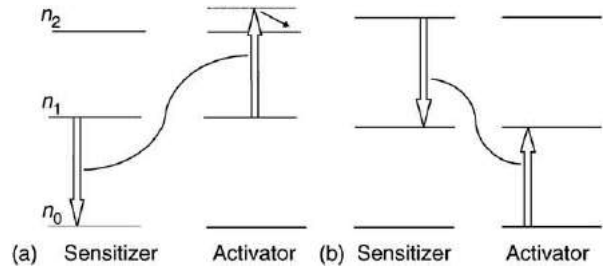

Fig. 1 (a) Up-conversion cross-relaxation energy transfer. (b) Down-conversion cross-relaxation energy transfer. The example in (a) is phonon assisted, the example in (b) is resonant.

The transitions are often not in perfect resonance but remain efficient enough to be used in practice. Then the energy mismatch is compensated by phonons, and the energy transfer is said to be phonon-assisted.

The energy transfer visualized in **Fig. 1(a)** is an up-conversion cross-relaxation. It is often met in the rare earth ions cited above when they are sufficiently concentrated in the material. The S and A ions can be of the same or of different chemical species. Successive energy transfers can promote the activator from its ground state towards higher energy levels, up to the one from which lasing can occur. This is the so-called 'Addition de Photons par Transferts d'Energie', sometimes appointed 'APTE'. Yb^{3+} is the most used ion as sensitizer and leads to the following transfers in examples of Er^{3+} or Ho^{3+} co-doping:

$$\begin{aligned} {}^2F_{5/2}(Yb) + {}^4I_{15/2}(Er) &\rightarrow {}^2F_{7/2}(Yb) + {}^4I_{11/2}(Er) \\ {}^2F_{5/2}(Yb) + {}^4I_{11/2}(Er) &\rightarrow {}^2F_{7/2}(Yb) + {}^4I_{7/2}(Er) \end{aligned} \tag{4}$$

$$\begin{aligned} {}^2F_{5/2}(Yb) + {}^5I_8(Ho) &\rightarrow {}^2F_{7/2}(Yb) + {}^5I_6(Ho) \\ {}^2F_{5/2}(Yb) + {}^5I_6(Ho) &\rightarrow {}^2F_{7/2}(Yb) + {}^5S_2(Ho) \end{aligned} \tag{5}$$

If several sensitizers give their energy simultaneously to a single activator, promoting it directly from its ground state towards a high energy level without intermediate levels, the sensitization is said to be 'cooperative'.

The energy transfer shown in **Fig. 1(b)** is a down-conversion cross-relaxation. It is the inverse of the previous one. At first sight it is, of course, undesirable in an up-conversion laser. However, we notice that one ion in the upper level will lead to two ions in the intermediate level. This fact can be exploited in special conditions (see below) and be beneficial for up conversion.

The dynamical study of the energy transfers, starting from the microscopic point of view contained in **Eq. (1)**, is very complicated because of the random locations of the ions or because of other processes, such as energy migration among sensitizers and back transfers. A popular method leading to useful predictions in practice is the rate equation analysis dealing with average parameters. The time evolution of the population densities n_i of the various levels i are described by first-order coupled equations, solved with adequate initial conditions and including the pumping rate. As an example, the rate equations for levels 1 and 2 in **Fig. 1(a)** take the form:

$$\frac{dn_1}{dt} = -\frac{n_1}{\tau_1} - 2kn_1^2 + \frac{\beta_{21}}{\tau_2} n_2 + W \tag{6}$$

$$\frac{dn_2}{dt} = -\frac{n_2}{\tau_2} + kn_1^2 \tag{7}$$

$$n_0 + n_1 + n_2 = 1 \tag{8}$$

where $\tau_{1,2}$ are the lifetimes of levels 1 and 2, k is the up-conversion transfer rate, β_{21} is the branching ratio for the $2 \rightarrow 1$ transition, and W is the pump rate (not shown in **Fig. 1**).

Sequential Two-Photon Absorption

Pumping the ground state of a luminescent ion (**Fig. 2(a)**) with a wave at angular frequency ω, leads to population with density n_1 of an excited state in a first step. Because the rare earth ions have numerous energy levels, it is possible for the pump to also be resonant with the transition from level 1 towards a higher energy level 2, and is further absorbed. Such a process leading to n_2 up-conversion population is a sequential two-photon absorption. If there is no $1 \rightarrow 2$ resonant transition for the ω frequency, it is possible to use a second pump wave at a different angular frequency ω' resonant with the $1 \rightarrow 2$ transition (**Fig. 2(b)**). Of course, in practice, it is advantageous when a single pump beam is required for up conversion.

Let us show the relevant terms of the rate equation analysis describing the process in **Fig. 2(a)**:

$$\frac{dn_1}{dt} = -\frac{n_1}{\tau_1} + I\sigma_0 n_0 - I\sigma_1 n_1 + \frac{\beta_{21}}{\tau_2} n_2 \tag{9}$$

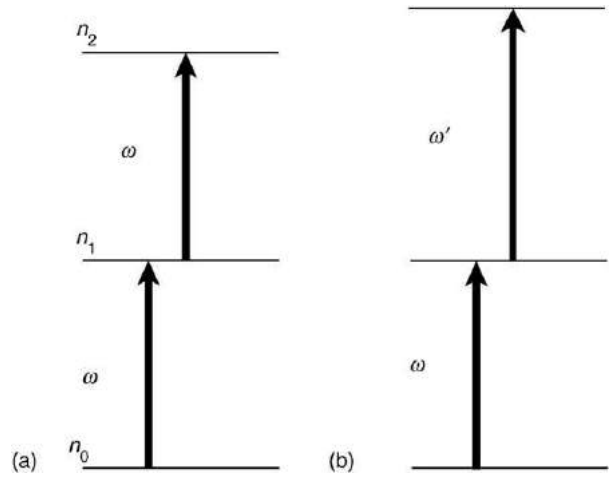

Fig. 2 Two-photon absorption up-conversion.

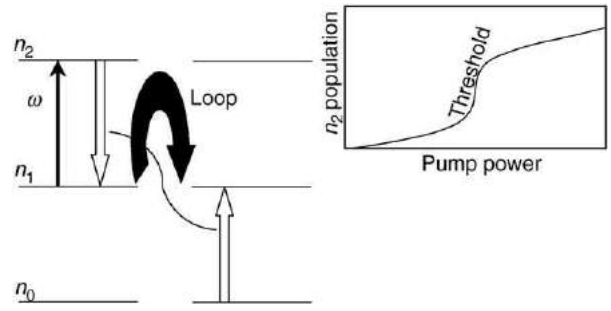

Fig. 3 Photon avalanche up-conversion.

$$\frac{dn_2}{dt} = -\frac{n_2}{\tau_2} + I\sigma_1 n_1 \tag{10}$$

$$n_0 + n_1 + n_2 = 1 \tag{11}$$

where I is the photon density of the pump (photons s^{-1} cm^{-2}) and σ_0 and σ_1 are the absorption cross-sections (cm^2) of the $0 \rightarrow 1$ and $1 \rightarrow 2$ transitions, respectively. The n_2 population density shows a nonlinear I dependence, most often close to I^2.

Photon Avalanche

In rare earth doped materials, an up-conversion emission, intense enough for visible lasing, can result from a more complex mechanism. In this case, absorption of the pump is not resonant with a transition from the ground state (weak ground-state absorption), but on the contrary, the photon pump energy matches the gap between two excited states (high excited state absorption). These levels are labeled 1 and 2 in Fig. 3. Level 1 generally has a rather long lifetime in order to accumulate population and level 2 is the initial up-conversion laser level. When the pump light is turned on, the populations of levels 1 and 2 first grow slowly, and then can accelerate exponentially. This is generally referred to as 'photon avalanche'. The term 'looping mechanism' also applies, because we can see in Fig. 3 that the excited state absorption $1 \rightarrow 2$ is followed by a cross-relaxation energy transfer (quantum efficiency up to two) that returns the system back to level 1. So after one loop, the n_1 population density has been amplified, and is further increased after successive loops. This pumping scheme was discovered in 1979, with $LaCl_3:Pr^{3+}$ crystal used as an infrared photon counter.

The rate equation analysis applied to the three-level system in Fig. 3 leads to the following time evolutions:

$$\frac{dn_1}{dt} = -\frac{n_1}{\tau_1} + I\sigma_0 n_0 - I\sigma_1 n_1 + \frac{\beta_{21}}{\tau_2} n_2 + 2kn_0 n_2 \tag{12}$$

$$\frac{dn_2}{dt} = -\frac{n_2}{\tau_2} + I\sigma_1 n_1 - kn_0 n_2 \tag{13}$$

$$n_0 + n_1 + n_2 = 1 \tag{14}$$

where the parameters have the same meaning as in Eqs. (6)–(11).

The steady state limit of the populations at infinite time, upon continuous excitation starting at $t=0$, can be obtained from canceling the time derivative in the left-hand side of the equation. Let us point out the interesting result. Two different dynamical regimes and analytical expressions for $n_2(\infty)$ are obtained. In the usual case of low pumping rate in the ground state and if the parameters of the systems satisfy: $k < \frac{1-\beta_{21}}{\tau_2}$, then the first regime is observed, whatever the pumping rate $I\sigma_1$ in the excited state.

If the parameters satisfy:

$$k > \frac{1-\beta_{21}}{\tau_2} \tag{15}$$

the first regime is still observed at low pumping rate $I\sigma_1$ in the excited state, more precisely at pumping rate low enough such that

$$I\sigma_1 < \frac{\left(k+\frac{1}{\tau_2}\right)\frac{1}{\tau_1}}{k-\frac{(1-b_{21})}{\tau_2}}$$

But interestingly, the second regime, the so-called photon avalanche, is obtained at high pumping rate $I\sigma_1$ in the excited state, that is to say if:

$$I\sigma_1 > \frac{\left(k+\frac{1}{\tau_2}\right)\frac{1}{\tau_1}}{k-\frac{(1-b_{21})}{\tau_2}} \tag{16}$$

So the right-hand side of **Eq. (16)** appears to be a pump threshold separating two dynamical regimes when the condition in **Eq. (15)** is satisfied, as the insert in **Fig. 3** shows.

Another way to distinguish between the two regimes is given by linearization of the system of **Eqs. (12)–(14)** ($n_0 \cong 1$). Then the description will only be valid close to $t=0$. We know from standard mathematical methods that the solutions of **Eqs. (12)–(14)** are then a combination of $\exp(\alpha t)$ terms. If the threshold condition in **Eq. (16)** is satisfied, a parameter α in the combination becomes positive, so the population of level 2 increases exponentially (photon avalanche) at early times and the dynamics of the system is unstable.

The physical meaning of **Eqs. (15)** and **(16)** is clarified by introducing the yield η_{CR} of the cross-relaxation process:

$$\eta_{CR} = \frac{k}{k+1/\tau_2} \tag{17}$$

and the yield η of the $2 \rightarrow 1$ de-excitation:

$$\eta = \frac{b_{21}/\tau_2}{k+1/\tau_2} \tag{18}$$

Then, the threshold condition in **Eq. (16)** can be written as:

$$\left(\frac{I\sigma_1}{I\sigma_1+1/\tau_1}\right)(2\eta_{CR}+\eta) > 1 \tag{19}$$

Starting with one ion in level 1, the first parenthesis in **Eq. (19)** is the yield with which it will be promoted to level 2, and the second parenthesis is the yield of the backward path $2 \rightarrow 1$. So, the left-hand side of **Eq. (19)** is the new number of ions after a looping path $1 \rightarrow 2 \rightarrow 1$, which has to be higher than 1 for the avalanche to occur. If we consider a lot of successive loops, the origin of the exponential behavior is clear. The role of the first ion in level 1 (and the weak pumping rate in the ground state $I\sigma_0$) is limited to initialize the process. Note that the value of the first parenthesis in **Eq. (19)** is less than 1, so the second parenthesis has to be higher than 1. This condition is equivalent to **Eq. (15)**.

Materials: The Energy Gap Law

As we can see from the above mechanisms, up conversion needs intermediate energy levels and is not efficient if these cannot accumulate populations due to a short lifetime. The spontaneous rate of de-excitation of an electronic level is the sum of a radiative part W^{rad} and of a nonradiative part W^{nrad}. The radiative part can be obtained through the absorption cross-section corresponding to the transition or from the Judd–Ofelt theory which also exploits the absorption spectrum. The nonradiative rate is then obtained by subtracting W^{rad} from the measured fluorescence decay rate of the level. Based on the measurements in many hosts for the different trivalent rare earth ions, at different temperatures, the nonradiative decay rate of a J-level towards the next lower J′-level was found to have the form:

$$W_{JJ'}^{nrad} = B\exp(-\beta\Delta E_{JJ'})\left(1-\exp\left(\frac{-\omega_{max}}{kT}\right)\right)^{-p} \tag{20}$$

where $\Delta E_{JJ'}$ is the energy gap between the two levels, $p=\Delta E_{JJ'}/\omega_{max}$, ω_{max} being the energy of an effective optical phonon of the host. The values of B, β and ω_{max} are found by fitting with experimental data. An example of such a fit is represented in **Fig. 4**.

Eq. (20) is the well-known and familiar 'energy gap law'. We have represented it for several oxide and nonoxide hosts at $T=300$ K in **Fig. 5**. It is clear that chloride, bromide, fluoride will be favorable hosts for up conversion due to their low phonon energy ($LaBr_3$: $\omega_{max}=175$ cm^{-1}, $LiYF_4$: $\omega_{max}=400$ cm^{-1}, $Y_3Al_5O_{12}$: $\omega_{max}=700$ cm^{-1}, silicate glass: $\omega_{max}=1100$ cm^{-1}). Let us mention also a fluorozirconate glass, ZBLAN, widely used in up-conversion fiber lasers.

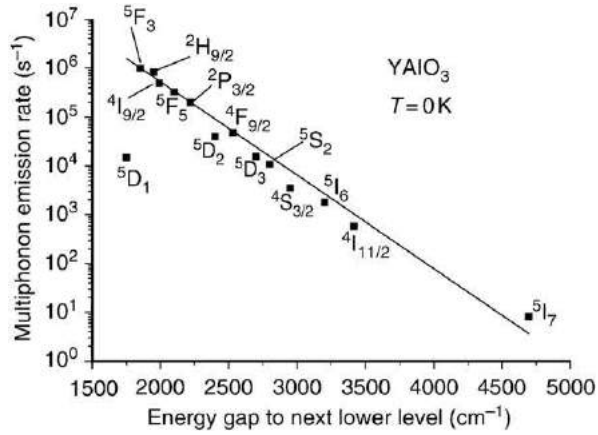

Fig. 4 Energy-gap dependence of nonradiative decay rates in $YAlO_3$.

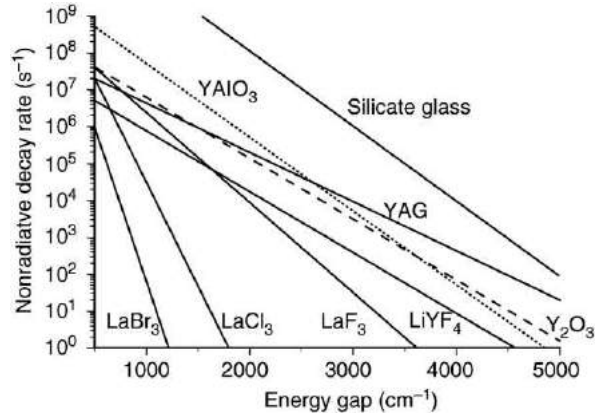

Fig. 5 Energy-gap law for different hosts at 300 K.

Table 1 Relevant terms for frequency up-conversion

Optical process	*Term*	*New wave*
Linear		
Refractive index, absorption, stimulated emission	$\chi^{(1)}(-\omega;\omega)$	
Second order nonlinear		
Second harmonic generation	$\chi^{(2)}(-\omega_3;\omega_1,\omega_1)$	$\omega_3=2\omega_1$
Sum-frequency mixing	$\chi^{(2)}(-\omega_3;\omega_1,\pm\omega_2)$	$\omega_3=\omega_1+\omega_2$

Up-Conversion from Second-Order Optical Nonlinearity

The inter-atomic electric field acting on the electrons inside a medium has a magnitude of $\sim 10^8$ V/cm and derives from a potential that is anharmonic. The electric field **E** of an electromagnetic wave propagating through the medium drives the electrons beyond the quadratic region of the potential in the case of high **E**. So, the electron response and the associated polarization **P** take the form of a function of **E**:

$$\mathbf{P}(\mathbf{E}) = \mathcal{E}_0(\chi^{(1)} : \mathbf{E} + \chi^{(2)} : \mathbf{E}^2 + \chi^{(3)} : \mathbf{E}^3 + \cdots) \tag{21}$$

The different terms in Eq. (21) are responsible for many optical phenomena because in Maxwell electromagnetic theory, the polarization is the source of the waves. We have selected in Table 1 the terms relevant for linear effects and for the purpose of frequency up conversions: second-harmonic generation (SHG) and sum-frequency mixing, due to second-order terms. It should be noticed that the latter have a nonzero value only in noncentrosymmetric materials.

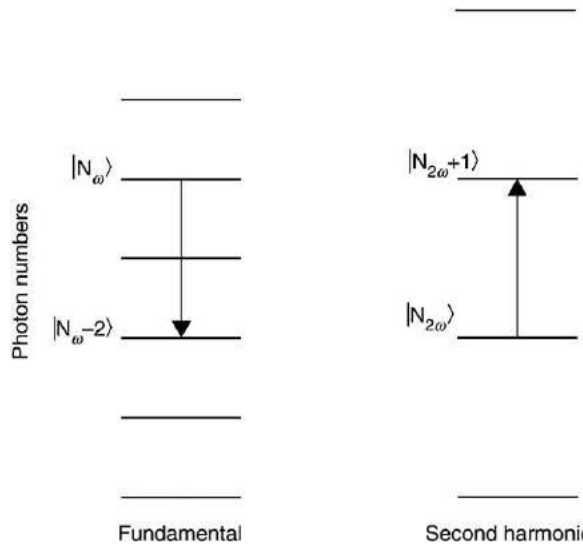

Fig. 6 Up-conversion from second harmonic generation.

Fig. 6 visualizes the up conversion from SHG: two photons of the fundamental field at angular frequency ω_1 are annihilated while one photon at twice the angular frequency is created. A similar picture could be drawn for sum-frequency up conversion: SHG is the degenerate case where $\omega_1 = \omega_2$.

The second-order nonlinear processes have an efficiency given by an effective coefficient d_{eff} given hereafter:

$$d_{\text{eff}} = \sum_{\text{ijk}} e_i(\omega_3) d_{\text{ijk}} e_j(\omega_1) e_k(\omega_2) \tag{22}$$

where $e_i(\omega_l)$ is the ith component of the unit vector of the electric field of the wave l at angular frequency ω_l and $d_{\text{ijk}} = (1/2)\chi_{\text{ijk}}$.

Moreover, the wave propagation equations in the slow varying envelope approximation, given hereafter for the E_3 electric field of the sum-frequency mixing wave:

$$\frac{\mathrm{d}E_3}{\mathrm{d}z} = \frac{i\omega_3 d_{\text{eff}}}{n_3 c} E_1 E_2 \quad \exp(\mathrm{i}(k_1 + k_2 - k_3)z) (\text{with } k_j = \omega_j n_j / c) \tag{23}$$

impose that the frequency conversion is efficient in practice, only if the following phase matching condition is satisfied:

$$\omega_1 n_1 \uparrow(\theta, \varphi) + \omega_2 n_2 \uparrow(\theta, \varphi) = \omega_3 n_3 \downarrow(\theta, \varphi) \tag{24}$$

where $n_i\uparrow$ and $n_i\downarrow$ are respectively the upper and lower refractive index in the direction of propagation (θ, ϕ) (n_i in Eq. (23) are the same with a simplified notation). Eq. (24) is restricted here for simplicity to collinear type I (waves 1 and 2 and the same polarization) and is the expression of photon momentum conservation. It is usually achieved from the birefringence of the crystals. When this is not possible, another technique can be used, namely quasi-phase matching, which involves reversing periodically the sign of the nonlinear optical coefficient.

In bifunctional crystals, the laser effect and the $\chi^{(2)}$ interaction occur simultaneously inside the same crystal. In the case of the self-frequency doubling laser, the angular frequency $\omega_1 = \omega_2$ in Eq. (24) is that of the fundamental laser wave. In the case of the self-sum frequency mixing laser, ω_1 corresponds to the fundamental laser wave and ω_2 to the pump wave. As we can see from Eqs. (22)–(24), the polarization of the laser emission is crucial in order for the device to work. So, the laser stimulated emission cross-section σ for propagation in the phase matching direction (θ, ϕ) in the adequate polarization has to be evaluated. This can be performed with the formula:

$$\sigma(\theta, \varphi) = \frac{\sigma_X \sigma_Y \sigma_Z}{\sqrt{\sigma_Y^2 \sigma_Z^2 e_X^2 + \sigma_X^2 \sigma_Z^2 e_Y^2 + \sigma_X^2 \sigma_Y^2 e_Z^2}} \tag{25}$$

where $\sigma_{X,\ Y,\ Z}$ are the emission cross-sections for X, Y, Z-polarizations and $e_{X,\ Y,\ Z}$ are the components of the unit vector or the electric field of the laser wave.

To summarize, the conditions necessary for a crystal to be a bi-functional material are: (i) it must be noncentrosymmetric; (ii) it must accept fluorescent doping (usually with Nd^{3+} or Yb^{3+}); and (iii) it must be phase matchable for its laser emission. In practice, only a few crystals satisfy these requirements and have been tried with some success. The main ones are $LiNbO_3$, $LaBGeO_5$, $Ba_2NaNb_5O_{15}$, β'-$Gd_2(MoO_4)_3$, $YAl_3(BO_3)_4$, $GdAl_3(BO_3)_4$, $CaY_4(BO_3)_3O$, and $CaGd_4(BO_3)_3O$.

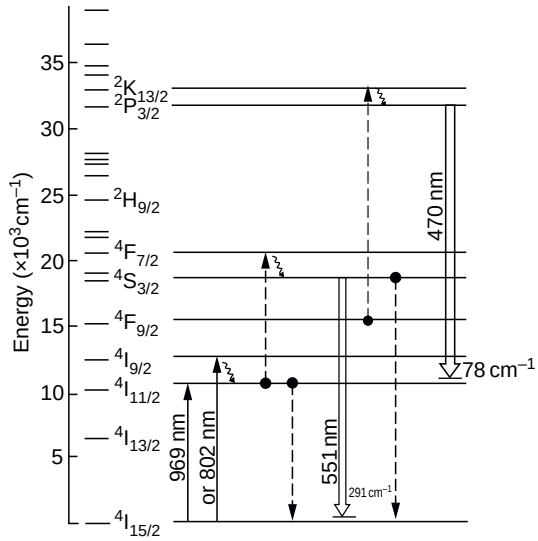

Fig. 7 Energy level diagram of $LiYF_4:Er^{3+}$. The dashed lines indicate the energy transfers responsible for the green and blue up-conversion lasing. Reproduced with permission of American Institute of Physics from Macfarlane RM, Tong F, Silversmith AJ and Lenth W (1988) Violet CW neodymium upconversion lasers. *Applied Physics Letters* 52(16): 1300.

Up-Conversion Lasers Based on Luminescence Mechanisms in Materials

Energy Transfers

Up-conversion lasers from energy transfers have the advantage of using a single pump laser. This latter populates firstly a long-lived intermediate level. A typical example, based on $LiYF_4:Er^{3+}$ crystal, is provided by Fig. 7.

The pump at 802 nm from a GaAlAs diode laser, or at 969 nm, populates efficiently at the 8 ms lifetime $^4I_{11/2}$ level. The green laser, at 551 nm, corresponds to the $^4S_{3/2} \rightarrow {}^4I_{15/2}$ transition and it can be shown that energy transfer is an essential part of the mechanism by stopping abruptly the pump laser: green lasing continues to occur for several hundred microseconds because the $^4S_{3/2}$ level is still fed (an excited-state absorption would stop immediately). The involved energy transfer is:

$$^4I_{11/2} + {}^4I_{11/2} \rightarrow {}^4F_{7/2} + {}^4I_{15/2} \tag{26}$$

and leads to a 20 mW laser threshold and 2.3 mW green output power at 50 K temperature upon 95 mW of incident power.

By using mirrors with high transmission in the green range but having high reflectivity at blue wavelengths, blue laser emission at 470 nm can be sustained, corresponding to the $^2P_{3/2} \rightarrow {}^4I_{11/2}$ transition. It originates from a mechanism involving three up-conversion energy transfers:

$$\begin{aligned} &^4I_{11/2} + {}^4I_{11/2} \rightarrow {}^4F_{7/2} + {}^4I_{15/2} \\ &^4I_{11/2} + {}^4I_{13/2} \rightarrow {}^4F_{9/2} + {}^4I_{15/2} \\ &^4S_{3/2} + {}^4F_{9/2} \rightarrow {}^4I_{15/2} + {}^2K_{13/2} \end{aligned} \tag{27}$$

The illustration in Fig. 7 shows up-conversion energy transfers between ions of the same chemical species, but co-doping the laser host with two ions of different chemical species provides the opportunity of separating their roles: one species is devoted to pump absorption (this is the sensitizer) and the other one is devoted to lasing (this is the activator). Due to the development of efficient semiconductor InGaAs diodes emitting near 980 nm, the most widely used sensitizer is Yb^{3+}. Table 2 provides a list of several up-conversion lasers based on energy transfer mechanisms, operating at room temperature. We can see that among the most efficient lasers, are those based on rare-earth ions doped glass fibers. The reason is that in fibers, the pump and laser waves remain confined inside a core (typically 5 μm diameter) and have high energy densities over a long length (tenths of cms), which is favorable for all up-conversion mechanisms.

Sequential Two-Photon Absorption

An example of an up-conversion laser pumped by a two-photon absorption mechanism is represented in Fig. 8. The material is $LaF_3:Nd^{3+}$ and violet lasing at 380 nm corresponds to the $^4D_{3/2} \rightarrow {}^4I_{11/2}$ transition.

Transition from the ground state up to the $^4F_{5/2}$ level is firstly obtained from an infrared beam at 790 nm in order to feed the $^4F_{3/2}$ intermediate level after a fast nonradiative de-excitation. The $^4F_{3/2}$ level has a rather long lifetime: 700 μs, so it accumulates

Table 2 Main room temperature up-conversion lasers operating from energy transfers

Laser material	*Laser wavelength (nm)*	*Pump wavelength (nm)*	*Output power*
$BaY_{1.4}Yb_{0.59}Ho_{0.01}F_8$	670	1540 + 1054	
$BaYYb_{0.998}Tm_{0.002}F_8$	649	1054	1%
$BaYYb_{0.99}Tm_{0.01}F_8$	799–649–510–455	960	
$BaYYb_{0.99}Tm_{0.01}F_8$	348	960	
$LiYF_4$:Er(1%):Yb(3%)	551	966	37 mW
$LiY_{0.89}Yb_{0.1}Tm_{0.01}F_4$	810–792	969	80 mW
$LiY_{0.89}Yb_{0.1}Tm_{0.01}F_4$	650	969	5 mW
$LiKYF_5$:Er (1%)	550	808	150 mW
Fiber:Yb:Pr	635	849	20 mW
Fiber:Yb:Pr	635	1016	6.2 mW
Fiber:Yb:Pr	521	833	0.7 mW
Fiber:Yb:Pr	635	860	4 mW

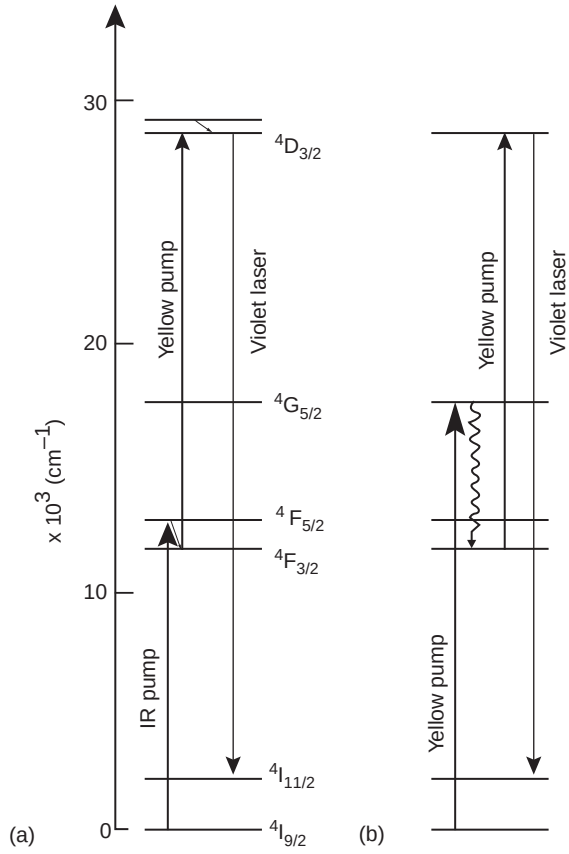

Fig. 8 Simplified energy level diagram of LaF_3:Nd^{3+} and up-conversion lasing from two-photon absorption. (a) Two different pump wavelengths, (b) doubly resonant single pump wavelength. Reproduced with permission of American Institute of Physics from Lenth W, Silversmith AJ and Macfarlane RM (1988) Green infrared pumped erbium up conversion lasers. In: TAM AC, Gole JL and Stwalley WC (eds) *Advances in Laser Science – III.* AIP Conference Proceedings n°172: 8–12.

enough population that a second yellow pumping at 591 nm (**Fig. 8(a)**) can efficiently populate the $^4D_{3/2}$ initial laser level. With 110 mW of infrared pump and 300 mW of yellow pump, 12 mW of violet output were obtained at 20 K temperature.

A simpler device is obtained if a single-pump beam is used. This is the case in **Fig. 8(b)**, where a yellow pump at 578 nm provides both ground state and excited state absorption. This doubly resonant scheme is less efficient in LaF_3:Nd^{3+} than the previous one but in **Fig. 9**, we show another example: $LiYF_4$:Er^{3+}, working at room temperature. The lasing $^4S_{3/2} \rightarrow {}^4I_{15/2}$ transition at 551 nm delivers 20 mW upon 1600 mW pumping at 974 nm.

The simplified Er^{3+} level scheme of **Fig. 9** contains four main levels with population densities: $n_0 = [^4I_{15/2}]$, $n_1 = [^4I_{13/2}]$, $n_2 = [^4I_{11/2}]$, $n_3 = [^4S_{3/2}]$. The rate equations of populations of this system can be written and solved because all the implied parameters (lifetimes, branching ratios, ground, and excited states absorption cross-sections) have been measured in this material.

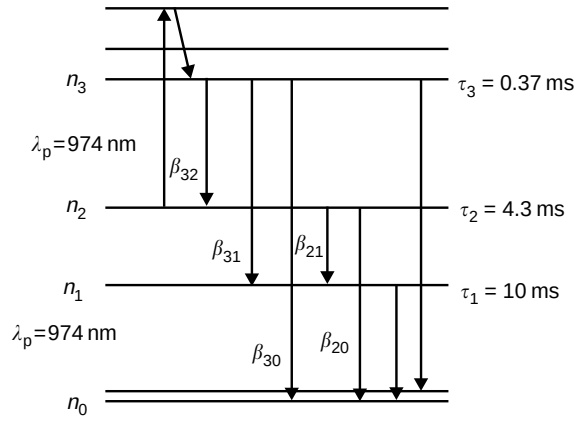

Fig. 9 Simplified energy level diagram of $LiYF_4:Er^{3+}$ and up-conversion lasing from two-photon absorption (doubly resonant). Reproduced with permission of Institute of Physics Publishing from Huber G (1999) Visible cw solid-state lasers. *Advances in Lasers and Applications*, Bristol and Philadelphia: Scottish Universities Summer School in Physics & Institute of Physics Publishing, 19.

Table 3 Main room temperature up-conversion lasers operating from two-photon absorption

Laser material	*Laser wavelength (nm)*	*Pump wavelength (nm)*	*Output power*
$Y_3Al_5O_{12}$:Er 1%	561	647 + 810	
$LiYF_4$:Er 1%	551	647 + 810	0.95 mJ
$LiYF_4$:Er 1%	551	810	40 mW
$LiYF_4$:Er 1%	551	974	45 mW
$LiYF_4$:Er	551	974	20 mW
KYF_4:Er 1%	562	647 + 810	0.95 mJ
$LiLuF_4$:Er	552	974	70 mW
$LiLuF_4$:Er	552	970	213 mW
$LiYF_4$:Tm 1%	453–450	781 + 649	0.2 mJ
Fiber:Tm	480	1120	57 mW
Fiber:Tm	480	1100–1180	45 mW
Fiber:Tm	455	645 + 1064	3 mW
Fiber:Tm	803–816	1064	1.2 W
Fiber:Tm	482	1123	120 mW
Fiber:Yb:Tm	650	1120	
Fiber:Ho	540–553	643	38 mW
Fiber:Ho	544–549	645	20 mW
Fiber:Er	548	800	15 mW
Fiber:Er	546	801	23 mW
Fiber:Er	544	971	12 mW
Fiber:Er	543	800	
Fiber:Pr	635	1010 + 835	180 mW
Fiber:Pr	605	1010 + 835	30 mW
Fiber:Pr	635	1020 + 840	54 mW
Fiber:Pr	520	1020 + 840	20 mW
Fiber:Pr	491	1020 + 840	7 mW
Fiber:Nd	381	590	74 μW
Fiber:Nd	412	590	500 μW

Predictive models of the up-conversion green laser are successful, matching experimental measurements, and confirm that the doubly resonant mechanism is realistic at low doping concentration.

Table 3 summarizes the performances of the main up-conversion lasers based on sequential photon absorption and working at room temperature.

Photon Avalanche

Photon avalanche up-conversion was observed mainly in Nd^{3+}, Tm^{3+}, Er^{3+}, and Pr^{3+} doped materials. As an illustration of this mechanism, let us consider the case of the continuous wave Pr^{3+}–Yb^{3+}-doped fluorozirconate fiber laser (3000 ppm Pr,

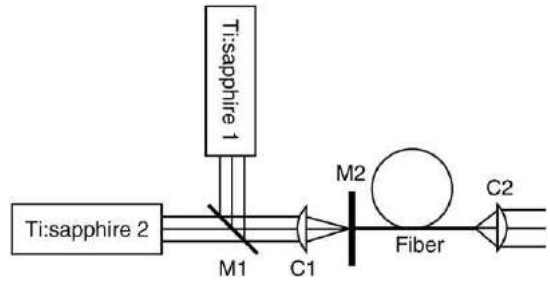

Fig. 10 Experimental set-up for Yb–Pr-doped fiber up- conversion laser. M1: dichroic mirror, C1, C2: lenses, M2 dichroic input mirror. Reproduced with permission of Optical Society of America from Sandrock T, Scheife H, Heumann E and Huber G (1997) High power continuous wave up-conversion fiber laser at room temperature. *Optics Letters* 22(11): 809.

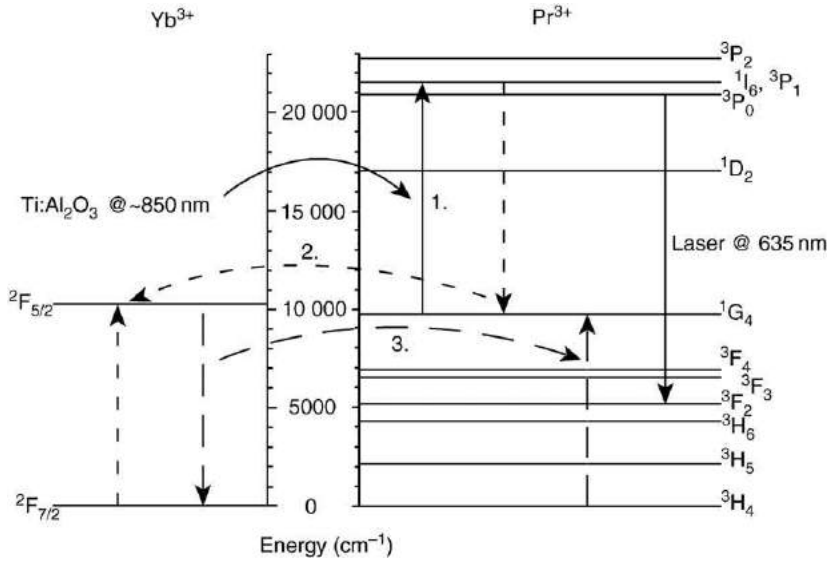

Fig. 11 Energy level diagram of fluorozirconate: Yb^{3+}:Pr^{3+} fiber. The dashed lines indicate the energy transfers. Reproduced with permission of the Optical Society of America from Sandrock T, Scheife H, Heumann E and Huber G (1997) High power continuous wave up-conversion fiber laser at room temperature. *Optics Letters* 22(11): 809.

20 000 ppm Yb). With two Ti:sapphire pump lasers operating at 852 and 826 nm (Fig. 10), an output power at 635 nm as high as 1.02 W was obtained with an incident pump power of 5.51 W (19% slope efficiency).

This remarkable result can be explained quantitatively by modeling the photon avalanche with a five-level diagram and restricted in Fig. 11 to a single laser pump at 850 nm. The laser emission corresponds to the $^3P_0 \rightarrow {}^3F_2$ transition. The five relevant levels have the population densities: $n_0 = [^3H_4(Pr)]$, $n_1 = [^1G_4(Pr)]$, $n_2 = [^3P_0(Pr)]$, $n_3 = [^2F_{7/2}(Yb)]$, $n_4 = [^2F_{5/2}(Yb)]$.

The two components of the mechanism are:

(i) the excited state absorption of the pump by the spin allowed transition: $^1G_4 \rightarrow {}^1I_6$

(ii) the process leading to the doubling of the $^1G_4$ population, in this instance through two energy transfers:

$$\begin{aligned} {}^1I_6(Pr) + {}^2F_{7/2}(Yb) &\rightarrow {}^1G_4(Pr) + {}^2F_{5/2}(Yb) \\ {}^2F_{5/2}(Yb) + {}^3H_4(Pr) &\rightarrow {}^2F_{7/2}(Yb) + {}^1G_4(Pr) \end{aligned} \tag{28}$$

The rate equation analysis predicts that the avalanche threshold occurs at about 1 W pump and the long fluorescence rise time is somewhat shortened at higher pump power, as observed experimentally.

Table 4 summarizes the performances of the main up-conversion lasers based on photon avalanche mechanisms.

Up-Conversion Lasers Based on Bi-Functional Crystals

The laser stimulated emission in a bi-functional crystal was first obtained in 1969 with Tm^{3+} in $LiNbO_3$ but the most used ion is, of course, Nd^{3+} working in the two channels:

$$^4F_{3/2} \rightarrow {}^4I_{11/2} \tag{29}$$

Table 4 Main room temperature up-conversion lasers operating from photon avalanche

Laser material	*Laser wavelength (nm)*	*Pump wavelength (nm)*	*Output power*
BaY_2F_8:Yb:Pr	607.5	822	55 mW
BaY_2F_8:Yb:Pr	638.7	841	26 mW
$LiY_{0.89}Yb_{0.1}Pr_{0.01}F_4$	720	830	1%
$LiY_{0.89}Yb_{0.1}Pr_{0.01}F_4$	639.5	830	
Fiber:Yb:Pr	635–637	780–880	300 mW
Fiber:Yb:Pr	605–622	780–880	44 mW
Fiber:Yb:Pr	517–540	780–820	20 mW
Fiber:Yb:Pr	491–493	780–880	4 mW
Fiber:Yb:Pr	635	850	1.02 W
Fiber:Yb:Pr	635	850	2.06 W
Fiber:Yb:Pr	520	850	0.3 W

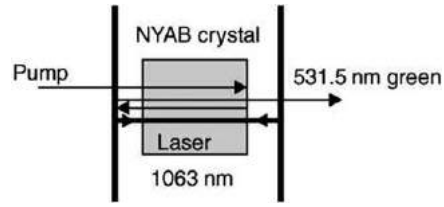

Fig. 12 Scheme of a self-frequency doubling laser. Note that in reality the three beams are superimposed.

$$^4F_{3/2} \rightarrow {}^4I_{13/2} \tag{30}$$

and more recently Yb^{3+} working in the channel:

$$^2F_{5/2} \rightarrow {}^2F_{7/2} \tag{31}$$

Channels in **Eqs. (29)** and **(31)** operate around 1060 nm wavelength and the channel in **Eq. (30)** near 1338 nm.

The Self-Frequency Doubling Laser

A typical laser scheme is shown in **Fig. 12**.

The laser beam cannot escape from the cavity and dichroic mirrors are used. The input mirror has high transmission at the pump wavelength and is highly reflective at laser- and second-harmonic wavelengths. The output mirror is highly reflective at laser wavelength and has high transmission for second harmonic. Channels in **Eqs. (29)** and **(31)** generate green light near 530 nm and the channel in **Eq. (30)** generates red light near 669 nm.

The most exploited bi-functional crystal is $YAl_3(BO_3)_4$:Nd^{3+}, well-known as NYAB. It is a negative uniaxial trigonal crystal (extraordinary index lower than the ordinary one) and its laser emission is easily observed in ordinary polarization with a high cross-section: 2×10^{-19} cm^2 at 1063 nm. Type I phase matching occurs at a polar angle $\theta = 30.7^\circ$ and 26.8° for channels in **Eqs. (29)** and **(30)** respectively. The effective nonlinear optical coefficient d_{eff} is close to 1.4 pm V^{-1} at azimuthal angle $\phi = 0^\circ$.

The Yb^{3+} ion has some advantages over the Nd^{3+} ion due to its very simple energy level scheme: there is no excited state absorption at the laser wavelength, no up-conversion losses, no concentration quenching, and no absorption in the green. The small Stokes shift between pump and laser emission reduces the thermal loading of the material during laser operation. A self-doubling laser based on $YAl_3(BO_3)_4$:Yb^{3+} of crystal has produced 1.1 W green power upon 11 W diode pumping.

The drawback of $YAl_3(BO_3)_4$ is that it is rather difficult to grow because it is not congruent. This difficulty was overcome by the discovery at Ecole Nationale Supérieure de Chimie de Paris of the rare-earth calcium oxoborate $CaGd_4(BO_3)_3O$ which can be grown to a large size by the Czokhralski method. It is a monoclinic biaxial crystal and doped with Nd^{3+}, it was firstly exploited (channel in **Eq. (29)**) in the *XY* principal plane at $\theta = 90^\circ$, $\phi = 46^\circ$ with ($d_{eff} = 0.5$ pm V^{-1}). It was soon recognized that the optimum phase matching direction ($d_{eff} = 1.68$ pm $\times V^{-1}$) occurred out of the principal planes, in the direction $\theta = 66.8^\circ$, $\phi = 132.6^\circ$, and high green power can be obtained: 225 mW under 1.56 W pump.

Table 5 summarizes the main self-frequency doubling results obtained with the channel in **Eq. (29)**.

The Self-Sum Frequency Mixing Laser

The example in **Fig. 13** gives the main features of such a laser. The input mirror has high transmission at the pump wavelength and both mirrors are highly reflective at laser wavelength. The output mirror has high transmission at sum frequency mixing wavelength.

Table 5 Main results of Nd^{3+} and Yb^{3+} green self-frequency doubling lasers near 530 nm (Eqs. (29) and (31))

Crystal	*Input power (mW)*	*Output power (mW)*	*Pumping*
$LiNbO_3$:MgO:Nd	215	1	Dye laser
$YAl_3(BO_3)_4$:Nd	870	10	Laser diode
$YAl_3(BO_3)_4$:Nd	280	3	Laser diode
$YAl_3(BO_3)_4$:Nd	400	69	Laser diode
$YAl_3(BO_3)_4$:Nd	1380	51	Laser diode
$YAl_3(BO_3)_4$:Nd	369	35	Laser diode
$LiNbO_3$:MgO:Nd	850	18	Dye laser
$LiNbO_3$:Sc_2O_3:Nd	65	0.14	Ti:sapphire laser
$LiNbO_3$:MgO:Nd	100	0.2	Laser diode
$YAl_3(BO_3)_4$:Nd	1600	225	Laser diode
$YAl_3(BO_3)_4$:Nd	2200	450	Ti:sapphire laser
$(Y,Lu)Al_3(BO_3)_4$:Nd	880	24	Laser diode
$LiNbO_3$:ZnO:Nd	430	0.65	Ti:sapphire laser
$Ba_2NaNb_5O_{15}$:Nd	270	46	Ti:sapphire laser
$CaY_4(BO_3)_3O$:Nd	900	62	Laser diode
$CaGd_4(BO_3)_3O$:Nd	1600	192	Ti:sapphire laser
$CaGd_4(BO_3)_3O$:Nd	1250	115	Laser diode
$CaGd_4(BO_3)_3O$:Nd	1560	225	Ti:sapphire laser
$YAl_3(BO_3)_4$:Yb	11 000	1100	Laser diode
$GdAl_3(BO_3)_4$:Nd	2.8 mJ/pulse	0.12 mJ/pulse	Pulsed dye laser

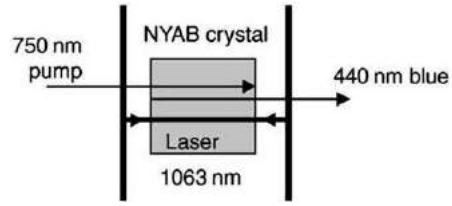

Fig. 13 Scheme of a self-sum frequency mixing laser. Note that in reality the three beams are superimposed.

Table 6 Main results of Nd^{3+} self-sum frequency mixing lasers based on Eq. (29)

Crystal	*Pump wavelength (nm)*	*Generated wavelength (nm)*	*Output power*
$YAl_3(BO_3)_4$:Nd	740–760	436–443	0.16 mJ/pulse
$GdAl_3(BO_3)_4$:Nd	740–760	436–443	0.403 mJ/pulse
$CaGd_4(BO_3)_3O$:Nd	811	465	1 mW
$YAl_3(BO_3)_4$:Nd	585–600	377–383	0.25 mJ/pulse
$GdAl_3(BO_3)_4$:Nd	587–597	378–382	0.105 mJ/pulse
$YAl_3(BO_3)_4$:Nd	488, 515	330, 380	0.2 mW

The laser wave at angular frequency ω_1 in Table 1 corresponds to the channel in Eqs. (29) or (30) in the case of Nd^{3+}. The wave at angular frequency ω_2 in Table 1 has two different roles: first it excites the Nd^{3+} laser center and secondly its nonabsorbed part is up converted by the second-order nonlinear process. ω_2 (corresponding to the wavelength λ_2) is then chosen to match the main Nd^{3+} absorption lines:

$$\begin{aligned} &{}^4I_{9/2} \rightarrow {}^4G_{5/2} - {}^2G_{7/2}\ (\lambda_2 = 590\ \text{nm}) \\ &{}^4I_{9/2} \rightarrow {}^4F_{7/2} - {}^4S_{3/2}\ (\lambda_2 = 750\ \text{nm}) \\ &{}^4I_{9/2} \rightarrow {}^4F_{5/2} - {}^2H_{9/2}\ (\lambda = 800\ \text{nm}) \\ &{}^4I_{9/2} \rightarrow {}^4F_{3/2}\ (\lambda_2 = 800\ \text{nm}) \end{aligned} \tag{32}$$

The most efficient self-frequency mixing lasers based on channels in Eqs. (29)–(31) are gathered in Table 6.

Further Reading

Brenier, A., 2000. The self-doubling and summing lasers: overview and modeling. Journal of Luminescence 91, 121–132.
Butcher, P.N., Cotter, D., 1990. The Elements of Nonlinear Optics. Cambridge: Cambridge University Press.
Huber, G., 1999. Visible cw solid-state lasers. In Advances in Lasers and Applications. Bristol and Philadelphia: Scottish Universities Summer School in Physics & Institute of Physics Publishing.
Joubert, M.F., 1999. Photon avalanche up-conversion in rare earth laser materials. Optical Materials 11, 181–203.
Kaminskii, A.A., 1996. Crystalline Lasers: Physical Processes and Operating Schemes. CRC Press, Inc.
Lenth, W., Macfarlane, R.M., 1992. Upconversion lasers. Optics & Photonics News. 3: 8–15.
Powell, R.C., 1988. Physics of Solid-State Laser Materials. New York: Springer-Verlag.
Risk, W.P., Gosnell, T., Nurmikko, A.V., 2003. Compact Blue-Green Lasers. Cambridge: Cambridge University Press.
Weber, M.J., 2001. Handbook of Lasers. CRC Press.
Yariv, A., 1995. Quantum Electronics, Third Edition John Wiley & Sons.

Thin Disk Lasers

Mikhail Larionov, Dausinger + Giesen GmbH, Stuttgart, Germany

Introduction

A thin disk laser (TDL) is a kind of solid state lasers, which uses a laser medium with a small thickness in the direction of the laser beam. One disk surface is cooled resulting in a heat flow, which is mainly parallel to the laser beam (**Fig. 1**). The term "TDL" using disk, rarer "disc," arises from the German word "Scheibenlaser" and is used in many publications, firstly in Giesen *et al.* (1994). The name "disk" presumes that the disk is circular shaped. The shape is indeed often used, but is not necessarily required. The disk is used in a laser as an amplifying mirror and is therefore often referred to as an active mirror.

The size of the laser disk and its thickness depend on many parameters. Typically, disks have thicknesses between 100 and 200 μm and diameters between 5 and 25 mm. Some pictures of laser disks mounted on heat sinks are shown in **Fig. 2**.

The idea of the TDL evolved in the early 1990s as the answer to the question: "How can the laser output power be increased, while preserving a good beam quality?" The beam quality is a measure of the beam focusability. The output beam quality of lasers often deteriorates at higher output power due to thermal issues. Part of the excitation energy is transferred into heat and deposited in the gain medium. The heat flow in the gain medium causes a temperature gradient, which, consequently, produces an optical parameters gradient in the gain medium.

In **Fig. 3**, left, a typical temperature distribution in a rod-shaped laser medium cooled from its cylindrical circumference is shown. This distribution results in a strong effective focusing lens of the rod. This lens is not completely spherical and disturbs the beam quality. Additionally, depolarization due to mechanical stress caused by thermal expansion produces losses for the laser beam, limiting the achievable output power. In order to avoid these limitations, cooling has to be improved by reduction of the material thickness. This is especially important, since all laser media have relatively low heat conductivity. Even for the best candidates with a heat conductivity of 10 $Wm^{-1}\,K^{-1}$ the temperature increase over 1 mm of the laser material for a heat flow of 5 $MW^{*}m^{-2}$ (which is typical for TDL) is 500K. In order to avoid damage of the laser medium the material thickness has to be reduced, either by reducing the thickness, thereby leading to the thin disk geometry or slab geometry, or by decreasing the diameter, thereby leading to a fiber geometry. Basically all of these geometries allow for a high beam quality: mode-selective light guiding in the fiber ensures a good output beam quality in a properly designed fiber. The thin disk geometry reduces distortions by

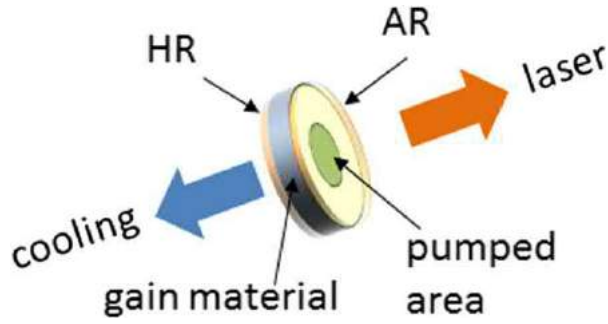

Fig. 1 Thin disk laser (TDL) geometry. AR, antireflective coating; HR, highly reflective coating.

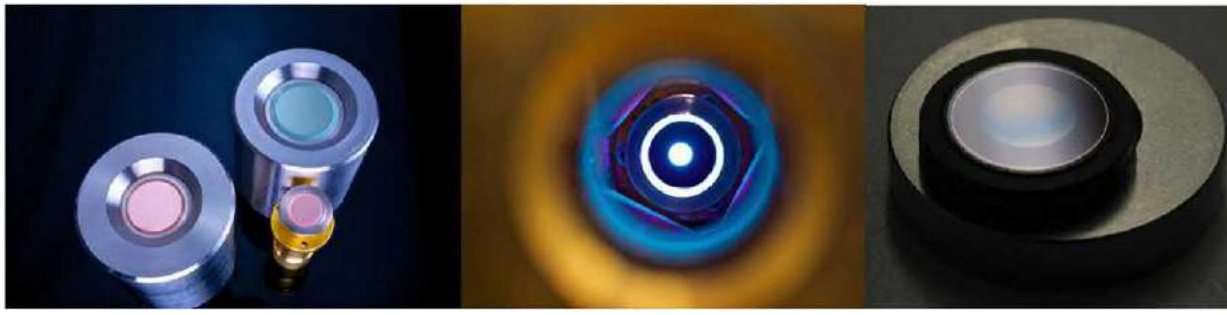

Fig. 2 Examples of disks with different coating, mounted on various heat sinks (left); disk in operation (middle); disk on a structured heat sink (right). Johannes Trbola and Dausinger + Giesen GmbH.

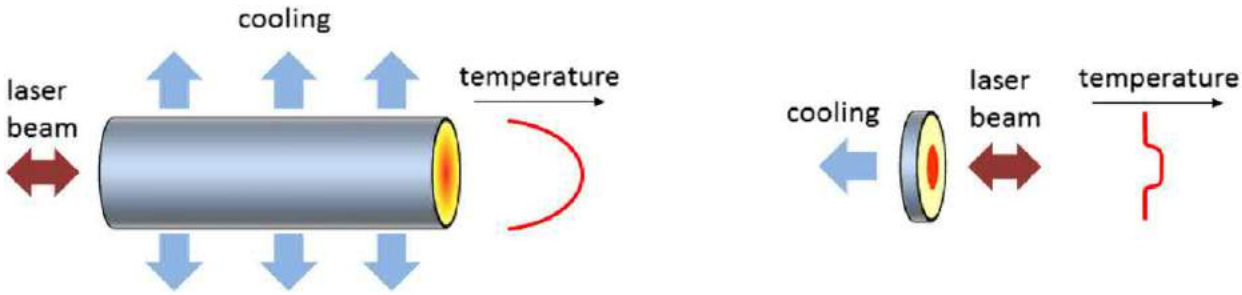

Fig. 3 Temperature distribution in a rod cooled at its cylinder surface (left) and in a disk cooled at one of its plane surfaces (right).

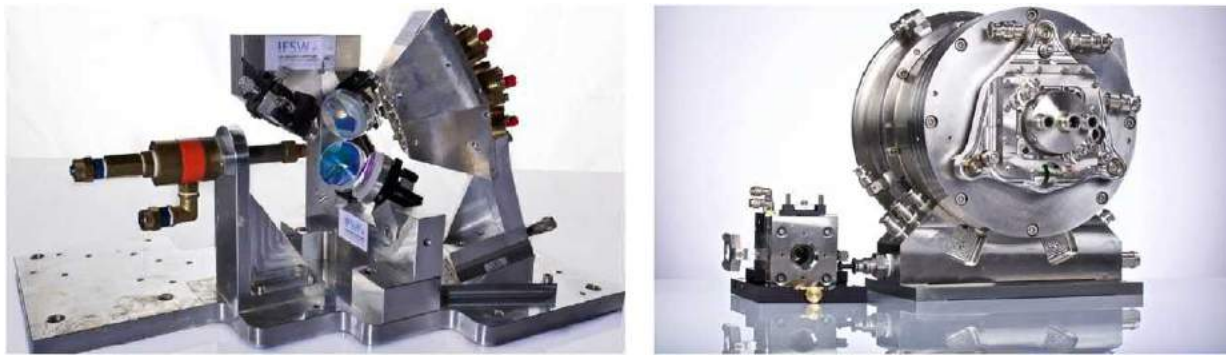

Fig. 4 Left: one of the first thin-disk laser (TDL) setups, demonstrated an output power of 100 W. This setup provides eight passes of the pump power radiation through the thin disk (courtesy of IFSW). Right picture shows two disk modules: the smaller one is designed for 1 kW pump power and the larger one – for 30 kW. Both modules provide 24 or 48 passes of pump radiation through the laser disk, depending on the pump beam quality. Dausinger + Giesen GmbH.

a strong reduction of the material thickness and the slab laser geometry allows for high-power, high beam quality output, however with some spatial anisotropy due to the different thicknesses in the cooling and the laser beam propagation directions.

The invention of the TDLs is closely connected to the ytterbium (Yb) laser ion. In the early 1990s the Nd:YAG gain material was broadly used in the flash-lamp pumped lasers. The advantages of using ytterbium were discussed, but only the availability of laser diodes with sufficient power and narrow spectrum enabled use of Yb in high-power lasers at that time. Ytterbium has several advantages over Nd: lower heat generation, absence of parasitic effects like up-conversion, cross-relaxation and concentration quenching. However, realization of this potential demands for a high excitation density, which is only possible with a sufficient cooling.

In 1992 a group of Dr. Giesen at the University of Stuttgart, IFSW institute invented the main principles of the TDL and started research on it. In 1993 IFSW together with DLR (German aerospace center) have applied for a patent, protecting the basic principles of TDL (Brauch *et al.*, 1995). One of the first TDL setups used at that time is shown in **Fig. 4**.

On the left photograph of the figure a disk holder with two water connections for cooling is shown. Rightward from the disk holder a set of three spherical mirrors and one flat mirror can be seen, reflecting the pump beam four times on the disk. The pump beam originates from an array of seven fiber bundles, placed in the upper right corner.

Support from the German ministry of science and education (BMBF) allowed a fast development, increase of the output power and commercialization of TDLs. Several companies in Germany and outside of Germany bring on the market laser products, based on the thin disk geometry, now. Nowadays, the TDL is a mature technology broadly used in material processing with continuous-wave and pulsed lasers, medicine, diagnostics, etc. In the research field TDL seems to be one of the most promising approaches for enabling of the next generation of research lasers (Fattahi *et al.*, 2014), providing powerful femto- and attosecond pulses. The TDL technology is regarded as enabling the third-generation femtosecond technology after dye (first generation) and titanium:saphire lasers (second generation).

Principle

The disk is coated highly reflective (HR) on the cooled side and serves as an active mirror in the laser setup, i.e., a mirror amplifying the incident beam within two passes through the laser active medium. The opposite side of the disk is typically antireflective (AR) coated for both, laser and pump wavelengths (**Fig. 1**). In this geometry, the heat generated in the laser is

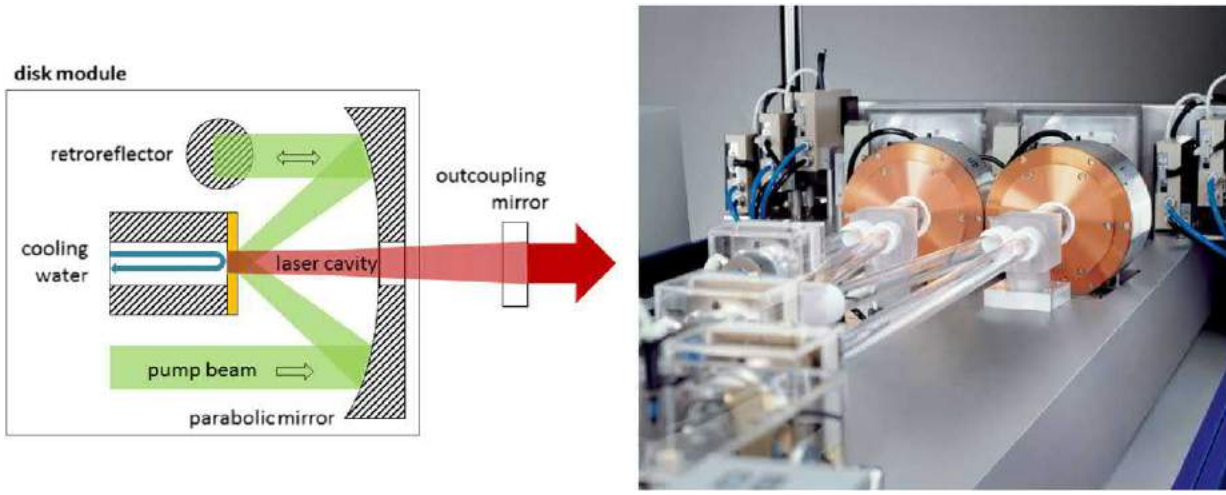

Fig. 5 Typical setup of a thin-disk laser (TDL) (left). Practical realization of a TDL using two disk modules in one laser cavity (right). TRUMPF group.

transported to the coolant on the shortest possible way. The heat flow is parallel to the laser beam. Thus, in the first approximation the radial temperature profile in the disk is nearly flat-top shaped, minimizing the thermal lens and other effects connected to the radial temperature distribution (like stress-induced depolarization).

A typical laser setup is shown in Fig. 5. The laser disk is fixed on a water cooled mount. The pump light is reflected on the disk several times in order to enhance absorption, using a parabolic mirror and several folding mirrors (retroreflectors). This setup is called disk module, pump module, pump chamber, or pump cavity. A laser of the company Trumpf GmbH & Co KG implementing two disk modules, where both disks are used as folding mirrors is shown on the right side of the figure. The cavity beam is encapsulated in transparent tubes.

Pumping Schemes

A thin laser medium provides low absorption and low gain. For the typically used disks absorption is in the range of 10–20%. This is not sufficient for an efficient laser operation. Therefore different multipass schemas are used for recycling the transmitted pump radiation, propagating it through the laser medium as often as possible.

Laser diodes are the most commonly used pump source of the TDL. Most of the schemes are using 4 f or Rayleigh imaging of the pump spot on the disk again and again to shape the same pump spot for each pass of the pump radiation. The simple 2 f-imaging used in the first TDL (as in Fig. 4) leads to an increase of the pump beam divergence from pass to pass, limiting the number of passes. In order to simplify the multipass setup a single parabolic mirror instead of many imaging lenses is typically used. Multiple passes are produced with different configurations of folding mirrors, placed near the focal plane of the parabolic mirror. Several patents dealing with this topic were issued.

An interesting idea is the usage of only two reflectors for a setup with an arbitrary number of absorbing passes through the laser medium. Each reflector consists of a mirror pair, with mirrors arranged at an angle of 90 degree between the mirrors. This idea is illustrated in Fig. 6. A collimated pump beam enters the setup between the reflectors and hits the parabolic mirror at position 1. The parabolic mirror focuses the beam onto the thin disk. The not absorbed part propagates to the position 2 on the parabolic mirror, is re-collimated there and propagates further passing the positions 7–12, where the propagation direction is reversed and the residual pump beam propagates the same way back to position 1 and exits the setup. In total the pump beam completes six reflections on the disk, passing 12 times through the laser medium. Changing of the angle between the symmetry axes of both reflectors changes the number of pump passes.

High pump power is typically available with a low beam quality only. Designing the pump optics for low-beam-quality pump radiation requires usage of larger optics and a reduction of the number of pump passes. In Fig. 4 on the right is a comparison of a pump optics designed for a pump power of 1 kW and a pump optics designed for a pump power of 30 kW.

An arrangement of the pump reflections on the parabolic mirror in two rings or even more complicated patterns can be used. Up to 48 passes of pump radiation in a commercially available multipass optics have been demonstrated. In general, properly designed pump optics can accept a rather low beam quality of the pump radiation. TDL efficiently converts this "low-quality" radiation to radiation with a high beam quality.

For low power lens systems can be also applied.

An obvious way to improve absorption is pumping of the disk through its thin side. This method is often referred as "radial" or "side" pumping. In this case a long way of the pump radiation in the disk plane, which is increased by zig–zag propagation due to total internal reflection, allows efficient absorption of the pump light. This scheme was proposed at the very beginning of the TDL

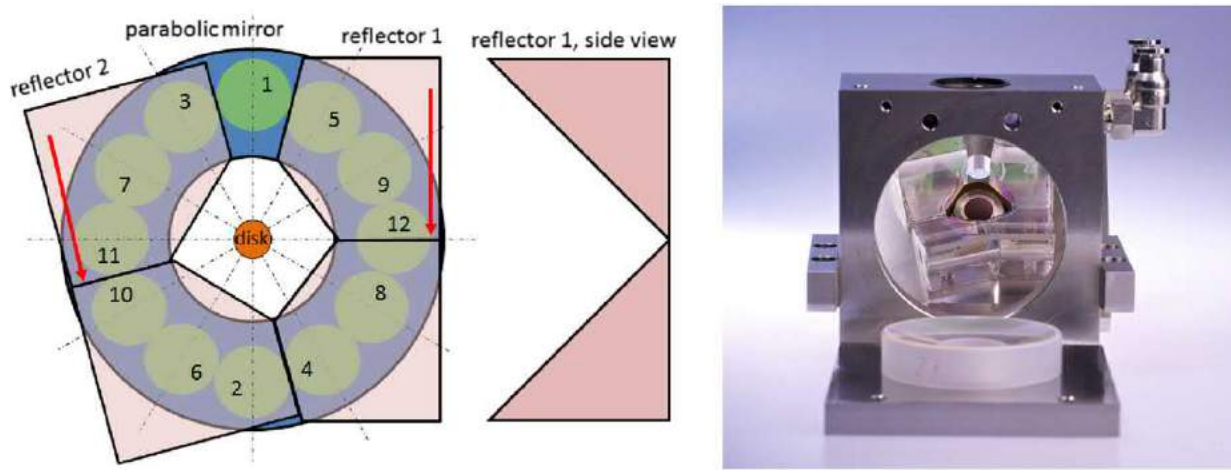

Fig. 6 Left: Principle of a pumping multipass setup, providing 12 reflections on the thin disk or 24 absorption passes of the pump radiation through the laser material. Beam positions on the parabolic mirror are shown with circles and numbered in the same order as beam passes them. Right: Picture shows a practical realization of this setup. TDM1.0, Source: Dausinger + Giesen GmbH.

development, but has not yet found any broad industrial application. Its main drawback is the nonuniformity of the absorbed power in the disk and its dependence on the changes of every emitter.

Gain

Typically the small signal gain of thin disks is low and does not exceed 30% (some exceptions are described in the next chapter). In order to deal with the low gain of the thin laser disk, special techniques have been developed. Multiple reflections on the laser disk in one round trip in the laser cavity or amplifier allow multiplying the gain. For example, using of the laser disk as a folding mirror increases the gain by a factor of 2 compared to the usage as an end mirror.

Assuming a laser disk gain of 8%, which is a typical value for >10 kHz laser operation, 30 reflections on the disk are necessary to achieve a total amplification factor of 10. For an amplification of 1,000,000 the number of reflections is 180. While the first can be realized with a mirror arrangement (multipass amplifier) the second one needs regenerative amplification, where the number of amplifier bounces is defined by a number of round trips in the amplifier cavity.

Mounting

The laser disk itself does not have sufficient mechanical strength to be directly cooled. Therefore, the thin disk is typically mounted on a mechanically stable carrier, which often works also as a heat spreader. Mounting to the carrier is the key technology for TDL. The requirements to the mounting are tough: low thermal resistance, plane or spherical shape of the disk after the mounting within sub 100 nm, high stiffness of the resulting sandwich, withstanding mechanical stress in laser operation, homogeneous mechanical and thermal contact, manufacturability, cost, etc. Several techniques (overview in, e.g., Larionov, 2009) have been broadly used in the last years. First disks were attached to copper heat sinks with a layer of indium. This technique is well suited for the laboratory use, but reaches its limitations for high-power and/or long-term operation due to low creeping resistance and mechanical fatigue of indium. Soldering and gluing of the laser disks were consequently developed and are the mostly used mounting technologies at the moment. Especially, gluing, which allows usage of diamond as a heat sink is a well-established technique for high-power operation. New techniques like diffusion bonding or some kind of "chemically activated" bonding, epitaxial growth – have been also proposed for mounting.

Amplified Spontaneous Emission

ASE plays an important role for TDL. A photon, which is emitted in the pumped volume of the thin disk, propagates at some angle to the disk axes. A large fraction of the photons does not escape from the disk due to total internal reflection and is guided to the edge. For example, for Yb:YAG this fraction is 84%. On its zig-tag way through the pumped volume the photon is amplified with a coefficient, which is much higher than the gain for laser radiation, since the way of the fluorescence photon in the gain medium is much longer. At the edge the photon can be scattered or reflected back in direction of the pumped volume. In this case it gathers even more amplification. This process reduces effective gain of the thin disk for high pump power densities and/or for large pump spots. Typical ways for reduction of ASE is usage of a thick (compared to the thin-disk) undoped cap at the top of the disk, usage of thicker disks, beveling of the disk edge, introducing absorption at the disk edge by co-doping or attaching of an absorbing material.

Recently, several groups published research on cryogenically cooled TDLs with an undoped cap. The undoped cap is used to reduce ASE. In this special architecture the laser disk can be quite thick (up to 1 mm). The laser disk is optically contacted to a thick undoped cap on the side opposite to cooling and cooled down to cryogenic temperatures (usually liquid nitrogen). The thickness of the undoped cap is even larger than that of the disk, allowing an efficient suppression of ASE, because most fluorescence photons pass the gain medium only once and propagate further in the undoped cap. This is a boundary case between thin-disk and bulk laser geometry since the heat flow is not one-dimensional. This leads to a stronger thermal lensing and depolarization compared to the thinner disks. This disadvantage can be mitigated by operation at low temperature.

Scaling

The output power or energy of a TDL can be increased by increasing the pumped area and/or by increasing the number of the used laser disks. Whereas scaling via the disk number is limited at least by the disk damage at high power or energy density, scaling via pumped area is not limited by temperature or mechanical stress as long as pump and laser power/energy densities remain constant. Practically, there are some issues limiting enlarging of the laser disk. Single crystals are available up to a diameter of roughly 100 mm. Sintering allows production of polycrystalline material (often called laser ceramics) with any dimensions and quality comparable to that of the single crystal.

Mounting of a larger thin disk is a technological challenge. Also challenging is designing and setup of the laser cavity for large disks. Once the problems are solved, only ASE limits the achievable energy and power. Several approaches were developed to estimate the limitations of the TDL geometry. An output power of 1 MW is feasible with an optical efficiency of 50% and a pump spot diameter of 20 cm, if the cavity internal losses remain low $L < 0.25\%$. For pulsed laser operation the expected losses are higher. For a loss of 4% an energy of 8.3 J can be extracted. For repetition rates below $1/\tau$ (where τ is fluorescence lifetime) the use of pulsed pumping allows even higher output energies. Lower losses would also allow extraction of pulses with higher energy.

Performance in Different Operation Modes

Continuous Wave

The typical transmission of the output coupling mirror of a TDL in cw operation is only a few percent. The power circulating in the laser cavity is typically 10–50 times higher than the output power. Therefore, the TDL design is sensitive to losses. The typical loss amounts to about 0.1%, requiring a very good mirror coating and a clean environment. At the moment the highest multimode power from one laser disk is 10 kW, as demonstrated by Schad *et al.* (2016).

Fundamental mode operation is typically achieved by adjusting the mode size on the laser disk to the size of the pumped area. The pumped area provides an effective aperture, which cuts off higher transversal cavity modes. In general, cavity design and its sensitivity to thermal lensing and misalignment scales with r_{mode}^2. Therefore, high-power fundamental mode operation was demonstrated later than high-power operation in multimode. The highest power of 4 kW was demonstrated.

In (Speiser, 2016) a way of scaling the output power to the multi-10 kW range is proposed. It includes using a high-power oscillator with a good beam quality and further amplification in a chain of multipass amplifiers. An output power of 12 kW with an M^2 of 10 has been generated using this approach.

Mostly, due to low gain of the thin disk, stable cavities are used for operation of TDLs. Unstable cavities provide high outcoupling ratio and cannot be readily used. The company Boeing used multiple disks to overcome this limitation. A record output power of 25 kW or even 30 kW according to the company's homepage with a "nearly diffraction limited" beam quality was demonstrated. Since the fundamental mode of an unstable cavity is not equal to fundamental Gaussian mode of a stable cavity, it is difficult to compare the beam quality of this result to other results in terms of beam quality.

Single-frequency

Oscillations in a TDL cavity can be limited to one longitudinal mode with suitable frequency-selective elements, allowing generation of high-power, spectrally narrow radiation. A Lyot filter together with an etalon ensure oscillations of only one longitudinal mode. The high laser power circulating in the laser cavity heats the frequency-selective elements used in transmission. A very promising approach is therefore the usage of reflective elements, like a highly efficient dielectric diffraction grating.

A typically negative effect of "spatial hole burning" reduces gain and laser efficiency in a laser operation with only one longtitudinal cavity mode. This effect works in TDL also as an additional etalon, allowing to narrowly select one mode with a small number of filtering elements.

Frequency-converted

TDLs offer an interesting opportunity for intra-cavity frequency conversion. Since the optimum outcoupling of a continuous-wave cavity is in the range of a few percent, only this amount of circulating radiation has to be converted for efficient operation. This allows the usage of nonlinear frequency-converting crystals at moderate power-densities, still providing high laser efficiency. The "green problem" causing output power instability, can be solved, operating the cavity in one longitudinal mode or at least with a sufficiently narrow spectrum. As for single frequency operation, using of an efficient reflective diffraction grating is beneficial

for high power output. 403 W of continuous-wave green radiation was generated using this approach, recently. A few-Watt laser based on this approach is built commercially in high numbers.

Q-Switch and Cavity Dumping

Modulation of the cavity losses allows pulsed operation of the TDL. Typically acoustooptical or electrooptical modulators are used. In Q-switched operation the losses are modulated between low loss and high loss. The output pulse length is defined by the cavity length and laser dynamics. Typical pulse lengths are between 0.2 μs and several microseconds. An output power of 190 W with a nearly diffraction-limited beam quality and pulse energy of maximum 18 mJ has been demonstrated.

In cavity-dumped operation, the losses of the cavity are modulated between roughly 0 and 1. Reduction of the losses to 0 for pulse build-up allows a fast increase of cavity internal power. Fast switching of the losses to 1 dumps the cavity power in one round trip and provides at the output a pulse with a pulse duration defined by the cavity round trip time. Typical pulse durations in this regime are 10–50 ns. Shorter pulse durations are hard to achieve at high power, due to the limitations of the available cavity length. An output power of 850 W with a beam quality of 20 mm × mrad and a pulse energy of up to 80 mJ is commercially available.

Frequency-converted

Modulation of the cavity losses can be combined with intracavity frequency conversion. The low gain of the thin disk allows efficient operation with a low conversion efficiency and thus at a moderate power density in the nonlinear conversion crystal. This allows high-power output with a pulse duration of several 100 ns. The pulse duration is longer than that of cavity-dumped lasers, because conversion extracts power from the laser cavity in multiple round trips and not in one round trip. Lasers based on this principle with an output power of 300 W and a pulse energy of maximum 7.5 mJ are commercially available.

Regenerative and Multipass Amplification

Shorter pulse durations can be obtained with amplifiers. A regenerative amplifier theoretically allows for an unlimited number of reflections on the laser disk. Amplification factors of up to 10^{11} have been experimentally demonstrated. A robust design with a small number of optical components made this principle commercially successful. The mode size on the laser disk is not limited and can be enlarged for higher pulse energies, avoiding damage of the optical components. For amplification of spectrally broad picosecond and subpicosecond pulses chirped-pulse amplification is used. For picosecond pulse duration an output powers of 220 W at 1 kHz, 300 W at 10 kHz and 1.1 kW at 100 kHz have been experimentally demonstrated.

Another interesting application is the amplification of spectrally narrow nanosecond or sub-nanosecond pulses. The output of the laser repeats temporal and spectral characteristics of the seed pulse and can be used for applications requiring high energy and high power pulses with narrow spectrum.

Gain narrowing during amplification can be compensated by nonlinear spectrum broadening in the electro-optical switch or another optical element used in transmission. Pulse duration of an amplifier using Yb:YAG is typically slightly below 1 ps. Using spectrum broadening a pulse duration of well below 300 fs has been shown.

For generation of even higher pulse energy or power the electro-optical switch may become the limiting factor. Therefore, multipass amplifiers, reflecting the signal beam on the laser disk repeatedly by means of turning mirrors, have been developed. Different concepts of beam propagation, including 4 f imaging and not imaging systems, have been built up and evaluated. For pulse amplification optimized for high output power an output power of 1.4 kW with a pulse energy of 6.3 mJ and picosecond pulse duration and, recently, an output power of 1.9 kW has been demonstrated. For high-energy operation an output energy of 1 J at a repetition rate of 100 Hz was achieved with an amplifier chain consisting of a regenerative amplifier and two multipass stages. Rather thick disks with a thickness of 0.75 mm and pulsed pumping have been used, allowing efficient ASE suppression and thermal management. Combination of both goals: high power and high energy remain challenging.

Mode Locking

Mode locking of TDLs allows the highest output-power and energy achieved in the mode-locked regime when compared with other lasers. The reason for this is a very low nonlinearity of the laser gain medium combined with the principal power-scalability of the TDL.

Typically passive mode-locking is realized using semiconductor saturable absorbing mirror (SESAM) or Kerr-lensing. Since the nonlinearity of the laser medium is small, the nonlinearity of other optical components and even of the air in the laser cavity becomes the critical factor for increasing of the output pulse energy using SESAM. This problem can be solved by operation in vacuum or in helium atmosphere or by increase of the output coupling rate, which decreases the energy of the circulating pulses in the cavity. Kerr lens mode-locking produces shorter pulses.

Materials

Yb:YAG

Yttrium aluminum garnet (YAG, $Y_3Al_5O_{12}$), where some yttrium ions are substituted with ytterbium ions (doping) is the mostly used and mature material for TDL. YAG has a heat conductivity of 11 Wm^{-1} K^{-1}. Doping decreases heat conductivity to, for

example, 6 Wm^{-1} K^{-1} for 10% doping concentration. Any doping concentration from 0 to 100% can be manufactured. Typically a concentration around 10% is used for thin disks.

The ytterbium ion has a very simple energy structure, consisting only of two multiplets, which are further split into seven energy levels by crystal field. The heat generation is mainly caused by relaxation in the multiplets, resulting in the quantum defect. At high excitation densities, usual for TDLs, not completely understood parasitic processes provide additional decay channels involving several excited ions.

There is a variety of other materials used in TDL. The reason for using another material is either enhancing of power capability and efficiency or demand for a different spectrum – a broader one for generation of shorter pulses.

Power capability of a laser material depends mainly on its heat conductivity. Therefore, several materials with a heat conductivity higher than that of YAG were proposed for usage in TDL. Among them only Yb:LuAG (lutetium aluminum garnet) found industrial application. It has a very similar structure and parameters as YAG, just Y is substituted by Lu. Since Lu ion has a larger ion radius than Y, doping with Yb does not decrease heat conductivity that much. Sesquioxides are cubic crystals with a formula X_2O_3. Especially Lu_2O_3, Sc_2O_3 and Y_2O_3 have been examined in TDL. The heat conductivity of a doped Lu_2O_3 (12 Wm^{-1} K^{-1}) is higher than that of doped YAG. Additionally, mixed oxides can be manufactured, allowing to broaden the gain spectrum. The main drawback impeding the usage of sesquioxides, so far, is their high melting temperature, making crystal growth difficult. Polycrystalline material (laser ceramics) does not have this drawback.

There are plenty of materials, which were examined in TDL in order to get a broader gain spectrum, maintaining amplification and generation of pulses shorter than possible with Yb:YAG. As a rule of thumb the emission cross-section is inverse proportional to the gain bandwidth. Thus, materials with broader gain bandwidth typically provide lower gain. Broader gain spectrum is mostly caused by some disorder in the crystal structure – anisotropy of presence of different crystal places for Yb ion. Such materials are typically more difficult to handle and to manufacture. Some of interesting candidates are tungstates ($KY(WO_4)_2$), calcium fluoride (CaF_2), CALGO ($CaGdAlO_4$). Only tungstates are used for industrial lasers, so far.

Nd Doping

In general, Neodymium has a more complicated energy structure allowing various parasitic decay channels, compared to Yb and a higher quantum defect. Therefore, the heat generation is much higher, limiting the output power. Due to the large radius of the Nd ion, maximum doping concentration is limited for most host materials, either reducing the laser efficiency or requiring thicker disks, which in turn limits the output power. Nd:YAG and YVO (YVO_4) are used in TDL, in order to get access to various laser wavelengths offered by the energy structure of Nd and to use them after frequency conversion. Different visible colors from blue to red can be produced using this approach. The typical output power is limited to a few dozen Watt.

Ho Doping

Holmium is used to get access to the "eye-safe" wavelength of 2.1 μm. Continuous-wave operation with an output power of 20 W and Q-switched operation with an output power of 9 W have been demonstrated. Ho can be pumped at 1.9 μm, providing a low quantum defect, which is beneficial for high-power operation. The principal possibility of pumping with laser diodes, which is a prerequisite for its commercial usage, has been demonstrated. However, availability and parameters of the available laser diodes are still not sufficient. Once this issue is solved, Holmium can be an interesting candidate even for high-power TDLs. At the moment Tm lasers are often used for pumping of Ho TDL.

Tm Doping

Emission wavelength of Thulium is 1.9 μm and is also "eye-safe." An output power of 21 W and a tuning range from 1899 to 1927 nm was demonstrated using $Tm:LiLuF_4$.

Thulium is often used as a co-doping to provide a pumping wavelength at 790 nm, which is easily accessible with laser diodes. However in TDL this scheme has not been demonstrated to work well.

Cr Doping

Chromium doped ZnSe has been demonstrated to achieve an output power of 5 W at 2.4 μm. The laser was tunable over more than 100 nm from 2.23 to 2.37 μm. The same pumping options as for Ho are available.

Semiconductor

TDL geometry is also used in vertical-external-cavity surface-emitting lasers (VECSELs), which are optically pumped semiconductor lasers (OPSLs). A semiconductor chip including an end Bragg mirror is grown to provide the required parameters, mounted on a heat sink and pumped from the opposite side. This geometry is often referred to as semiconductor disk laser (SDL), underlining its connection to TDL geometry. Using a grown semiconductor structure as a laser medium provides a very interesting possibility to engineer the laser wavelength. Intracavity frequency doubling allows to access the visible spectrum range. The inherent "green

problem″ is not present in SDL design allowing for a very simple cavity design. Such lasers are commercially available in a variety of wavelengths with an output power of up to 20 W.

References

Brauch, U., Giesen, A., Voss, A., Wittig, K., 1995. Laser Amplifier System. Idea of Thin-Disk Laser, European Patent EP 0 632 551 B1.
Fattahi, H., Barros, H.G., Gorjan, M., *et al.*, 2014. Third-generation femtosecond technology. Optica 1, 45–63.
Giesen, A., Hügel, H., Voss, A., *et al.*, 1994. Scalable concept for diode-pumped high-power solid-state-lasers. Applied Physics B 58, 365–378.
Larionov, M., 2009. Kontaktierung und Charakterisierung von Kristallen für Scheibenlaser. Munich: Herbert Utz Verlag.
Schad, S., Gottwald, T., Kuhn, V., *et al.*, 2016. Recent development of disk lasers at TRUMPF. In: Proceedings of the SPIE. 9726, Solid State Lasers XXV: Technology and Devices.
Speiser, J., 2016. Thin disk lasers: History and prospects. In: Proceedings of SPIE 9893, Laser Sources and Applications III, 98930 L.

Further Reading

Contag, K., Erhard, S., Giesen, A., *et al.*, 2000. Laser Amplification System. Pumping of Multiple Disks. International Patent WO 00/08726 A3.
Erhard, S., Giesen, A., Stewen, C., 2002. Laser Amplifier System. Pump Optics With Two Reflectors. European Patent EP1252687 B1.
Erhard, S., Giesen, A., Karzewski, M., Stewen, C., Voss, A., 2000. Laser Amplification System. Pump Optics With Parabolic Mirror. International Patent WO 00/08728.
Hartke, R., Baev, V., Seger, K., *et al.*, 2008. Experimental study of the output dynamics of intracavity frequency doubled optically pumped semiconductor disk lasers. Applied Physics Letters 92, 101107.

Relevant Websites

www.dausinger-giesen.de
Dausinger + Giesen GmbH.
www.jenoptik.com
Jenoptik.
www.trumpf.com
Trumpf.
www.rp-photonics.de
RP Photonics.

Microchip Lasers

John J Zayhowski, Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, MA, United States

Introduction

Microchip lasers are a rich family of solid-state lasers defined by their small size, robust integration, reliability, and potential for low-cost mass production. In its simplest embodiment, a microchip laser is a monolithic device that consists of a small piece of solid-state gain medium polished flat and parallel on two opposing sides. Dielectric cavity mirrors are deposited directly on the gain medium and the laser is pumped with a diode laser, either directly, as shown in Fig. 1, or via an optical fiber. In other embodiments, two or more materials are joined together before being polished and dielectrically coated to form a composite-cavity microchip laser with added functionality or enhanced mode properties. For example, the additional material could be an electro-optic material to enable frequency tuning of the laser, a nonlinear material for intracavity wavelength conversion, a saturable absorber to Q switch the device, or a passive transparent material to increase both the transverse mode area and the output power of the laser.

Continuous-wave (CW) microchip lasers cover a wide range of wavelengths, often operate single frequency in a near-ideal mode, and can provide a modest amount of tunability. Q-switched microchip lasers provide the shortest output pulses of any Q-switched solid-state laser, with peak powers up to several hundred kilowatts. The average output power of microchip lasers is typically in the range from several tens to several hundreds of milliwatts.

Background

Most solid-state lasers are built from discrete optical components that must be carefully assembled and critically aligned. Laser assembly is typically performed by trained technicians, and is time consuming and expensive. As a result, the cost of most solid-state lasers makes them unattractive for a wide range of applications for which they would otherwise be well suited. Additional characteristics that have historically impeded the widespread use of solid-state lasers include their size and reputation for being fragile and unreliable. This was even truer in the early 1980s, the infancy of microchip laser development, than it is today. Microchip lasers were developed to overcome these limitations – cost, size, robustness, and reliability – and thereby become viable components for a variety of large-volume applications. The term "microchip laser" was coined at MIT Lincoln Laboratory in the early 1980s to draw an analogy between this new class of lasers and semiconductor electronic microchips with their inherent small size, reliability, and low-cost mass production.

Consider the simplest of all possible microchip lasers, a small piece of solid-state gain medium polished flat and parallel on two opposing sides, with dielectric cavity mirrors deposited directly onto the polished faces, as shown in Fig. 1. Fabrication of the laser starts with a large boule of gain material, such as Nd:YAG. The boule is sliced into wafers about 0.5-mm thick. The wafers are then polished and dielectrically coated before they are diced into 1-mm-square pieces, with each piece being a complete laser. One boule can produce thousands of lasers, and throughout the fabrication process the lasers never need to be handled independently.

To complete a microchip laser system, the laser must be coupled to a pump source. Microchip lasers are pumped with semiconductor diode lasers. The 1980s were a time of rapid development in diode lasers. The amount of power that was available from commercial diode lasers was rapidly increasing and the cost per watt of output power was quickly decreasing, with projections of extremely inexpensive, high-power diode lasers in the near future. As a result, diode lasers fit nicely into the picture of low-cost microchip laser systems. To keep the cost of the system low, it is important that the coupling of the diode to the microchip laser be performed inexpensively. The use of a flat–flat laser cavity eliminates any critical alignment between the diode and the laser and makes the assembly of the system quick and simple, with the potential for inexpensive automation.

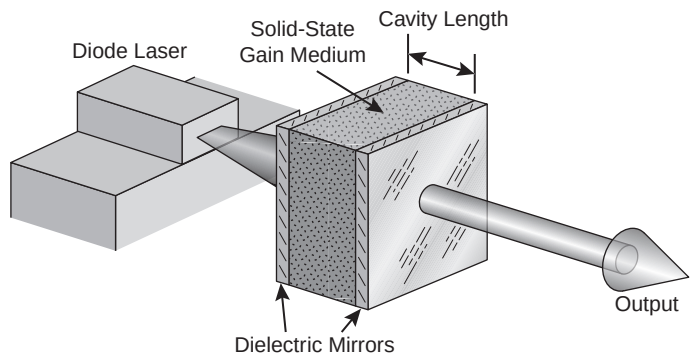

Fig. 1 Illustration of a monolithic microchip laser.

The use of simple, small, monolithic, mass-produced solid-state lasers noncritically coupled to low-cost semiconductor diode pump lasers gives microchip laser systems their defining characteristics – small size, robust integration of components, reliability, and the potential for low-cost mass production – and differentiates them from other miniature laser systems that are designed and constructed using more conventional techniques.

Since their early development, microchip lasers have evolved into a rich family of devices with capabilities that often exceed those of conventional lasers. The first microchip lasers developed operated CW, and a variety of CW microchip lasers were quickly demonstrated, covering a wide range of wavelengths. Some of the early applications required a modest amount of tunability, and several tuning mechanisms were incorporated into the laser structure. It was not long before researchers realized that the short cavity lengths inherent to microchip lasers gave them tremendous potential as pulsed devices. This led to the development of actively Q-switched microchip lasers, quickly followed by the most successful variation of microchip laser, the passively Q-switched microchip laser, which has demonstrated the shortest output pulse of any Q-switched solid-state laser.

Overview

Monolithic Microchip Lasers

The earliest microchip lasers were monolithic devices based on optical transitions in the Nd^{3+} ion near 0.94, 1.06, or 1.32 μm, using a variety of gain media including Nd:YAG, NPP, LNP, Nd:GSGG, Nd:YVO_4, Nd:$LaMgAl_{11}O_{19}$, Nd:YCeAG, Nd:YLF, $Nd_xY_{1-x}Al_3(BO_3)_4$, Nd:La_2O_2S, Nd:MgO:$LiNbO_3$, and Cr,Nd:YAG. It was not long, however, before other active ions were investigated and monolithic microchip lasers were constructed at a wide variety of wavelengths. These include Cr:LiSAF microchip lasers operating in the 0.8- to 1.0-μm spectral region, Yb:YAG microchip lasers at 1.03 μm, Yb,Er:glass microchip lasers at 1.5 μm, Tm and Ho microchip lasers operating near 2 μm, Cr-doped chalcogenides microchip lasers near 2.3 and 2.5 μm, and Er-doped microchip lasers at 3 μm. Of the gain media demonstrated in microchip lasers, Nd:YAG and Nd:YVO_4 are the most commonly used.

All of the lasers listed above operate in a near-ideal fundamental transverse mode, and several operate single frequency (in a single longitudinal mode) with a narrow linewidth. The $Nd_xY_{1-x}Al_3(BO_3)_4$ microchip laser has another very interesting characteristic. In addition to being a good gain medium, $Nd_xY_{1-x}Al_3(BO_3)_4$ has a second-order nonlinearity that allows it to perform harmonic generation, and the microchip laser is self frequency doubled to produce green output at 531 nm. The Nd:MgO:$LiNbO_3$ microchip laser is electro-optically tunable. The doubly doped Cr,Nd:YAG laser is the only one of the lasers listed above that is not CW – it is self passively Q switched.

Composite-Cavity Microchip Lasers

The search for multifunctional gain media – gain media that simultaneously act as harmonic converters, electro-optic material, or passive Q switches – has led to several interesting monolithic devices. However, the multifunctional media are often less than ideal in one or more of their functions, and/or are difficult or expensive to grow. Higher-performance devices can be obtained by combining two or more specialized materials within the same composite cavity. In composite-cavity microchip lasers the constituent materials are bonded together to form a quasi-monolithic device, with dielectric mirrors deposited (or bonded) on the outer surfaces.

Several issues must be addressed in the design of composite-cavity microchip lasers. Different materials have different refractive indices and different thermal-expansion coefficients. The optical, mechanical, and thermal properties of the material interface must be dealt with in a way that satisfies the optical requirements of the laser cavity, is robust enough for the intended application, and is cost-effective. Otherwise, the less-than-ideal performance of a monolithic device, if one can be constructed, may be the best solution.

Composite cavities have been successfully employed in coupled-cavity microchip lasers to obtain single-frequency operation from gain media with extremely broad gain bandwidths, harmonically converted green and blue CW microchip lasers, electro-optically tunable microchip lasers, actively Q-switched microchip lasers, passively Q-switched microchip lasers, and frequency-converted passively Q-switched microchip lasers.

Pumping

Microchip lasers are typically pumped by semiconductor diode lasers. The simplest configuration places the microchip laser in close proximity to the output facet of the pump diode with no intervening optics. As the amount of pump power increases, the diameter of the oscillating mode in a microchip laser usually decreases and, for this proximity-coupled configuration, the typically large divergence of the diode output overfills the oscillating-mode volume resulting in inefficient operation or multi-transverse-mode oscillation. The situation can be improved by putting a lens between the diode and the microchip cavity, which is common practice in moderate- or high-power microchip laser systems (output powers in excess of several tens of milliwatts).

An alternative configuration for pumping microchip lasers uses fiber-coupled diode lasers. This configuration decouples the diode-laser system from the microchip cavity and offers several practical advantages. It separates the engineering of the pump subsystem from the microchip cavity; it makes it easy to independently control the temperature of the diodes and the microchip

cavity; and it facilitates an extremely compact laser head that can fit into small spaces, coupled to the rest of the system by a single, flexible optical fiber. High-power microchip lasers are often pumped by diode-laser arrays. In this case, fiber coupling serves the additional function of shaping the combined diode-laser output.

Transverse Mode Definition

Most microchip lasers use a flat–flat cavity design. The eigenmodes of a flat–flat cavity are plane waves, yet microchip lasers typically operate in a near-ideal, fundamental transverse mode with a well-defined mode radius. The mode-defining mechanisms vary depending on the gain medium, other media within a composite-cavity device, and pump power.

For many microchip lasers the transverse mode is determined primarily by thermal effects. When a microchip laser is longitudinally pumped, the pump beam deposits heat. In materials with a positive change in refractive index with temperature, such as Nd:YAG, this creates a thermal lens that stabilizes the cavity and defines the transverse mode. When the cavity mirrors are deposited directly on the gain medium the heat also induces curvature of the mirrors. For materials with a positive thermal-expansion coefficient this contributes to the stabilization of the transverse mode.

In three-level or quasi-three-level lasers there can be significant absorption of the oscillating light in unpumped regions of the gain medium. This creates a radially dependent loss that can restrict the transverse dimensions of the lasing mode. The mere absence of gain in the unpumped regions of the gain medium can have a similar effect. In passively Q-switched devices bleaching of the saturable absorber results in a dynamic aperture that opens as the Q-switched pulse forms and is an important mode-defining mechanism.

Optical gain provides dispersion. Gain-related index guiding has been shown to play an important role in Nd:YVO_4 microchip lasers, and can lead to interesting effects such as self Q switching.

Parallelism between the cavity mirrors can be critical for the creation of a circularly symmetric fundamental transverse mode in a flat–flat cavity. At high pump powers the mode-defining mechanisms in microchip lasers are usually thermal, and the requirement on parallelism is relaxed as the power of the laser increases. The requirement is less severe in microchip lasers based on three-level gain media or employing a saturable absorber.

Methods consistent with low-cost mass fabrication have been developed to put curved mirrors on microchip lasers. Curved mirrors can stabilize the transverse mode of the laser when the mechanisms discussed above are not strong enough to do so. They can reduce the threshold of CW microchip lasers and make low-power operation more consistent from device to device. For a large variety of gain media they are not needed, especially when the laser is pumped with medium- or high-power diodes (typically more than several tens of milliwatts).

Typical mode radii of microchip lasers fall in the range from 20 to 150 μm, depending on the gain medium, pump power, cavity length, and several other factors. To ensure oscillation in the fundamental transverse mode, that mode must use most of the gain available to the laser. If the radius of the fundamental mode is much smaller than the radius of the pumped region of the gain medium, higher-order transverse modes will oscillate.

Spectral Properties

Single-Frequency Operation

As the cavity length of a laser decreases, its free spectral range increases. The original microchip-laser concept included making the laser cavity sufficiently short that its free spectral range is comparable to the gain bandwidth of the gain medium. Robust single-frequency operation can then be obtained as long as one of the cavity modes falls near the peak of the gain profile. This requires precise, sub-wavelength control of the cavity's optical length (length times refractive index), and is often accomplished through thermal control of the cavity.

Typically, each pulse of a Q-switched microchip laser is single frequency. However, at high repetition rates (when the interpulse period is short compared to the relaxation time of the gain medium) consecutive pulses may correspond to different longitudinal modes.

Fundamental Linewidth

One contribution to the spectral width of all lasers is the coupling of spontaneous emission to the oscillating mode. This gives rise to the Schawlow–Townes linewidth, which has a Lorentzian power spectrum with a width that scales inversely with the output power of the laser. In microchip lasers, thermal fluctuations of the cavity length at a constant temperature can result in a much larger fundamental linewidth. These fluctuations result in a Gaussian power spectrum that, for monolithic devices, scales as the inverse square root of the oscillating-mode volume. Monolithic CW Nd:YAG microchip lasers with a cavity length of ~ 1 mm have a Gaussian spectral profile with a linewidth of several kilohertz, with spectral tails corresponding to a Lorentzian contribution of only a few hertz.

Single-frequency Q-switched microchip lasers have a Fourier-transform-limited optical spectrum that is much broader than the fundamental linewidth of CW devices.

Frequency Tuning

Microchip lasers are frequency tuned by changing the cavity's optical length. The optical length can be changed using a variety of techniques including thermal tuning, stress tuning, and electro-optic tuning. Pump-power modulation represents a special case of thermal tuning. Each of these techniques allows continuous frequency modulation of a single longitudinal cavity mode.

Changing the temperature of an element in a laser cavity, or the entire cavity, is often the simplest way to tune a laser. A change in temperature results in a change to both the physical length of the component and its refractive index.

The cavity modes of a resonator also tune as elements within the cavity are squeezed. Squeezing transverse to the resonator's optical axis results in an elongation of the material along the axis. Superimposed on this is the stress-optic effect. In crystals with cubic symmetry, the stress-optic effect can split the frequency degeneracy of orthogonally polarized optical modes. For squeezing along the optical axis of the cavity there is a compression of the squeezed elements and the frequency degeneracy of orthogonally polarized modes remains unchanged. By using a piezoelectric transducer to squeeze a monolithic Nd:YAG microchip laser, researchers demonstrated tuning at modulation frequencies up to 20 MHz, although nonresonant response was limited to $\sim$80 kHz. The nonresonant tuning response was 300 kHz V^{-1}.

Many applications require high rates of frequency tuning that can only be achieved electro-optically. For high-sensitivity tuning it is desirable to fill the cavity with as large a fraction of electro-optic material as possible. However, it is often still important to keep the total cavity length as short as possible, to ensure single-frequency operation, to maximize the tuning range, and to minimize the response time. Composite-cavity electro-optically tuned Nd:YAG/$LiNbO_3$ microchip lasers have been continuously tuned over a 30-GHz range with a tuning sensitivity of $\sim$14 MHz V^{-1}. The tuning response was relatively flat for tuning rates from DC to 1.3 GHz. Monolithic Nd:MgO:$LiNbO_3$ and Nd:$LiNbO_3$ electro-optically tuned microchip lasers have also been demonstrated.

Changes in pump power induce frequency changes in the output of solid-state lasers. As the pump power increases more heat is deposited in the gain medium, causing the temperature to rise and changing both the refractive index and length. Because frequency tuning via pump-power modulation relies on thermal effects, it is often thought to be too slow for many applications. In addition, modulating the pump power has the undesirable effect of changing the amplitude of the laser output. However, for microchip lasers significant frequency modulation can be obtained at relatively high modulation rates with little associated amplitude modulation. For example, pump-power modulation of a 1.32-μm microchip laser has been used to obtain 10-MHz frequency modulation at a 1-kHz rate and 1-MHz frequency modulation at a 10-kHz rate, with an associated amplitude modulation of less than 5%. This technique has been employed to phase lock two microchip lasers and introduced less than 0.1% amplitude modulation on the slave laser. When it can be used, pump-power modulation has advantages over other frequency-modulation techniques since it requires very little power, no high-voltage electronics, no special mechanical fixturing, and no additional intracavity elements.

Polarization Control

Linear polarization is easily achieved in microchip lasers that employ an anisotropic gain medium, but can be problematic for monolithic microchip lasers with isotropic gain media. For lasers with isotropic gain media, the polarization degeneracy of the gain medium can often be removed by applying uniaxial transverse stress or asymmetric heat sinking. In the absence of any other polarizing mechanism, the polarization of the pump light, or asymmetry in its transverse mode profile, may determine the polarization of the laser output. In either of these cases, the polarization selectivity can be weak and the polarization of the laser may be sensitive to perturbations, including external feedback.

In passively Q-switched microchip lasers that employ an isotropic gain medium it is often the properties of the saturable absorber that determine the polarization of the laser.

Pulsed Operation

Pulsed output has been obtained from microchip lasers using a variety of techniques, including active Q switching, passive Q switching, and mode locking.

The shortest output pulse that can be obtained from a Q-switched laser has a full-width at half maximum of 8.1 times the round-trip time of light in the laser cavity divided by the natural logarithm of the round-trip gain. Since microchip lasers are physically very short, they have very short round-trip times. This leads to the possibility of producing very short Q-switched output pulses.

Active Q Switching

Actively Q-switched microchip lasers can be realized using the coupled-cavity configuration illustrated in Fig. 2. An etalon containing an electro-optic element ($LiTaO_3$) serves as a variable-reflectivity output coupler for a gain cavity defined by the two reflective surfaces adjacent to the gain medium (Nd:YAG). The reflectivity of the etalon for the potential lasing frequencies of the device is controlled by a voltage applied to a pair of electrodes in contact with the electro-optic material.

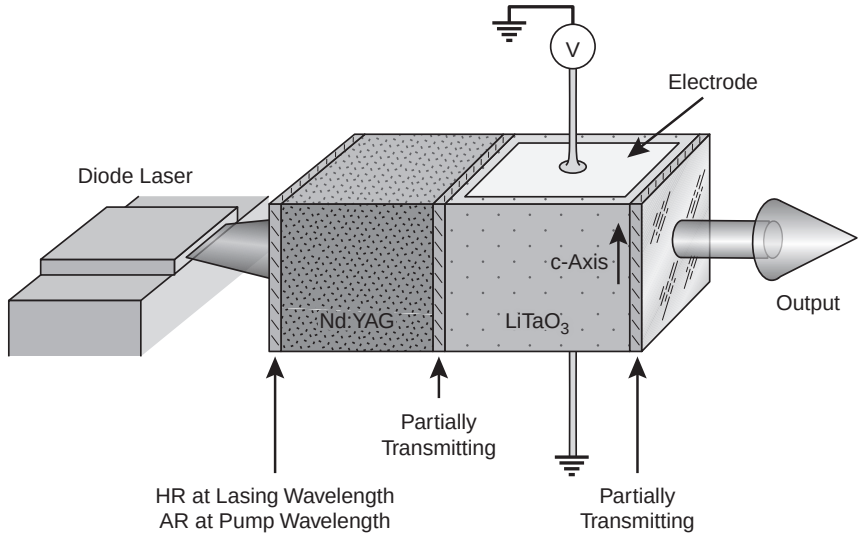

Fig. 2 Coupled-cavity electro-optically Q-switched microchip laser. AR, antireflective; HR, highly reflective.

Coupled-cavity electro-optically Q-switched microchip lasers have demonstrated the shortest Q-switched pulses obtained from any actively Q-switched solid-state laser to date. A coupled-cavity Nd:YAG device, pumped with a 500-mW diode laser, produced 270-ps pulses with a pulse energy of 6.8 μJ at a repetition rate of 5 kHz. A coupled-cavity Nd:YVO_4 microchip laser pumped at the same power level produced 115-ps pulses with a pulse energy of 12 μJ at a 1-kHz repetition rate. Pumped slightly harder, with a pump power of 1.2 W, a Nd:YVO_4 microchip laser demonstrated 8.8-ns pulses at pulse repetition rates as high as 2.25 MHz.

Passive Q Switching

Passively Q-switched microchip lasers contain a gain medium and a saturable absorber, as shown in the top of Fig. 3. When the gain medium absorbs sufficient pump energy that the gain in the laser cavity is greater than the loss, an intracavity optical field forms. The optical field saturates the loss of the saturable absorber and the gain of the gain medium. When the materials are chosen properly, the loss of the saturable absorber saturates more quickly than the gain of the gain medium and the laser will Q switch, generating a short, intense pulse of light without the need for high-voltage or high-speed electronics.

In addition to simplicity of implementation, the advantages of a passively Q-switched laser include the generation of pulses with a well-defined energy and duration. The pulse energy and duration are determined by the design of the laser cavity and the material properties of the gain medium and passive Q switch. Passively Q-switched microchip lasers have demonstrated pulse energies and pulse widths with stabilities of better than 1 part in 10^4. This is achieved at the expense of pulse-to-pulse timing jitter, caused primarily by fluctuations in the pump source.

To manage thermal effects, high-power passively Q-switched microchip lasers are often pulse pumped. A common implementation uses an external clock to turn the pump diodes on and a signal generated by the Q-switched output pulse to turn them off. By using high-power pump diodes at a low duty cycle, larger amounts of energy can be stored in the gain medium of the laser with reduced thermal loading. This results in a larger oscillating-mode diameter and more energetic pulses.

Q switching with bulk saturable absorbers

The most commonly used bulk saturable absorber for passive Q switching of microchip lasers is Cr^{4+}:YAG. It has been used to Q switch Nd:YAG microchip lasers operating at 946 nm, 1.064 μm, and 1.074 μm; Nd:YVO_4 microchip lasers operating at 1.064 μm; Nd:$GdVO_4$ microchip lasers operating at 1.062 μm; and Yb:YAG microchip lasers operating at 1.03 μm.

The combination of Cr^{4+}:YAG and a doped YAG gain medium, such as Nd:YAG, is particularly attractive from the point of view of an extremely robust device. Since both materials use the same host crystal, YAG, they can be diffusion bonded to each other in a way that blurs the distinction between a monolithic and a composite-cavity device. Both materials have the same thermal and mechanical properties and the same refractive index, and the bond between them can be sufficiently strong that the composite device acts in all ways as if it were a single crystal. An early, low-power, diffusion-bonded Nd:YAG/Cr^{4+}:YAG microchip laser is shown in the bottom of Fig. 3.

As an alternative to diffusion bonding, Nd:YAG can be epitaxially grown on Cr^{4+}:YAG and vice versa, using nonequilibrium growth techniques that can produce Nd or Cr^{4+} concentrations that could not otherwise be achieved. Cr^{4+},Nd:YAG can also be grown as a single crystal, although this approach precludes some of the device optimization that can be achieved when the two materials are physically distinct.

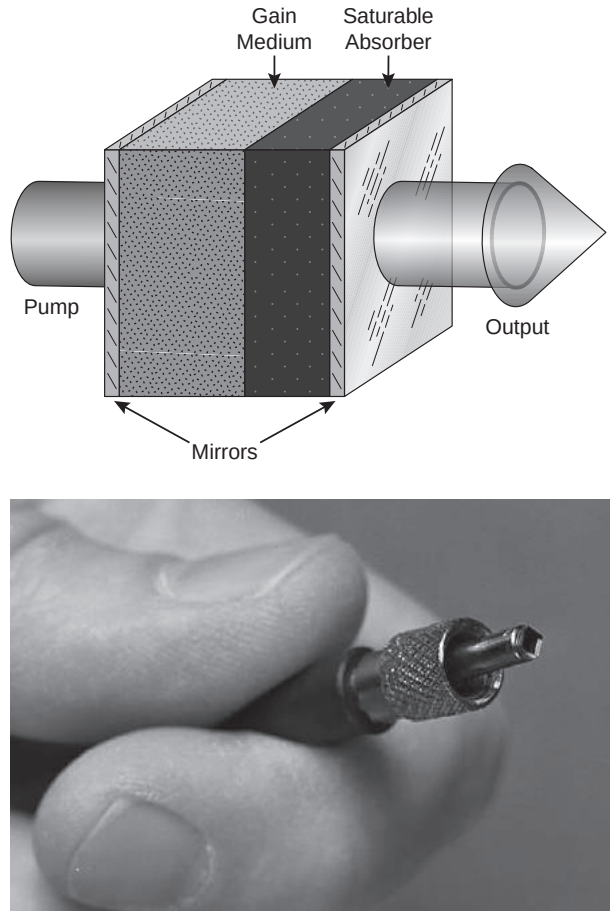

Fig. 3 Composite-cavity Nd:YAG/Cr^{4+}:YAG passively Q-switched microchip laser: (top) schematic and (bottom) photograph of laser bonded to ferrule of pump fiber.

Typically, Nd:YAG/Cr^{4+}:YAG passively Q-switched microchip lasers are operated with an output coupling approximately equal to the single-pass loss of the unsaturated saturable absorber. The minimum pulse width that can be obtained is limited by the average inversion density that can be achieved within the oscillating-mode volume, pump-induced bleaching of the saturable absorber, and a reduction in the gain cross section caused by heating of the gain medium as the laser is pumped harder. Passively Q-switched Nd:YAG/Cr^{4+}:YAG microchip lasers typically produce pulses with full-widths at half maximum between 300 ps and $\sim$1 ns, and pulses as short as 150 ps have been demonstrated.

The maximum pulse energy that can be obtained from passively Q-switched Nd:YAG/Cr^{4+}:YAG microchip lasers is limited by the amount of stored energy the fundamental mode can access, and is strongly influenced by thermal effects. Devices pumped with a 1-W diode laser can generate 1.064-μm pulses with energies up to $\sim$15 μJ and peak powers up to $\sim$30 kW, at pulse repetition rates of $\sim$10 kHz. Less energetic pulses, with lower peak powers, have been demonstrated at repetition rates up to 110 kHz. By pulse pumping microchip lasers with high-power diode-laser arrays, pulse energies up to $\sim$30 μJ with peak powers up to $\sim$100 kW are achieved at repetition rates of $\sim$10 kHz; pulse energies of $\sim$250 μJ with peak powers up to $\sim$500 kW are obtained at repetition rates of $\sim$1 kHz; and pulse energies of several millijoules with peak powers of several megawatts are achieved at repetition rates up to $\sim$100 Hz.

The combination of Yb:YAG and Cr^{4+}:YAG has the same potential as Nd:YAG and Cr^{4+}:YAG for quasi-monolithic and monolithic integration. The longer upper-state lifetime of Yb:YAG is attractive for low-repetition-rate systems since it allows the gain medium to accumulate energy for a longer time, making it possible to use lower-power, less-expensive pump diodes.

To date, the best results in the eye-safe spectral region were obtained from passively Q-switched microchip lasers that use Yb,Er:glass as the gain medium and Co^{2+}:$MgAl_2O_4$ (Co^{2+}:MALO) as the saturable absorber. When pumped with a 1-W 975-nm diode laser, these devices have produced output pulses of $\sim$5-ns duration at repetition rates up to 25 kHz, with peak powers as high as 1.6 kW and average output powers up to 150 mW. At low duty cycles, pulses as short as 880 ps, pulse energies as high as 110 μJ, and peak powers in excess of 35 kW have been demonstrated.

Other combinations of gain medium and bulk solid-state saturable absorber have been used in Q-switched microchip lasers operating at a variety of wavelengths. Cr^{4+}:YAG has been used with numerous gain media at wavelengths between 0.91 and 1.08 μm. The Cr^{5+}-vanadates are a relatively new family of saturable absorbers that are used in the same spectral region. Although lasers based on them have not yet demonstrated the short pulse durations or high peak powers that are obtained with Cr^{4+}:YAG,

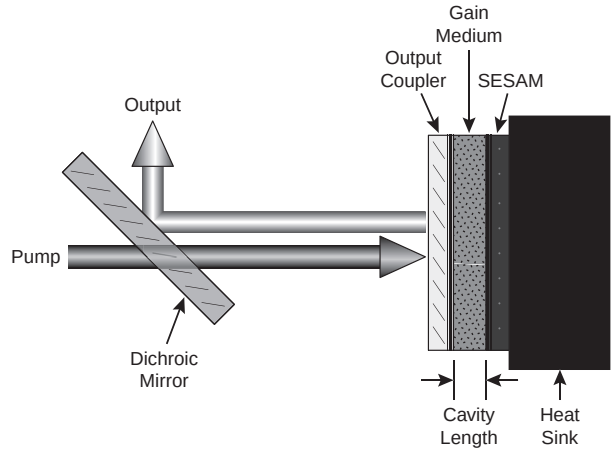

Fig. 4 Typical configuration for a passively Q-switched microchip laser employing a semiconductor saturable-absorber mirror (SESAM) Q switch.

they offer the potential for quasi-monolithic and monolithic integration with vanadate gain media. V^{3+}:YAG has been used as a saturable absorber over the spectral range from 0.93 to 1.44 μm. It is most useful in the long-wavelength portion of this range, where Cr^{4+}:YAG is not an option. Co^{2+}:$LaMgAl_{11}O_{19}$ (Co^{2+}:LMA) and Co^{2+}:MALO extend the coverage of solid-state saturable absorbers into the eye-safe spectral region.

Bulk semiconductor saturable absorbers have been used to Q switch miniature solid-state lasers at wavelengths from 1 to 2 μm, but have a much lower damage threshold than the solid-state saturable absorbers discussed above and are therefore rarely used. They also have very different thermal and mechanical properties than solid-state gain media, making robust integration of miniature devices challenging.

Q switching with semiconductor saturable-absorber mirrors

Semiconductor saturable-absorber mirrors (SESAMs) contain quantum-well saturable absorbers. When high-intensity light at the proper wavelength is incident on the mirror the absorption of the quantum wells saturates and the reflectivity of the mirror increases. In a passively Q-switched microchip laser employing a SESAM, the SESAM is used as one of the cavity mirrors.

When a SESAM is used to Q switch a microchip laser, the physical length of the saturable-absorber region of the microchip cavity is small and its contribution to the round-trip time of light in the laser cavity is negligible, resulting in the shortest possible Q-switched pulses. SESAMs have extremely short upper-state lifetimes, which allow Q-switched lasers employing them to operate at very high pulse repetition rates. One of the main limitations of SESAMs is their relatively low damage threshold.

A practical consideration when using SESAMs is that they typically cannot be used as input or output couplers. As a result, the output coupler is usually on the pump-side face of the laser, as shown in Fig. 4. Also, the thermal-expansion coefficients of SESAMs are not well matched to solid-state gain media, making robust bonding of the gain medium to the saturable absorber challenging.

As a result of their advantages and limitation, SESAMs are most attractive in applications requiring short, low-energy pulses, and where requirements on the system's robustness can be relaxed. In this regime, they can be operated at very high repetition rates, can produce extremely short pulses, and can be engineered to work with gain media at many different wavelengths.

Nd:YVO_4 microchip lasers passively Q switched with SESAMs have produced 1.064-μm output pulses as short as 16 ps, and have been pulsed at repetition rates up to 7 MHz. SESAMs have also been used to Q switch microchip lasers operating near 1.03, 1.34, and 1.5 μm. The largest pulse energy reported for a passively Q-switched microchip laser using a SESAM is 4 μJ, with tens to hundreds of nanojoules being more typical.

Mode locking

Although many microchip lasers are designed to operate in a single longitudinal mode, they need not be. Optical-domain generation of millimeter-wave signals for fiber radio led to the development of monolithic electro-optically mode-locked 1.085-μm Nd:$LiNbO_3$ microchip lasers with pulse widths as short as 18.6 ps and repetition rates up to 20 GHz. In this application, the pulse repetition rate of the laser is the radio carrier frequency.

Frequency Conversion and Amplification

Nonlinear frequency generation is an important adjunct to microchip laser technology. The high peak powers obtained from Q-switched microchip lasers have enabled a variety of miniature nonlinear optical devices. Harmonic generation, frequency

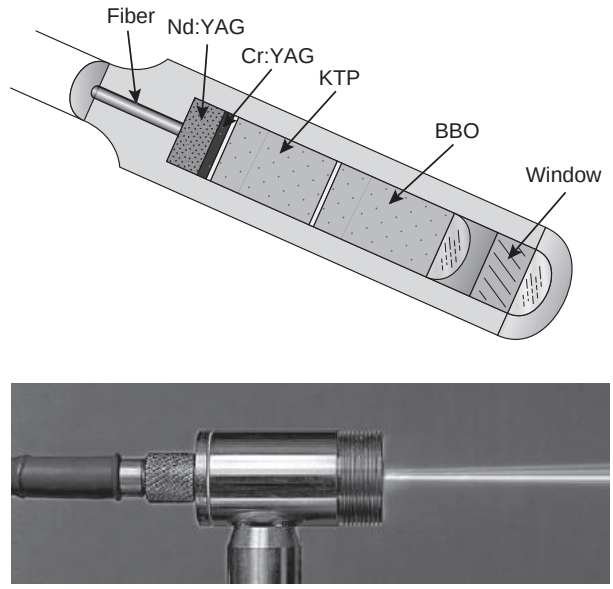

Fig. 5 Fiber-coupled 266-nm frequency-quadrupled passively Q-switched microchip laser: (top) schematic and (bottom) photograph of working device mounted on 12.7-mm-dia. post.

mixing, parametric conversion, and stimulated Raman scattering have been used with passively Q-switched microchip lasers for frequency conversion to wavelengths covering the entire spectrum from 213 nm to 8.1 μm in extremely compact optical systems.

Many applications for microchip lasers require harmonic conversion of the laser's infrared output. Because of its high peak intensity, the output of even low-average-power Nd:YAG/Cr^{4+}:YAG passively Q-switched microchip lasers can be efficiently harmonically converted by placing the appropriate nonlinear crystals near the output facet of the laser with no intervening optics, as shown in **Fig. 5**. With this approach, a 1-W-pumped 1.064-μm Nd:YAG/Cr^{4+}:YAG passively Q-switched microchip laser has been frequency converted to produce 7 μJ of second-harmonic (532-nm green), 1.5 μJ of third-harmonic (355-nm UV), 1.5 μJ of fourth-harmonic (266-nm UV), and 50 nJ of fifth-harmonic (213-nm UV) light at a typical pulse repetition rate of 10 kHz.

As an alternative to nonlinear frequency generation, the output of passively Q-switched microchip lasers and its harmonics have been used as a pump to gain switch miniature lasers in the infrared, visible, and ultraviolet portions of the spectrum. The high peak powers of Q-switched microchip lasers have also been exploited in fibers to generate white-light continua, and sub-200-fs pulses through a process that involves self-phase modulation, spectral filtering, and pulse compression.

For many applications the output power, or energy, produced directly by a microchip laser is sufficient. For numerous others, some amplification is required. A wide variety of amplifiers have been used in conjunction with microchip lasers, in systems that take advantage of the waveforms and near-ideal mode properties that they produce.

Applications

Within their range of capabilities, microchip lasers are the simplest, most compact, and most robust implementation of solid-state lasers. CW microchip lasers face strong competition from diode lasers and fiber lasers. Diode lasers are more efficient, smaller, and simpler, and are available at a greater variety of wavelengths. Fiber lasers also tend to be more efficient and can produce higher output powers in a fundamental transverse mode. To compete against either of these technologies, CW microchip lasers need to find a niche application that exploits their unique spectral properties. Nonetheless, the CW microchip geometry has established itself as a testing ground for newly developed gain media.

On the other hand, Q-switched microchip lasers provide capabilities that cannot be matched by semiconductor devices or fiber lasers. Semiconductor diode lasers have a very limited capacity to store energy, and their facets are damaged at very modest optical intensities. Fiber lasers have much longer cavities, which prevent them from producing short Q-switched pulses and make them susceptible to undesirable nonlinear effects at peak powers well below those demonstrated with Q-switched microchip lasers. Fiber amplifiers can be used to increase the peak power from pulsed semiconductor diodes, but this requires several stages of amplification, results in more complicated systems, and is still limited by fiber nonlinearities. Additionally, the amplifier output has interpulse amplified spontaneous emission that may be detrimental for some applications.

As a result of their small size, robust construction, reliability, and relatively low cost, coupled with their ability to produce energetic, diffraction-limited, Fourier-transform-limited, sub-nanosecond pulses, passively Q-switched microchip lasers have been embraced for applications in high-resolution time-of-flight three-dimensional imaging. Because their high peak output intensity allows for efficient nonlinear generation of ultraviolet light in very compact and reliable formats, they have also been well accepted

in the field of ultraviolet fluorescence spectroscopy, and are an integral part of numerous fielded spectroscopic instruments. In addition, passively Q-switched microchip lasers have made inroads in laser scribing and marking, laser-induced breakdown spectroscopy, and most recently laser ignition. The continued development of the technology will be driven by applications.

Further Reading

Zayhowski, J.J., 2013. Microchip lasers. In: Denker, B., Shklovsky, E. (Eds.), Handbook of Solid-State Lasers: Materials, Systems and Applications. Cambridge: Woodhead Publishing (Chapter 14).

Zayhowski, J.J., Welford, D., Harrison, J., 2007. Miniature solid-state lasers. In: Gupta, M.C., Ballato, J. (Eds.), The Handbook of Photonics, second ed. Boca Raton, FL: CRC Press (Chapter 10).

Supercontinuum Generation

James R Taylor, Imperial College London, London, United Kingdom

Introduction

Although nonlinear optical effects had been investigated in the 19th century, for example, Kerr in 1875 examined induced birefringence in isotropic media as a result of an applied DC electric field, showing it to be proportional to the field intensity, while 20 years later, Pockels demonstrated that the birefringence induced in piezoelectric crystals linearly depended on the applied DC field, however, it was not until just after the invention of the laser in 1960, particularly in its early pulsed or relaxation oscillation manifestations, that the electric field associated with the optical pulse was sufficiently intense to allow nonlinear optical conversion of the incident laser signal to be observed directly. Within a few years of the demonstration of the ruby laser by Maiman, the techniques of Q-switching and mode-locking were reported, which allowed the controlled generation of nanosecond pulses using the former method and picoseconds with the latter. Subsequently, over the following three decades, innovative mode-locking schemes were devised in various families of lasers to routinely allow the generation of controlled pulse durations of only a few femtoseconds, noting that in the visible around 600 nm a single optical cycle is approximately 2 fs.

Consequently, for relatively modest pulse energies, it is possible to generate peak powers in the focal region of a conventional convex lens exceeding a terawatt per square centimeter, with corresponding electric field strength over 10 MV per centimeter. In a transparent dielectric medium, electric field strengths of such magnitude give rise to a nonlinear response, in that the induced polarization has to be described by an expansion that includes higher order terms, such that

$$P = \varepsilon_0\left(\chi^{(1)}E + \chi^{(2)}E^2 + \chi^{(3)}E^3 + ...\right)$$

where P is the polarization, ε_0 is the permittivity of free space, and $\chi^{(n)}$ is the nth order susceptibility. The first term on the right hand side represents the linear response of the medium, which until 1960 was the most commonly encountered case, where the electric field strength E was relatively low and higher order terms could be neglected.

If one examines the second term in E^2 and considers an input optical field varying as $\sin\omega t$, where ω is the angular frequency of the radiation, an expansion of this term will give rise to components varying as $\sin 2\omega t$, a nonlinear response, leading to frequency doubling or so called second harmonic generation. This was the first laser-generated nonlinear process observed, using a simple focused pump scheme in crystal quartz. In media with a center of symmetry, the second order term vanishes to zero and second harmonic generation is generally not observed. The process of second harmonic generation is also not particularly relevant in the contribution to supercontinuum generation, where the more important effects arise from the third order term. Driven by at a fundamental field at $\sin\omega t$, the third order term gives rise to a component response at $\sin 3\omega t$, generating the third harmonic. Once again, this is of little impact for supercontinuum generation, however, other nonlinear contributions arise from the third order term, such as the optical Kerr effect or intensity dependent refractive index, four-wave mixing, modulational instability, stimulated Raman scattering, and stimulated Brillouin scattering, although the latter also plays a minimal role in supercontinuum generation. Through a judicious choice of system parameters, such as pump wavelength, pump duration, bandwidth, system dispersion, and power, a particular nonlinear process may be chosen to dominate or alternatively many of the processes can be selected to occur simultaneously.

Spectral Broadening and Early Supercontinuum Sources

In early realizations of pulsed, Q-switched and mode-locked solid state lasers, anomalous spectral broadening of the output, much greater than the expected gain linewidth of the transition, was widely reported upon. This was also observed in oscillator–amplifier configurations and in particular where catastrophic optical damage to the gain medium was simultaneously recorded. Spectral broadening was also observed through focusing in highly nonlinear liquids, such as CS_2. It was theoretically proposed that the nonlinear refractive index, both in the temporal and the spatial regimes, was responsible for these observations. If one considers the overall refractive index n_{tot}

$$n_{tot} = n_0 + n_2 I(x,t)$$

where n_0 is the refractive index as the intensity of the propagating field approaches zero, the low power case that is more commonly encountered and n_2 is the space and time intensity $I(x,t)$ dependent refractive index. For silica n_2 is approximately 3.2×10^{-20} m^2 W^{-1}, while in CS_2 it is approximately two orders of magnitude greater. In the spatial domain if one considers a Gaussian transverse mode profile propagating in a nonlinear medium, the intensity at the center is greater than in the wings. Consequently the total refractive index is greater centrally than in the wings of the mode, such that the wings will travel faster than the center. This causes the mode to collapse in on itself – to self-focus. It was observed that significant spectral broadening was observed beyond a critical power where this catastrophic collapse led to the formation of self-trapped filaments. Of course, these self-trapped filaments were highly irreproducible in space and in time, as were the associated spectra.

 doi:10.1016/B978-0-12-803581-8.09366-8

In the time domain, the accumulated phase difference on propagating over a distance L of a pulse of intensity $I(t)$ is given by $n_2kLI(t)$, where k is the wavenumber. Since a frequency change is simply the negative time derivative of phase and as n_2, k, and L are constants, the frequency shift is simply proportional to the negative time derivative of the pulse intensity, while assuming that the time response of the third order nonlinearity is ultrafast ($\sim$fs) compared to the incident pulses, as was first explained by Shimizu and labeled self-phase modulation (SPM). As a consequence of SPM, the front of a pulse is frequency down shifted, i.e., red shifted, while the rear is frequency up (blue) shifted and a quasi-linear frequency shift exists throughout the central region of the pulse. The spectral broadening experienced is proportional to the propagation length and the intensity. In the time domain, SPM plays a vital role in the generation of temporal optical solitons, where in the anomalously dispersive regime for a specific power a balance can exist between dispersion and nonlinearity such that a pulse can propagate without dispersive temporal broadening. In optical fiber, solitons play one of the most important roles in supercontinuum generation.

Despite many reports of anomalous spectral broadening in laser systems, the first recognized study and report of supercontinuum generation, although not thus designated, was carried out by Alfano and Shapiro in 1970 by focusing the frequency doubled output of a pulsed, mode-locked Nd laser into bulk samples of BK7 glass at a power density of $\sim$1 GW cm^{-2}. The resulting beam filamentation gave rise to a white light spectrum that covered the range from 400 to 700 nm, which was substantially greater in extent than anything previously reported. Four-wave mixing was proposed as the principal formation process. Although very simple in the experimental configuration deployed, because of the uncontrollability of the formation mechanism, the generated spectra were not particularly reproducible. Stable supercontinua covering the range from 190 to 1600 nm were achieved by focusing the output of an amplified 80 fs dye laser at 627 nm to a power density of 10^{13}–10^{14} W cm^{-2} into a 500 μm thick jet stream of ethylene glycol, however, the energy density in the extremes of supercontinuum were up to five orders of magnitude lower than that around the pump and the overall average power was low. Other liquids such as water, carbon tetrachloride, and various fluorocarbons have also been used. In bulk samples, self-focusing and SPM have been proposed and identified as the primary spectral broadening mechanisms, however, the process is somewhat complicated by the formation of an optical shock generated at the rear of the pump due to spatial and temporal focusing as well as self-steepening. Visible supercontinuum have also been generated by focusing amplified femtosecond dye laser pulses into cells of high pressure gases, such as Xe, H_2, and N_2 and in millimeter lengths of optical fibers, however, the large footprint of the pump laser/amplifier schemes together with bulk lens coupling to the samples restricted their application primarily to research laboratories.

Modeling of Supercontinuum Generation

The formation of optical solitons plays an underpinning role in supercontinuum generation. Essential for the generation process is the balance that arises between linear and nonlinear effects. This can happen both in the spatial domain, where the balance is achieved between self-focusing and diffraction or in the temporal domain where it is between SPM and dispersion. In optical fiber-based systems, temporal soliton formation is the dominant process, while, for example, in early experimental schemes where supercontinuum generation occurred through the focusing of a pulsed laser into a cell containing a nonlinear liquid or into bulk material, spatial soliton formation principally contributed. These simple systems can be described by the nonlinear Schrödinger equation (NLSE). In the case of a complex optical pulse envelope $U(x,t)$ propagating in a fiber, normalized such that $|U(z,\tau)|^2 = P(z,\tau)$, where $P(z,\tau)$ is the instantaneous power in Watts, z is the position in the fiber, and τ is a co-moving time frame, the NLSE can be written as

$$\partial_z U = -i\frac{\beta_2}{2}\partial_\tau^2 U + i\gamma|U|^2 U$$

where $\beta_2(\omega) = \partial^2\beta/\partial\omega^2$ and $\beta(\omega)$, the axial wavenumber, $= n_{\text{eff}}\omega/c$, where n_{eff} is the effective modal refractive index, ω the angular frequency and c the speed of light. The nonlinear coefficient $\gamma = n_2\omega/cA_{\text{eff}}$ (where n_2 is the nonlinear refractive index and A_{eff} the effective modal area). The first term on the right hand side of the NLSE above relates to the dispersive term and the second term is the nonlinear balancing contribution and the equation gives rise to solutions of the form

$$U(z,\tau) = \sqrt{P_0}\,\text{sech}\left(\frac{\tau}{\tau_0}\right)\exp\left(i\gamma\frac{P_0}{2}z\right)$$

where the soliton peak power $P_0 = |\beta_2|/\gamma{\tau_0}^2$ and τ_0 is the soliton duration. For powers launched into a fiber where $P = N^2P_0$, where N is an integer, high order soliton behavior is observed. In the early stages of propagation this leads to a rapid temporal compression, however, as the simple NLSE description is unable to adequately describe the evolution of such broad bandwidth pulses, additional terms are needed. For example, the effects of high order dispersion need to be included as excessive bandwidth is encountered, while broad bandwidth operation also gives rise to the need to consider Raman scattering effects, in particular self-Raman interaction or the so called soliton self-frequency shift, where short wavelength components in a pulse can provide Raman gain for the longer wavelengths. Historically, each of these modifications to consider each effect were introduced separately, however, all effectively reduce to the same equation, the generalized nonlinear Schrödinger equation (GNLSE), which is valid for pulse durations down to a couple of optical cycles. In the frequency domain it is given by

$$\partial_2\tilde{U} = i\left(\beta(\omega) - \frac{\omega}{v_{\text{ref}}}\right)\tilde{U} + i\frac{n_2\omega}{cA_{\text{eff}}}\mathcal{F}\left[\left(R \ast\!\ast |U|^2\right)U\right]$$

where $\tilde{U}(z,\omega)=\mathcal{F}[U(z,t)]$ and $\mathcal{F}$ is the Fourier Transform, v_{ref} a chosen reference velocity and $R(t)$ is the nonlinear response function. For silica glass $R(t)$ takes the form $R(t)=(1-f_r)\delta(\tau)+f_r h_R(t)$, an instantaneous Kerr contribution and a delayed Raman response described by h_R. This GNLSE includes the full modal dispersion, the Raman effect and the effect of self-steepening, which leads to optical shock formation. The equation can also be modified to include, for example, the frequency dependence of the effective area of the fiber deployed or third harmonic generation.

Another vitally important ingredient in the modeling of supercontinuum generation is the inclusion of the contribution of noise, this is particularly relevant when concerned with the process of modulational instability. Most commonly this is managed through the addition of quantum noise fluctuations to the initial conditions, through the addition of a field with a spectral power equivalent of one photon per mode with random phase. In addition the noise on the actual pump source needs also to be included and this is of particular relevance to cw pumped systems.

Supercontinuum Generation in Optical Fiber

Optical fiber is an ideal medium for the study of nonlinear optics. In bulk samples, the nonlinear interaction length is limited by the confocal parameter of the focusing lens, which is the order of a millimeter in the visible for a 10 μm waist. Driven by the needs of telecommunications, silica-based glass fibers have been produced with losses better than 0.2 dB km^{-1}. Consequently, signals can propagate tens of kilometers, confined to core diameters of a few microns without significant reduction in the signal intensity. Neglecting processes that require phase matching, since nonlinearity is a power times length process, in-fiber generation can have an enhancement of up to 10^6–10^7 compared to that in the bulk, or equivalently the peak power requirement to observe an effect is reduced by this factor and nonlinear effects can be observed at modest pump power levels.

A nitrogen laser pumped broad band dye laser ($\sim$15 nm) was used as the fundamental source for the first supercontinuum generation in optical fiber as reported by Lin and Stolen (1976). A power of 1 kW in a 10 ns pulse around 434 nm launched into a 20 m long multimode 7 μm diameter fiber generated a "white light" spectrum extending from 434 to 614 nm. The dominant contributing process was cascaded, stimulated Raman scattering and consequently the generated supercontinuum was only to the long or Stokes wavelength side of the pump. By 1978, the more common approach of using a pulsed Nd:YAG laser was used to produce a supercontinuum that extended from about 700 to 2100 nm, using 50 kW, 20 ns pulses from a Q-switched system to excite 315 m of multimoded fiber with a 33 μm core diameter. Again cascaded Raman generation was the principal generation mechanism. **Fig. 1** shows a typical supercontinuum generated in a standard 7 μm core silica fiber, with a dispersion zero around 1310 nm, using 80 ps pulses from a mode-locked Nd:YAG laser at 1060 nm.

The cascade of Raman orders is clearly seen in **Fig. 1** evolving sequentially from the pump at 1060 nm in this example, from first to fourth labeled 1S through to 4S, respectively. SPM and cross phase modulation from the pump and between the Raman orders can give rise to additional spectral broadening but the individual orders are clearly resolved, each separated by approximately 440 cm^{-1}. Beyond the dispersion zero of the fiber around 1310 nm, the fourth Stokes component is seen to evolve into a continuum. This soliton-Raman continuum is made up of randomly distributed femtosecond soliton pulses, with the evolution proceeding via modulational instability. Closely related to soliton generation and four-wave mixing, modulational instability results from the interplay of anomalous dispersion and the intensity dependent refractive index. Amplitude or phase modulations on an effective continuous wave background, for example, the femtosecond amplified spontaneous emission (ASE) noise fluctuations on a tens of picosecond pulse, exhibit an exponential growth rate that is accompanied by a spectral sideband evolution at a frequency separation from the carrier that is proportional to the optical pump power. For exponential growth, the upper and

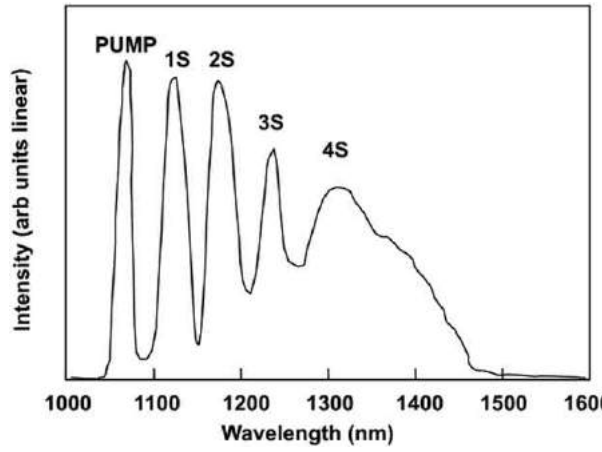

Fig. 1 Representative supercontinuum spectrum generated in a standard telecommunications fiber pumped by the 80 ps, mode-locked pulses from a Nd:YAG laser. Reproduced from Alfano, R.R. Supercontinuum Laser Source the Ultimate White Light, third ed. New York, NY: Springer. ISBN 978-1-4939-3324-2.

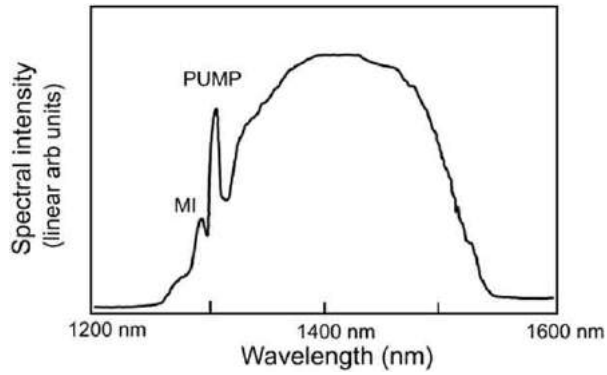

Fig. 2 Modulational instability initiated high average power soliton-Raman supercontinuum generated in 200 m of standard silica fiber at an average power of 1 W. Reproduced from IEEE Journal of Quantum Electronics 24, 332–340 (1988).

lower sideband frequency separation from the carrier must be less than a critical frequency given by $(4\gamma P_0/\beta_2)^{1/2}$, where P_0 is the incident power, γ is the nonlinear coefficient and β_2 is the group velocity dispersion parameter, while the maximum sideband growth occurs at a frequency separation of $(2\gamma P_0/\beta_2)^{1/2}$. The process is usually self-starting and from system noise but can also be seeded using an external signal lying within the gain bandwidth. Modulational instability plays an underpinning role in the initiation, amplification, and evolution of femtosecond pulses, which with subsequent self-Raman interaction and collisions form the basis of supercontinuum generation under picosecond, long pulse, or even cw pumping.

The role of modulational instability can be seen in **Fig. 2**, which is a soliton-Raman continuum generated from a cw mode-locked Nd:YAG laser generating 100 ps pulses at 1318 nm, with a launched power of approximately 1 W ($\sim$ 100 W peak power) in 200 m of silica fiber with a dispersion zero at 1290 nm. Significant depletion of the pump is observed and the anti-Stokes component of the originating modulational instability signal around 1295 nm is apparent. Following the rapid evolution of femtosecond solitons, the so-generated pulses exhibit self-Raman interaction (soliton self-frequency shift), where the bandwidth of the short pulse soliton is sufficient for the short wavelength component to provide significant Raman gain for the long wavelength components. This gives rise to a continuous long wavelength spectral shift with propagation length and one that also increases with pump power. The effects of noise amplification and soliton–soliton collisions also contribute to the spectral broadening and to the random temporal soliton structure of the output. On spectral filtering such a continuum, pulses as short as 50–80 fs can be selected but with large temporal jitter.

It should also be noted that only significant continuum generation is observed to the long (Stokes wavelength side) of the pump.

Solitons and Photonic Crystal Fiber

Although proposed by Hasegawa in 1973 and most certainly they were generated in the fiber-based supercontinua published by Lin *et al.* (1978), the optical soliton was not demonstrated unequivocally until 1980 by Mollenauer and colleagues in a series of innovative experiments that characterized their unique properties. The reason for the delay was two-fold. Only by the late 1970s were adequate lengths of low loss single mode optical fiber available and since the minimum dispersion zero wavelength of conventional silica-based fiber is 1270 nm, significant effort had been placed on the development of tuneable picosecond pulse sources to instigate soliton operation above that wavelength in the anomalously dispersive regime.

By the mid-1980s many laboratories were directing their efforts toward fiber-based sources of ultrashort pulses utilizing soliton effects for extreme compression. The power P_0, required to establish a fundamental soliton that will propagate over its characteristic nonlinear length without change in the pulse shape and duration is fixed and proportional to $A_{\text{eff}}\lambda^3 D/\tau^2$, where A_{eff} is the effective core area of the fiber, λ the wavelength of the radiation, D the group delay dispersion at that wavelength, and τ the pulse duration. If one launches a pulse power N^2P_0, where N is an integer, a high order soliton is established, which is a nonlinear superposition of N fundamental solitons. It has been demonstrated both theoretically and experimentally that on propagation, pulse narrowing and periodic splitting and recombination takes place as a result of the periodic interference between these N solitons. An optimal compression ratio of $4.1N$ is achieved with a corresponding increase in the associated spectral width. For the lower orders, for example, $N=2$ or 3, etc., periodic breathing of the compression and restoration of the original pulse width is observed on propagation, however, for relatively high order solitons, after extreme compression, instability occurs through self-Raman effects and higher order dispersion perturbations and break-up of the high order soliton into numerous single solitons occurs. Originally termed colored soliton generation, the process has been renamed soliton fission. Using higher order soliton compression, pulses as short as 18 fs, only four optical cycles, have been generated centered around 1318 nm with an associated spectral half intensity width of about 100 nm and baseline width around 300 nm.

It was not possible to observe soliton effects in conventional silica-based fibers using the more common short pulse sources, such as the Nd-doped glass and crystal lasers, Yb fiber lasers, or the Ti:sapphire laser. This was to dramatically change with the

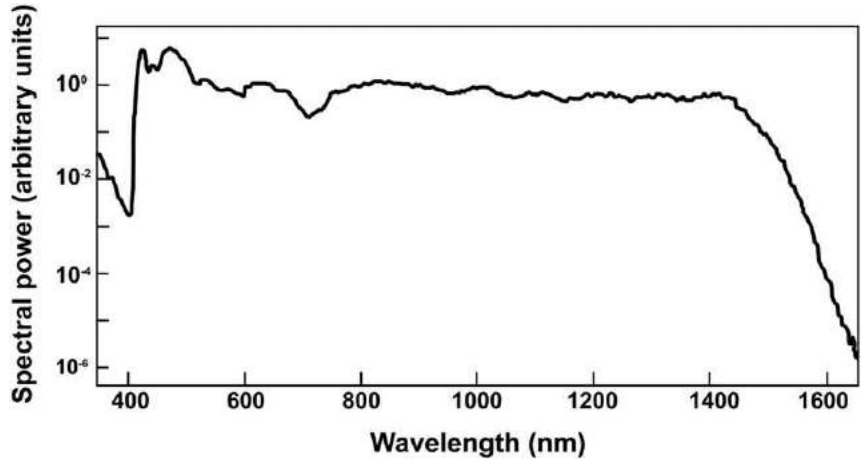

Fig. 3 Supercontinuum generated in a photonic crystal fiber with a dispersion zero at 770 nm, pumped by the 8 kW, 100 fs pulses of a Ti: sapphire laser at 790 nm. After Ranka, J.K., Windeler, R.S., Stentz, A.J., 2000. Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800 nm. Optics Letters 25, 25–27.

introduction of the photonic crystal fiber (PCF) by the Russell group at the University of Bath in 1996. Through the adjustment of the pitch and diameter of the crystalline-like hole structure around the core, it became possible to precisely engineer the dispersion zero wavelength even into the visible spectral region. Enhanced nonlinearity can also be achieved through the controlled manufacture of smaller effective mode field diameters, while fibers can also be made to operate with single mode characteristic throughout extended spectral regions from the visible to the near infrared, in fact covering the complete window of transmission of silica. This culminated in 2000 in a report by Ranka and colleagues on the generation of a relatively modest average power supercontinuum extending approximately from 400 to 1600 nm (see **Fig. 3**) from a 75 cm length of photonic crystal fiber with a dispersion zero at 770 nm, pumped by the 8 kW, 100 fs 790 nm pulses from a mode-locked Ti:sapphire laser. This result was to rejuvenate research into supercontinuum generation, lead to renewed studies of nonlinear optics in fiber and ultimately led to the development of an extremely successful commercial product.

Femtosecond Pulse Pumped Supercontinua

A remarkable feature of the femtosecond laser pumped supercontinuum shown in **Fig. 3** is the relative flatness of spectral intensity, however, it must be remembered that this results from the temporal integration of many millions of spectra generated by the individual pulses of the pumping mode-locked signal. Also of particular note is the complete depletion of the pump signal around 790 nm and importantly the extension of the generated supercontinuum to the short wavelength side of the pump, effectively covering the complete visible region.

The principal mechanism initiating the generation process is the temporal compression of the high order soliton that takes place over a length scale given approximately by $\tau_0^2/N\beta_0$, where N is the soliton order, τ_0 the duration of the input soliton, and β_0 the dispersion parameter. This is accompanied by inherent spectral broadening. In **Fig. 4**, which is a theoretical simulation of the spectral evolution with length of an input, $N \sim 9$, 10 kW peak power, 50 fs soliton at 835 nm propagating in a PCF with a dispersion zero in the region of 780 nm, the result of the rapid temporal compression is evidenced within an initial 1 cm of propagation.

Pumping in the region of anomalous dispersion and following rapid compression, the soliton spectrum ingresses the normal dispersion regime and dispersive waves feed off. Beyond the point of maximum compression the system is affected by system instabilities that inhibit the breathing and periodic restoration of the high order soliton. Modulational instability, higher order dispersion and the soliton self-frequency shift result in the spectral and temporal fragmentation into numerous solitons, with the associated spectrum being further complicated by the effects of soliton collisions, four-wave mixing, for example, between the soliton and the dispersive wave, and cross phase modulation. Despite the apparent complications of the generation process, the physical processes and dynamics of formation are very well understood with theoretically predicted spectra agreeing exceptionally well with experimental measurement.

Cross phase modulation between solitons and dispersive waves in the region of the dispersion zero induces a chirp on the dispersive wave, with the rear of the dispersive wave being frequency up shifted and as a consequence of propagation in the normal dispersion regime is de-accelerated and retarded. The short pulse soliton on the other hand experiences the soliton self-frequency shift, frequency down shifting via the Raman gain process. As operation is in the anomalously dispersive regime the soliton too is de-accelerated, consequently overlapping temporally with the up shifted dispersive wave. This process of dispersive wave generation and trapping by solitons can lead to significant power transfer to the dispersive wave and is the principal

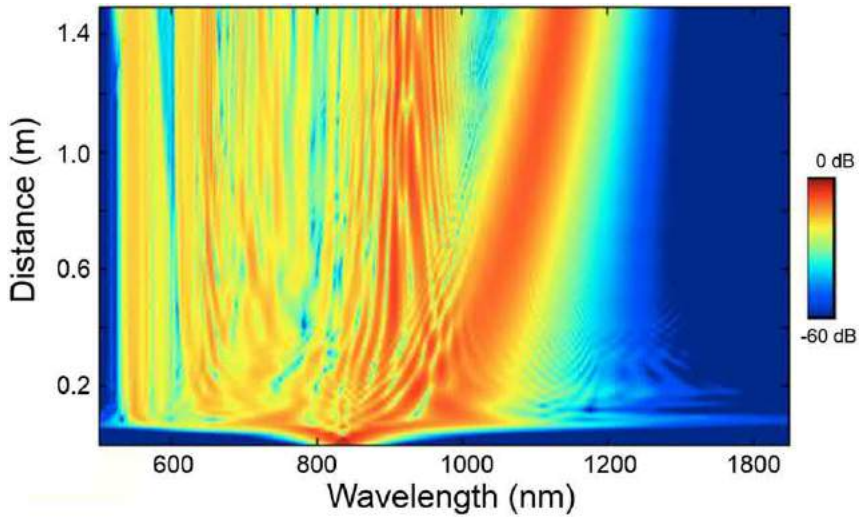

Fig. 4 Simulation of supercontinuum evolution with fiber length through high order femtosecond soliton compression and subsequent soliton fission. Courtesy J.C. Travers unpublished.

mechanism in the short wavelength extension of supercontinua. In Fig. 4, the result of soliton self-frequency shift is apparent. Following fragmentation of the highly compressed pulse after approximately 1 cm, a continuous long wavelength evolution with length can be seen. In many fiber structures, in the anomalously dispersive regime the group velocity dispersion increases with increasing wavelength. As a consequence, the required soliton power increases ($P_0 \propto D/\tau_0^2$) and the soliton accommodates for its now inherently low power by increasing in duration, however, with increased duration the effect of the soliton self-frequency shift is reduced ($\Delta\nu \propto \tau_0^{-4}$). Thus, an increasing anomalous dispersion with wavelength will inhibit the soliton self-frequency shift, long wavelength extension and consequently short wavelength extension of the supercontinuum.

Following rapid compression and soliton fission, signal noise can have a dramatic influence on the seeding of the modulational instability process. If intense femtosecond scale structures are present in the noise signature these can be rapidly amplified to soliton powers and can enhance the long wavelength self-frequency shifted edge of the supercontinuum, with the shortest most intense seeded signals giving the greatest spectral shift. Noise will strongly affect the evolving spectrum and considerable variation occurs in the spectra of the individual supercontinua generated by each individual pulse of the exciting mode-locked train – although the subtly of this is lost in the integrated spectra as is normally recorded (see Fig. 3). This role of noise in the seeding of the long wavelength component of the supercontinuum has been likened to the generation of "rogue waves" in hydrodynamics.

Beyond the point of optimum pulse compression in high order soliton driven supercontinuum generation, the effects of higher order dispersion, self-Raman interaction, soliton fission, modulational instability, soliton–soliton collisions, and noise lead to spectral and temporal instability and spectral coherence is lost. To minimize the effects of noise and instability it is best to utilize fiber lengths cut to the optimal high order soliton compression length. Soliton effects can of course be totally negated by operating in a region of normal dispersion, such that the process of SPM dominates the spectral broadening process, however, for equivalent pump powers, the spectral coverage is generally reduced compared to that attainable with solitons.

Picosecond Pulse Pumped Supercontinua

Although providing adequate spectral coverage under femtosecond solid state laser pumping, the necessity of bulk coupling into PCF using lenses does not lend the technique to high stability in a relatively hostile environment and diminishes potential commercial impact and routine application. In addition, the use of femtosecond pulses also inhibits the average power scaling of the spectral power density of the generated supercontinua, since nonlinearity and dispersion in the amplifiers affects the efficiency of the amplification process. With the development of picosecond, master oscillator power fiber amplifier (MOPFA) configurations, primarily based upon picosecond pulse seeded Yb-doped schemes operating around 1060 nm, power scaling to hundreds of watts average power in compact, all-fiber geometries was possible, while it was also possible to completely fiber integrate the MOPFA and the PCF to allow high stability and reproducibility. Initial configurations utilizing average pump powers of a few watts allowed a minimum spectral power density of 1 mW nm^{-1} to be achieved throughout the spectral region 500–2000 nm. Since then, technological improvements in pump source technology have allowed the supercontinuum spectral power density to be increased to greater than 100 mW nm^{-1}. Fig. 5 shows the supercontinuum generated in 30 m of PCF with a zero dispersion wavelength at 1040 nm pumped by 10 ps, 10 kW pulses at 1058 nm from an Yb fiber laser system.

Since the fundamental soliton power is proportional to τ^{-2}, where τ is the pulse duration, it may be thought that the process of high order soliton fission would play an important role in the development of the associated supercontinuum, however, the

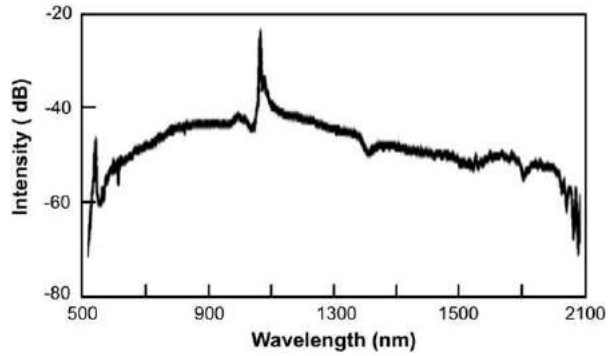

Fig. 5 Experimentally measured supercontinuum generated with 10 ps pulse pumping at a peak power of 10 kW in 30 m of photonic crystal fiber (PCF) with a zero dispersion wavelength of 1040 nm.

associated soliton period or nonlinear length scales as τ^2 and is many orders of magnitude greater than the few meters generally employed to observe substantial supercontinuum generation, hence soliton effects are not initially observed. For picosecond pumping in the region of the dispersion zero wavelength, modulational instability and four-wave mixing provide the principal initiation mechanisms. Modulational instability leads to the rapid break up of input picosecond scaled pulses into femtosecond scale subpulses, which are rapidly amplified to soliton powers. The modulational instability process can be significantly enhanced if the input pulses exhibit amplitude or phase fluctuations on their profile, which lie within the gain bandwidth of the modulational instability process and this noise can play an important role in the seeding of modulational instability and enhanced evolution of soliton structures. Subsequent four-wave mixing and soliton–soliton collisions lead to further spectral broadening, while the soliton self-frequency shift and soliton–dispersive wave trapping gives enhanced long and short wavelength extension respectively, as was described previously.

In early experimental realizations using pumping around 1060 nm in fibers with a dispersion zero wavelength around 1040 nm it was observed that the short wavelength extension of the supercontinua did not extend below about 550 nm, see, for example, **Fig. 5**. This characteristic, initially and surprisingly, was relatively independent of the pump power and or the fiber length employed, but can be explained simply by imagining the requirement to match the group velocity of the short and long wavelength components of the supercontinuum in the soliton–dispersive wave interaction. To the short wavelength side, dispersion is dominated by the host material, while to the long wavelength side the waveguide plays a dominant role. A group velocity match is achieved straddling the dispersion zero at a fixed pair of wavelengths. As a femtosecond scaled soliton experiences the soliton self-frequency shift it moves to a longer wavelength, trapping with it the dispersive wave component that is locked to the corresponding short wavelength with a matching group velocity. In silica-based fibers, the waveguide and material losses increase as the wavelength approaches 2 μm and as the loss increases soliton propagation can no longer be sustained because of energy loss and associated temporal broadening. Therefore it is the limitation on the long wavelength operation that through the group velocity matching condition restricts the short wavelength extent of the trapped dispersive waves in the supercontinuum spectrum. **Fig. 6** shows a theoretical simulation of a spectrogram – the spectral distribution with time of a supercontinuum generated after 0.5 m of fiber by a 10 kW peak power 3 ps pulse at 1060 nm launched just above the dispersion zero wavelength 1040 nm. Discrete soliton structures are observed in the anomalously dispersive regime together with the corresponding trapped dispersive wave structures in the normally dispersive regime. In the temporal domain the overall format is noise-like with numerous randomly distributed intense soliton structures, which are also clearly identifiable in the spectrogram. The overall spectrum, which extends from about 550 to 2000 nm and is the result of a single input picosecond pulse is noisy and structured. Only with the integration of many pulses from the driving mode-locked laser source, does a supercontinuum take on the smooth structure representative of **Fig. 5,** which is more commonly encountered.

Several techniques have been introduced to extend supercontinua spectra below 500 nm using 1 μm pumping. The first employed two concatenated PCFs. The first with a zero dispersion around 1040 nm and the second around 780 nm. When the continuum in the first fiber extends to the region around 780 nm the radiation around these wavelengths can successfully act as the pump in the second fiber, generating soliton-like structures around the dispersion zero and the process of soliton–dispersive wave trapping extends the spectrum. Since the shift in the zero dispersion of the second fiber is offset to shorter wavelengths the group velocity matching is similarly offset allowing matching to substantially shorter wavelengths. Pumping the fiber with a short zero dispersion wavelength alone would not give the same effect as the pump is too far away and any generated soliton structures would not give rise to dispersive waves in the normal dispersion regime. In such a case Raman would dominate the generation process, giving rise to a continuum that extended solely to the Stokes side of the pump pulse.

The technique was further extended using a length of tapered PCF, such that a continuous shift of the dispersion zero was produced over the length of the fiber. Such a structure enhances the group velocity matching condition, in addition, the tapered structure also leads to a decreasing group velocity with length. This leads to a de-acceleration of the soliton, equivalent to the action of the soliton self-frequency shift which is essential for soliton–dispersive wave interaction. Consequently, the tapered structures enhance the short wavelength achievement and supercontinuum generation has been demonstrated in 1.5 m long tapers

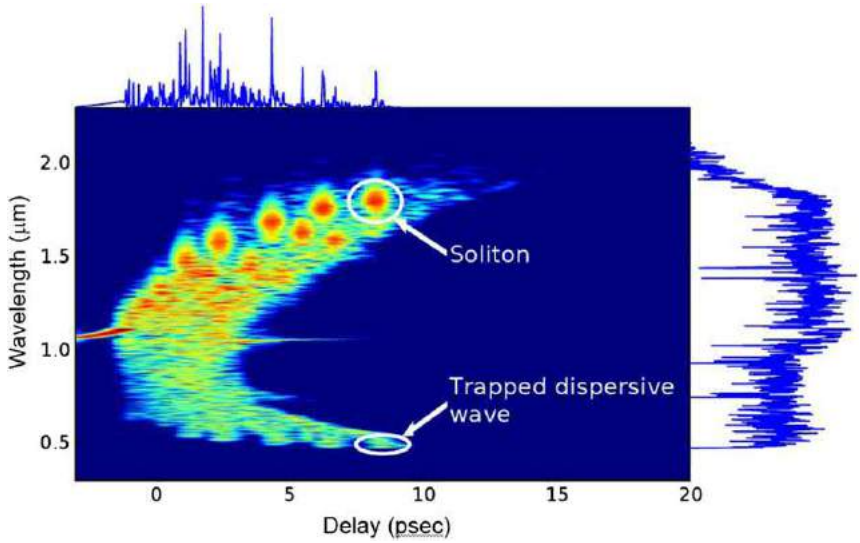

Fig. 6 Theoretical simulation of a spectrogram of a supercontinuum generated in 0.5 m of photonic crystal fiber (PCF) with a zero dispersion at 1040 nm pumped by 10 kW, 3 ps pulses at 1060 nm. The temporal structure of the output is shown above and the generated single shot supercontinuum spectrum is illustrated on the right. Also highlighted is the process of soliton–dispersive wave trapping.

pumped at 1060 nm from 320 to 2300 nm, effectively covering the complete window of transmission of silica fiber, with up to 2 mW nm^{-1} in the ultraviolet.

Similar operation can also be obtained in conventional PCF structures by modifying the waveguide structure, such that core resembles a few micron silica strand surrounded by air, by producing a PCF with a large hole pitch and a high air fill fraction. Operation from below 400 to 2500 nm has been shown in single constant core fibers.

Continuous Wave Pumped Supercontinua

With advances in broad stripe semiconductor pump laser technology, double clad doped fiber manufacture and innovative procedures for multimode combination of pumps, single transverse mode operation of cw Yb-fiber lasers at the 10 kW level has been demonstrated, however, rather modest average powers are quite sufficient to surprisingly observe supercontinuum generation in relatively short lengths of both photonic crystal and conventionally structured silica-based fibers. The process proceeds through the mechanism of modulational instability and consequently, the pump radiation wavelength must be in the region of anomalous dispersion. In the temporal domain, the intensity profile of the nominally cw pump radiation exhibits a noise-like structure arising from mode beating of the large number of randomly phased modes lying under the lasing spectral profile of the pump. The shortest temporal feature is given by the inverse of the frequency bandwidth of the laser radiation and this noise can enhance or seed the modulational instability process as has been described above. Through this mechanism the amplitude or phase perturbations rapidly evolve into fundamental soliton structures of femtosecond duration. Once established, these further exhibit the soliton self-frequency shift, extending the long wavelength extent of the supercontinuum. As soliton dynamics play a pivotal role in the overall generation process, it can be envisaged that the bandwidth, or related temporal structure, of the pump source is an important consideration in the generation process. For example, with a narrow bandwidth, the related temporal structure will be long. This means that the associated soliton lengths can be excessive and so soliton characteristics will not be established in the short fiber lengths utilized for supercontinuum generation, while long duration solitons would also not exhibit efficient self-Raman interaction. Likewise, for broad spectra, the associated structures are short, consequently, the required powers for soliton generation are excessively high and in this case solitons would be difficult to establish in the fibers. Empirically it can be argued and it has been shown theoretically that an optimum bandwidth exists for cw pumped supercontinuum generation.

The overall generation process is very similar to that using long picosecond pulse pumping. When the pump wavelength is far from the zero dispersion wavelength the generated supercontinuum is dominated by soliton-Raman shaping and the soliton self-frequency shift, with the generated radiation solely lying at Stokes wavelengths from the pump. This can be seen in trace 2 of Fig. 7, which shows the continuum generated by a 170 W cw signal at 1060 nm in a PCF with a dispersion zero wavelength at 830 nm and pump depletion is also evident. A spectral power density of 100 mW nm^{-1} was possible and the supercontinuum extended to above 2000 nm, but no short wavelength components were generated.

Similar to picosecond pulse pumping, when the generated short pulse soliton structures arising from the modulational instability process are in the region of the dispersion zero wavelength ingression to the normally dispersive region and the action

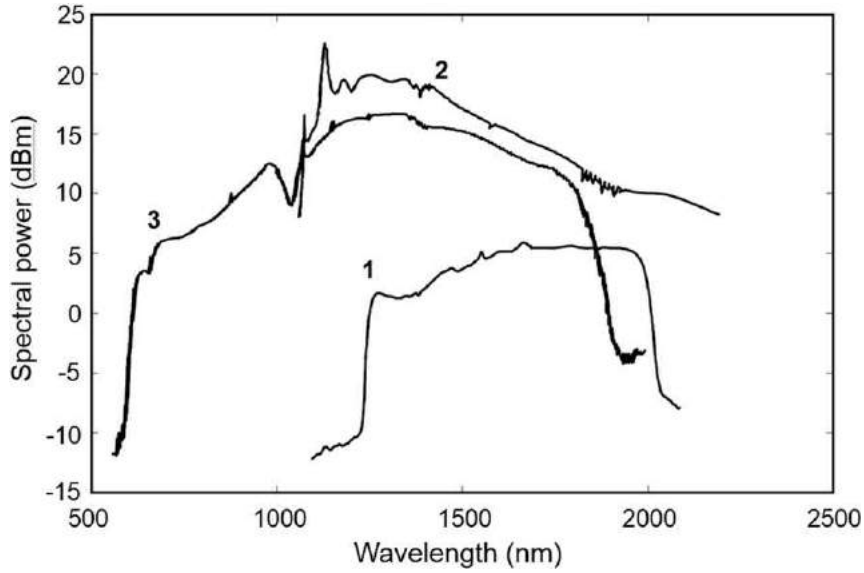

Fig. 7 Representative supercontinua generated under continuous wave (cw) pumping. (1) Conventionally structured highly nonlinear fiber pumped at 1.55 μm. (2) Photonic crystal fiber (PCF) with a dispersion zero at 830 nm pumped by 170 W at 1060 nm. (3) PCF with a dispersion zero at 1050 nm pumped by 230 W at 1060 nm.

of soliton–dispersive wave coupling leads to significant expansion of the supercontinua to shorter wavelengths. This process under cw pumping was first demonstrated in a 600 m length of conventionally structured highly nonlinear silica fiber with an effective core diameter of 2.1 μm and a zero dispersion wavelength of 1553 nm, which was excited by an Er-doped fiber-based ASE source at 1560 nm, amplified to 10 W average power. Plot 1 in **Fig. 7** shows the generated supercontinuum, which is marked by its relative spectral flatness, which is the result of the integration of the noise-like spectra, depletion of the pump and spectral expansion to wavelengths shorter than the pump. By employing a similar technique only pumping a 50 m length of a PCF with a zero dispersion at 1050 nm pumped by 230 W at 1060 nm, extension to the visible region was achieved with cw pumped supercontinua, as represented by plot 3 in **Fig. 7**. Through optimizing fiber structures and employing tapered geometries it has been possible to achieve operation down to 470 nm and with considerably lower pump power requirement of only 40 W under cw pumping.

A characteristic of the CW pumped supercontinua is the smoothness of generated spectra and this again is a result of the integration of the highly noisy spectra. This can be seen in **Fig. 8** which is a theoretical simulation of a spectrogram generated after 25 m propagation in a PCF with a zero dispersion at 1050 nm, pumped by 170 W at 1060 nm from an Yb fiber laser. In the figure, the one to one correspondence of the intense self-frequency shifted solitons to long wavelengths and the trapped dispersive waves to shorter wavelength can be clearly seen. The resultant spectrum shown to the right extends from 600 to 1800 nm, exhibits a relatively smooth structure while in the time domain, the output shown at the top of the figure exhibits a noise-like structure dominated by intense, randomly positioned solitons of varying duration and intensity. Power scaling has allowed spectral power densities in excess of 100 mW nm^{-1} to be obtained from cw pumped systems.

Wavelength Extension

The introduction of PCF and its integration with MOPFA assemblies underpinned the renaissance in both experimental and theoretical studies of supercontinuum generation, which subsequently led to the commercial development of rugged compact configurations that have found diverse application. To date, silica-based fibers have dominated the commercial product such that the spectral range from 320 to 2400 nm is covered, effectively the transmission window of silica. The average power capability has also been scaled such that in excess of 100 W has been achieved with both picosecond and continuous wave pumping. For many applications in spectroscopy, extension, particularly to the mid-infrared is desirable.

One possibility is to use fibers that are heavily doped with germanium. Beyond 2000 nm germanium oxide-based fibers exhibit a lower loss than silica, although it still is quite substantial, exceeding in the range of 100 dB km^{-1}, however, germanium oxide also exhibits a Raman gain coefficient that is nearly one order of magnitude greater than silica. As has been shown, the long wavelength extension in supercontinuum generation is a consequence of self-Raman interaction – the so called soliton self-frequency shift. As a result of the higher Raman gain coefficient, shorter gain lengths can be used, hence propagation losses are also reduced. Using pumping from a modestly powered sub-picosecond Tm-fiber laser-chirp pulse amplifier configuration, operation beyond 3000 nm has been achieved in these relatively robust Ge doped ($\sim$75% GeO_2) fibers. In order to further extend sources to the interesting mid-infrared "molecular fingerprint" region, it is essential to make use of the soft glasses, such as the fluorides or the

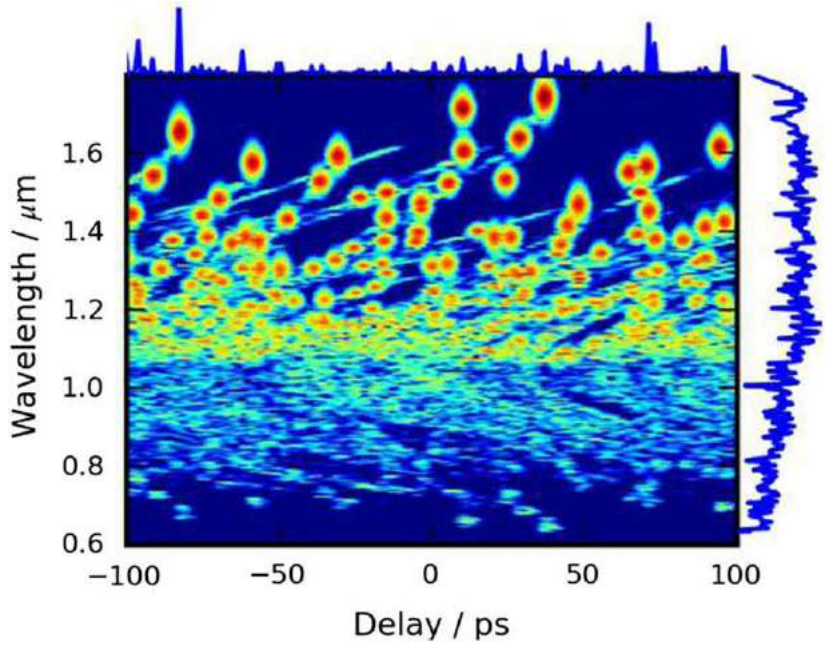

Fig. 8 Simulated spectrogram of a cw pumped supercontinuum generated after 25 m propagation in a photonic crystal fiber with a dispersion zero at 1050 nm pumped by 170 W at 1060 nm. A snapshot of the temporal output is shown to the top, while the associated supercontinuum is shown on the right.

chalcogenides, while to operate in the extreme ultraviolet, gas-filled, hollow-core fiber structures need to be utilized as the nonlinear medium.

The manufacture of single mode fluoride fibers of conventional structure is a well-established technology, with a material transmission window extending from 200 to 8000 nm, however, waveguide losses can significantly reduce the upper wavelength operational limit. In early work ZrF_4–BaF_2–LaF_3–AlF_3–NaF (ZBLAN) fiber pumped by the 1.6 MW, 900 fs pulses from a mode-locked Er-fiber laser at 1550 nm allowed supercontinuum generation from around 1400 to 3400 nm. Average power scaling has been undertaken with in excess of 10 W average power obtained in a fluoride fiber-based supercontinuum extending from 800 to 4000 nm. Through pumping with peak powers of 50 MW, obtained from a femtosecond optical parametric amplifier system at 1450 nm and a 2 cm length of ZBLAN fiber, spectral coverage from 350 to 6280 nm has been achieved. In the past few years, fluoride-based PCF has been produced using the stack and draw technique rather than by extrusion which had previously been used for the soft glasses and this has allowed submicron structures to be utilized to enhance the power density, permitting spectral coverage in supercontinua extending from 400 to 2400 nm in 4.3 cm fiber samples when pumped at 1040 nm, with the potential for power dependent further spectral extension.

The soft glasses exhibit nonlinear coefficients that are many orders of magnitude greater than silica, consequently for similar peak power pump pulses substantially shorter lengths of fiber are required to establish continuum generation. Generally, the soft glass fibers have not been demonstrated to handle the high operational average powers that their silica-based analogues have, which generally favors pumping with femtosecond pulses and consequently short fiber lengths. Most commonly, these host fibers have been manufactured by extrusion techniques. An 8-mm-long tellurite fiber pumped by a femtosecond optical parametric oscillator at 1550 nm generated a continuum from 789 to 4870 nm, while a 4 cm SF6 PCF pumped by 60 fs pulses at 1060 nm produced a spectrum spanning from 600 to 1450 nm.

Chalcogenide fibers have successfully demonstrated their potential for operation in the 2–12 μm range. Conventionally structured As-S fiber pumped by 100 fs pulses at 2500 nm exhibited operation from 2000 to 3600 nm. In a 9 cm length of suspended core chalcogenide fiber pumped at 3500 nm, a supercontiuum extending from 2000 to 6000 nm was reported. Through pumping at 6.3 μm with the 100 fs pulses derived through difference frequency generation of pump and signal pulses derived from an optical parametric amplifier, albeit at an average power of 0.75 mW, a supercontinuum extending from 1.4 mm to 13.3 μm has been achieved in an 85-mm-long step index chalcogenide fiber. Research is on-going to simplify the pump and generation process, as well as to scale the average power levels. Chalcogenide rib waveguides have also been successfully employed, generating spectra coving the 2–10 μm range.

In the ultraviolet, silica PCF has permitted the generation of 280 nm radiation in a supercontinuum, however, practical application and long term operation is prevented by photo-damage of the glass. ZBLAN PCF, however, has allowed operation down to 200 nm without any reported short wavelength radiation induced damage problems. To generate supercontinua below this, the glass path has to be removed and gas used as the nonlinear medium. Most commonly PCF with Kagomé structure is used with hollow-core diameters in the 10–50 μm range. The associated waveguide dispersion of these structures is anomalous with a

relatively low dispersion slope. Since the dispersion of the gas fill at several atmospheres pressure is normal, the net dispersion can be tuned and the position of the zero dispersion wavelength selected through variation of the gas pressure. As a result, a small anomalous dispersion can be obtained in the visible/near infrared, and a zero dispersion in the near-UV. High peak power pumping can be achieved without optical damage problems and high order soliton dynamics observed under femtosecond pumping with amplified Ti:Sapphire laser systems. Over a 15 cm length of a 28 μm diameter Kagomé-PCF filled with hydrogen at 5 bar pressure, 2.5 μJ, 35 fs pulses at 800 nm generated a supercontinuum extending from 124 to 1200 nm. In He, at 28 bar pressure, operation down to 110 nm has been observed.

To date, however, only the compact visible/near infrared supercontinuum PCF-based sources, pumped primarily by Yb-MOPFA configurations have had widespread commercial success and have underpinned a broad, turn-key applications base.

See also: Nonlinear Optics

References

Lin, C., Nguyen, V.T., French, W.G., 1978. Wideband near IR continuum (0.7–2.1 μm) generated in low loss optical fibers. Electronics Letters 14, 822–823.
Lin, C., Stolen, R.H., 1976. New nanosecond continuum for excited-state spectroscopy. Applied Physics Letters 28, 216–218.

Further Reading

Agrawal, G.P., 2001. Nonlinear Fiber Optics. San Diego, CA: Academic Press.
Agrawal, G.P., 2008. Applications of Nonlinear Fiber Optics. San Diego, CA: Academic Press.
Alfano, R.R. (Ed.), 2016. The Supercontinuum Laser Source – The Ultimate White Light, third ed. New York, NY: Springer.
Alfano, R.R., Shapiro, S.L., 1970. Emission in the region 4000 Å to 7000 Å via four-photon coupling in glass. Physical Review Letters 24, 584–587.
Dudley, J.M., Genty, G., Coen, S., 2006. Supercontinuum generation in photonic crystal fiber. Reviews of Modern Physics 78, 1135–1184.
Dudley, J.M., Taylor, J.R. (Eds.), 2010. Supercontinuum Generation in Optical Fibers. Cambridge: Cambridge University Press.
Genty, G., Coen, S., Dudley, J.M., 2007. Fiber supercontinuum sources. Journal of the Optical Society of America B 24, 1771–1785.
Mollenauer, L.F., Gordon, J.P., 2006. Solitons in Optical Fibers. San Diego, CA: Academic Press.
Ranka, J.K., Windeler, R.S., Stentz, A.J., 2000. Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800 nm. Optics Letters 25, 25–27.
Rulkov, A.B., Vyatkin, M.V., Popov, S.V., Taylor, J.R., Gapontsev, V.P., 2005. High brightness picosecond all-fiber generation in 525–1800 nm range with picosecond Yb pumping. Optics Express 13, 2377–2381.

Infrared Transition Metal Solid-State Lasers

Kenneth L Schepler, University of Central Florida, Orlando, FL, United States

Motivation

Broadband tunability of laser sources is of interest for a number of applications where broadband emission, wavelength selection, or wavelength scanning is needed. The uncertainty principle states that the product of the pulse width and bandwidth ($\Delta t \Delta \nu$) of a laser pulse has a minimum value. This requires that ultrashort pulses have broad bandwidth. Thus Ti^{3+} transition metal ions doped into a sapphire crystal have an emission bandwidth of 128 THz which could, in theory, be modelocked (ML) to produce a 3.4 fs pulse; 5 fs pulses have been demonstrated. Active remote sensing requires spectral scans over large frequency ranges and provides high signal to noise ratios with detection of multiple molecules of interest. Molecules with carbon–ligand bonds have vibration–rotation transition energies that correspond to wavelengths throughout the infrared spectral region. The atmosphere also has broad windows of infrared transmission (Fig. 1) as well as molecular constituents like water, carbon dioxide, carbon monoxide, etc. with optical transitions in the infrared region. Thus broadband infrared laser sources are needed for atmospheric remote sensing, molecular spectroscopy, medical sensing/detection of biomolecules, surgery, dentistry, gas leak detection, target designation/recognition, countermeasures to heat seeking missiles, and many others.

Early Transition Metal Lasers

The first demonstrated laser (1960), a ruby laser, employed a radiative transition between two electronic energy levels in Cr^{3+}, a transition metal ion doped into alumina (Al_2O_3). However, this transition metal laser had a fixed (not tunable) wavelength of 693.4 nm. There were also some early investigations of Ni^{2+} (1963) and Co^{2+} (1964) as lasing ions but operation required cryogenic cooling and tuning performance was poor due to excited state absorption (ESA). But the early transition metal lasers actually turned out to be exceptions to the rule that transition metal lasers have broadband tunability. This untunable nature of, particularly, the well-known ruby laser may have considerably delayed the development of broadly tunable transition metal lasers, such as alexandrite (1979) and Ti^{3+}-doped alumina (1982). So, where does the tunability of transition metal ions come from? After all, the rare earth lasers (e.g., Nd, Er, Tm, and Ho) have sharp energy transitions with quite limited laser tunability. The

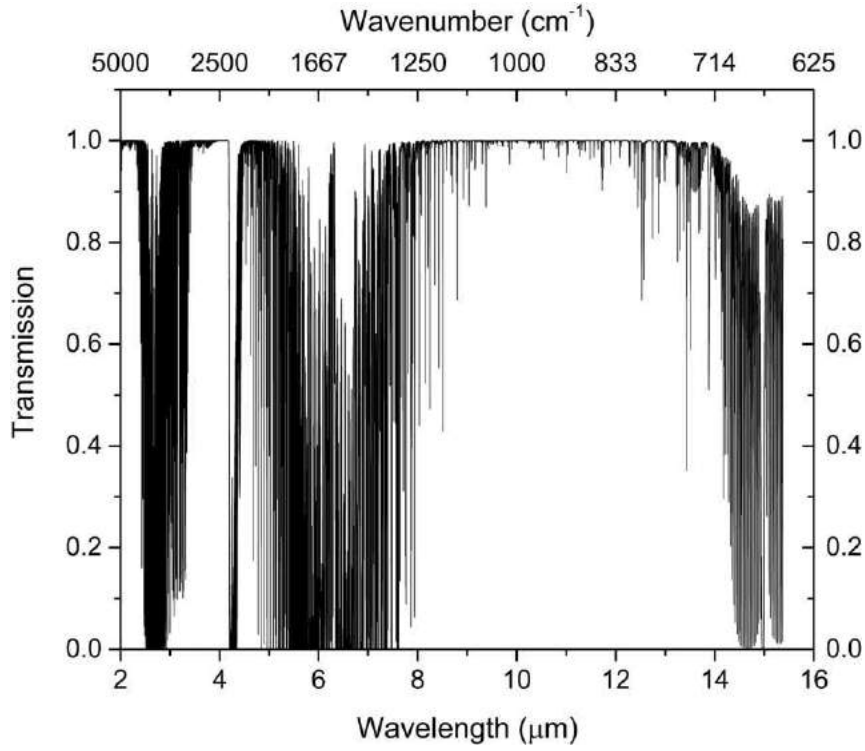

Fig. 1 Transmission through 10 m of mid-latitude summer atmosphere using HiTRAN on the Web.

difference lies in the nature of the energy levels involved. The partially filled electron orbitals in rare earths are f-shell orbitals, while the partially filled electron orbitals in transition metals are d-shell orbitals.

Transition Metal Spectroscopy

Crystal Field Theory

The IUPAC definition of a transition metal is an element whose atom has a partially filled d sub-shell, or which can give rise to cations with an incomplete d sub-shell. The d-shell electrons in a transition metal ion located in a crystalline solid are exposed to the influences of surrounding atoms and their electrons. Thus, for example, a Cr^{3+} ion substituting for an Al^{3+} ion in alumina (Al_2O_3), has six O^{2-} ions octahedrally arranged around it. These six negative charges form a crystal field which breaks the degeneracy of the d-electrons of an isolated chromium ion. Thus we now have the possibility of energy transitions (excitation, spontaneous emission, and stimulated emission) between d-shell orbitals. The symmetry of the surrounding ions and their proximities determine the strength of the crystal field. The crystal field and the nature of the transition metal ion determine the splitting of the d orbitals. However, the crystal field felt by f-shell electrons in rare earth ions is very weak. This is due to the fact that a filled s shell extends beyond the f shell and shields the f electrons from external fields. In general, for rare earths, Coulomb splitting of electronic energy levels is the strongest effect followed by spin–orbit splitting with crystal field being by far the weakest. Thus rare earth ions like Nd^{3+} have energy level splitting's that differ very little from host to host, for example, Nd:YAG (1064 nm), Nd:YVO_4 (1064 nm) Nd:YLF (1047, 1053 nm), and Nd:GSGG (1061 nm). Here, YAG (yttrium aluminum garnet [$Y_3Al_5O_{12}$]), yttrium vanadate [YVO_4], YLF (yttrium lithium fluoride [$LiYF_4$]), and GSGG (gadolinium scandium gallium garnet [$Gd_3Sc_2Ga_3O_{12}$]) are common crystalline hosts for rare earth ion dopants.

A simple example is given below to illustrate the concept of crystal field theory splitting of degenerate d-shell orbitals. Let us consider a single 3d electron in an octahedral field, for example, a Ti^{3+} ion substituted for Al^{3+} in Al_2O_3. The degeneracy of the free ion is $(2s+1)(2l+1)=(2(1/2)+1)$ spin states x $(2(2)+1)$ orbital states $=10$. Let us now see how the orbital degeneracy is partially removed by the crystal field (Fig. 2).

The electrostatic potential from ligand ions approximated as point charges along the x direction is

$$V_x = \frac{-Ze}{4\pi\varepsilon_o}\left[\frac{1}{\sqrt{r^2+a^2-2ax}}+\frac{1}{\sqrt{r^2+a^2+2ax}}\right]$$

where $r^2 = x^2+y^2+z^2$ is the position of the 3d electron and Ze is the charge of each ligand ion. We have similar terms for V_y and V_z. The total potential is $V = V_x + V_y + V_z$. After expanding in terms up to the 6th degree we have

$$V(x,y,z) = \frac{-Ze}{4\pi\varepsilon_o}\left\{\frac{6}{a}-\frac{35}{4a^5}\left[x^4+y^4+z^4-\frac{3}{5}r^4\right]-\frac{21}{2a^7}\left[x^6+y^6+z^6+\frac{15}{4}\left(x^2y^4+x^2z^4+y^2x^4+y^2z^4+z^2x^4+z^2y^4\right)-\frac{15}{14}r^6\right]\right\}$$

The Hamiltonian describing the energy of this system in terms of spherical harmonics $Y_l^m(\theta,\phi)$, $(-l\le m\le l)$ and neglecting terms greater than r^4 is

$$H_c^{O_h}(\mathrm{r}) = -\mathrm{eV} = \frac{Ze^2}{4\pi\varepsilon_o}\left\{\frac{6}{a}+\frac{7r^4}{2a^5}\left[Y_0^4(\theta,\varphi)+\sqrt{\frac{5}{14}}\left(Y_4^4(\theta,\varphi)+Y_{-4}^4(\theta,\varphi)\right)\right]\right\}$$

Let us define $D \equiv \frac{1}{4\pi\varepsilon_o}\frac{35Ze^2}{4a^5}$ and $q \equiv \frac{2}{105}\langle r^4\rangle$. Then using the properties of spherical harmonics the $<3dm_l'|H|3dm_l>$ matrix elements (where m_l is the orbital angular momentum quantum number) are: as shown in Fig. 3(a).

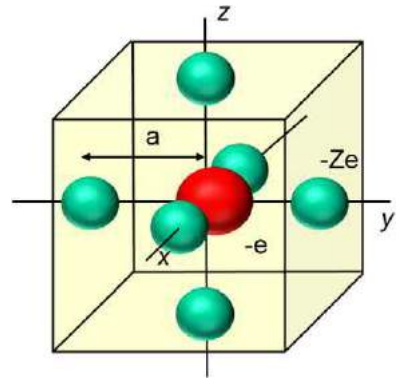

Fig. 2 Crystal field diagram for octahedral symmetry. Six negative ions (green) each with charge –Ze are located on the faces of a cube each a distance a from the central transition metal ion (red) with a single d-shell electron.

Obviously, $|3d\text{-}1>$, $|3d0>$, and $|3d1>$ are eigenstates since only diagonal matrix elements in Fig. 3(a) are nonzero. But the $m_l = \pm 2$ states are mixed. After diagonalizing the $m_l = \pm 2$ states we get degeneracy $g=3$ eigenstates (labeled T_2 in group theory terminology) with energy $-4Dq$ and $g=2$ eigenstates (labeled E) with energy $6Dq$. Thus as shown in Fig. 3(b) the octahedral crystal field has split the 10 degenerate electronic energy levels by $10Dq = \Delta_o$ with four levels in the upper level and six in the lower level. (Remember that the two electron spin states double the total degeneracy of each orbital degeneracy.) The calculated energy splitting makes sense physically when we look at the orbital orientations. (see Fig. 4). The T_2 orbitals have lower energy because they avoid the negative ions. The E orbitals have higher energy because they overlap neighboring negative ions.

At this point we will generalize the above discussion to state that a crystal field splits degenerate d orbital electronic energy levels into a set of multiple levels. Details for a specific transition metal ion depend upon how many d-level electrons are present and the strength and symmetry of the crystal field. For the case of Ti^{3+} there is only one electron in the 10d orbitals (d^1) and the crystal field energy levels are as just calculated. For other transition metal ions the number of electrons in d orbitals is different. For example, Cr^{2+} has four electrons in d orbitals (d^4) which is daunting to calculate until we realize that four of the five orbitals will be singly occupied (spin pairing is avoided) and we have one empty orbital. Thus the crystal field splitting is that of a single hole (a positive charge) and the T_2 and E sets of levels switch places but the size of the splitting for a given crystal field symmetry does not change. Details of crystal field splitting have been calculated for each of the d^q configurations ($q=1$–9) of the transition metal d orbital electrons and can be conveniently displayed in Tanabe–Sugano energy level diagrams. These useful diagrams can be found in Henderson and Bartram (2000).

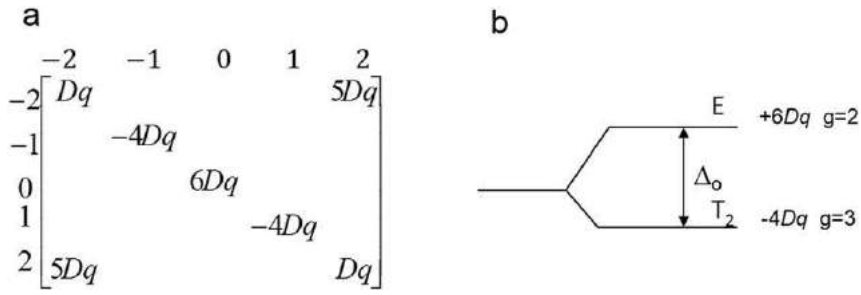

Fig. 3 (a) Crystal field Hamiltonian matrix elements for a single d-shell electron in octahedral symmetry with m_l eigenstates. (b) Crystal field splitting of the d-shell orbitals into a ground state triplet and an excited state doublet.

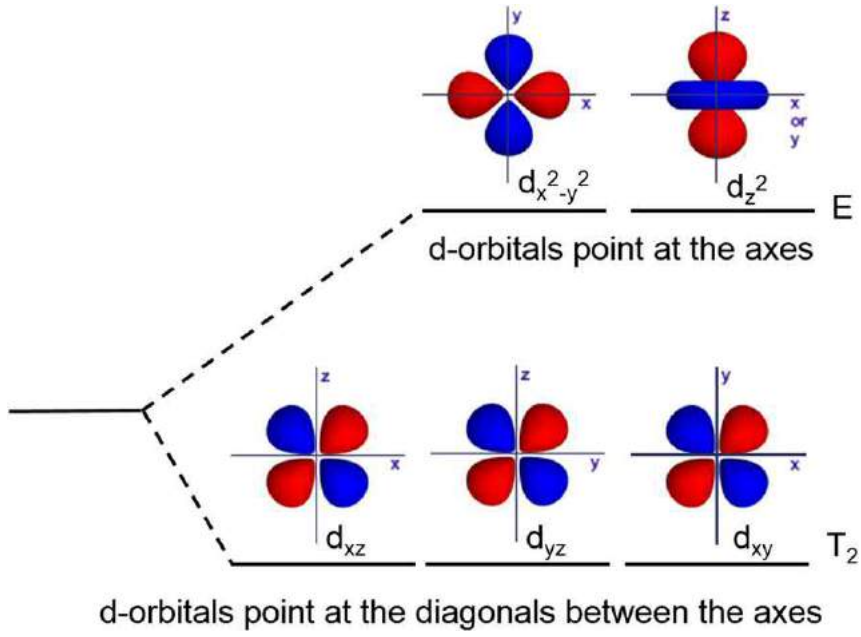

Fig. 4 Shapes of d orbitals and energy level splitting for octahedral crystal fields into two higher energy orbitals that point toward the surrounding negative charges and three lower energy orbitals that point diagonally between them.

Single Configuration Coordinate Model

But we have not yet explained the origin of broadband laser tunability of transition metal lasers. A crystal field not only splits the energies of the d orbitals but also couples the electronic motion to the motion (vibrations) of the surrounding crystal lattice ions. We can still assume that the electrons move much faster than the ions and that they adiabatically adjust to the varying ionic positions. This allows us to treat the full Hamiltonian of the electronic energies as a sum of the Hamiltonians for the electronic motions and the ionic vibrations. The ions can be modeled as masses connected to each other with springs, i.e., harmonic oscillators. Now instead of concentrating on the complex motion of many ions we can simplify things by considering the characteristic normal modes of vibration of the central transition metal ion and the nearest shell of surrounding ions. Each mode is a pattern of motion in which all parts of the system move sinusoidally with the same frequency and with a fixed phase relation. Each mode has a harmonic oscillator energy level structure of $h\Omega\left(n+\frac{1}{2}\right)$ where Ω is the vibrational frequency, n is the number of phonons (vibrational quanta), and h is Planck's constant. In any real crystal system, many normal modes of vibration must be taken into account. However, we will make the simplifying assumption that we can confine our attention to one representative mode only. It is conceptually useful to consider this the breathing mode where the surrounding shell of negative ions are moving in phase toward and away from the stationary central ion with the separation distance labeled as Q. In the single configuration coordinate (SCC) model, Q oscillates about its equilibrium value Q_o. Each electronic energy level has its own set of phonon sublevels spaced by energy $h\Omega$ and Q_0 values that are typically not equal due to different electron–lattice couplings; for example, consider how the two T_2 and E electronic levels have very different interactions with the surrounding electron clouds (Fig. 4).

A generalized example of the SCC energy curves is shown below (Fig. 5) for two electronic levels each of which has multiple vibrational sublevels. Each sublevel is an eigenstate of the system (an approximation that assumes electrons move much faster than the ions and that they adiabatically adjust to the varying ionic positions). The wavefunctions of the states can be represented as a product of an electron wavefunction ψ and a harmonic oscillator wavefunction χ. Probability of a photon absorption transition from electron–vibration state $|\psi_a\chi_n>$ to electron–vibration state $|\psi_b\chi_m>$ is proportional to:

$$|\langle\psi_b\chi_b(m)|\mu|\psi_a\chi_a(n)\rangle|^2$$

where μ is the appropriate electronic dipole operator. The harmonic oscillator states, χ, do not depend upon μ so we can separate the matrix element evaluation (Condon approximation) The transition probability is then:

$$W_{an-bm} = P_{ab}|\langle\chi_b(m)||\chi_a(n)\rangle|^2$$

where P_{ab} is the purely electronic transition probability and is the same for all n and m. The χ's are the harmonic oscillator eigenfunctions, but defined with different equilibrium points for electronic levels a and b so the overlap integrals (matrix elements) are in general not zero for $m \neq n$.

Notice that if electronic levels a and b do have the same Q_o equilibrium point, then the mth and nth χ eigenfunctions are orthogonal unless $m=n$ so the matrix element $|\langle\chi_b(m)||\chi_a(n)\rangle|^2 = \delta_{nm}$. We have uncoupled the electrons from the lattice vibrations. Rare earth ions act this way because the f electronic orbitals are shielded from the crystal field of the ligand ions.

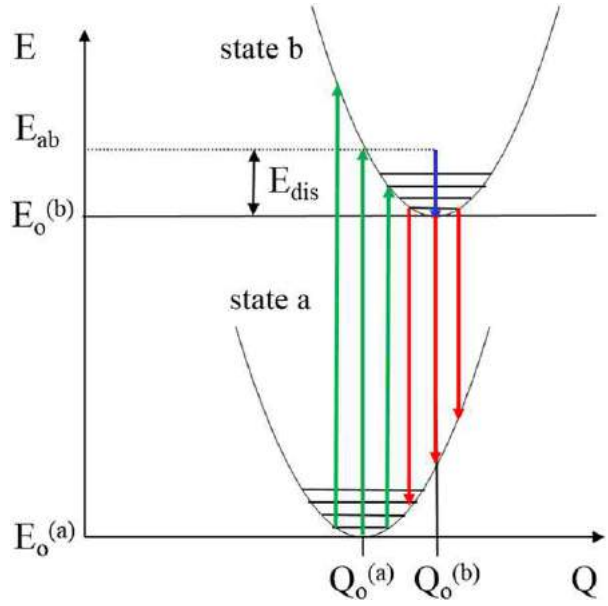

Fig. 5 Single configuration coordinate model vibronic energy levels. The green up-arrows are absorption transitions and the red down-arrows are emission transitions.

Classically we can think of the harmonic oscillator nuclei spending most of their time at the extrema of their oscillations so transitions take place primarily from one parabolic curve to another (Franck–Condon principle). Quantum mechanically one calculates the overlap integral of the product of the initial and final oscillator wavefunctions. These wavefunctions vary rapidly between the parabolic end points again resulting in transitions primarily happening at energies corresponding to going from one parabolic curve to the other. A closed form solution of the overlap integral is

$$|\langle\chi_a(m)|\chi_b(n)\rangle|^2 = \exp[-S/2]\sqrt{n!/m!}\left(\sqrt{S}\right)^{m-n} L_n^{m-n}(S)$$

L_n^{m-n} is an associated Laguerre polynomial and S is the Huang–Rhys factor, a dimensionless constant characterizing the strength of the electron–lattice coupling.

$$S \equiv \frac{E_{\text{dis}}}{\hbar\Omega}$$

E_{dis} is defined in **Fig. 5**. Note that E_{dis} and thus S increase quadratically as the Q offset of the upper and lower parabolas increases.

At temperature $T=0$K only the $n=0$ vibrational state is occupied so the matrix elements become

$$F_m(0) = |\langle\chi_b(m)\rangle|\ |\langle\chi_a(0)\rangle|^2 = \frac{\exp[-S]S^m}{m!}$$

and the absorption bandshape is

$$I_{ab}(E) = I_o \sum_m \frac{\exp[-S]S^m}{m!}\delta(E_{bm} - E_{a0} - E)$$

Since $\sum_m |\langle\chi_b(m)||\chi_a(n)\rangle|^2 = 1$ the intensity of the full band is I_o and is independent of S. This also means that the total intensity is independent of temperature. Intensity of the zero-phonon line ($m=0$, $n=0$, i.e., energy difference between the two bottom levels of the two parabolas in the SCC model diagram) is $I_o e^{-S}$. If $S=0$, then all of the intensity is contained in the zero-phonon line and there is no lateral displacement of the harmonic oscillator parabolae, i.e., $Q_o^{(a)} = Q_o^{(b)}$. As S increases, the intensity in the zero-phonon line decreases, but this is compensated by the appearance of vibrational sidebands observed at energies spaced by $\hbar\Omega$. The larger S is, the broader the absorption spectrum becomes; similarly emission is also broadened as S increases. These trends are shown in **Fig. 6** for several values of S. S values are typically in the 5–10 range for transition metal ions.

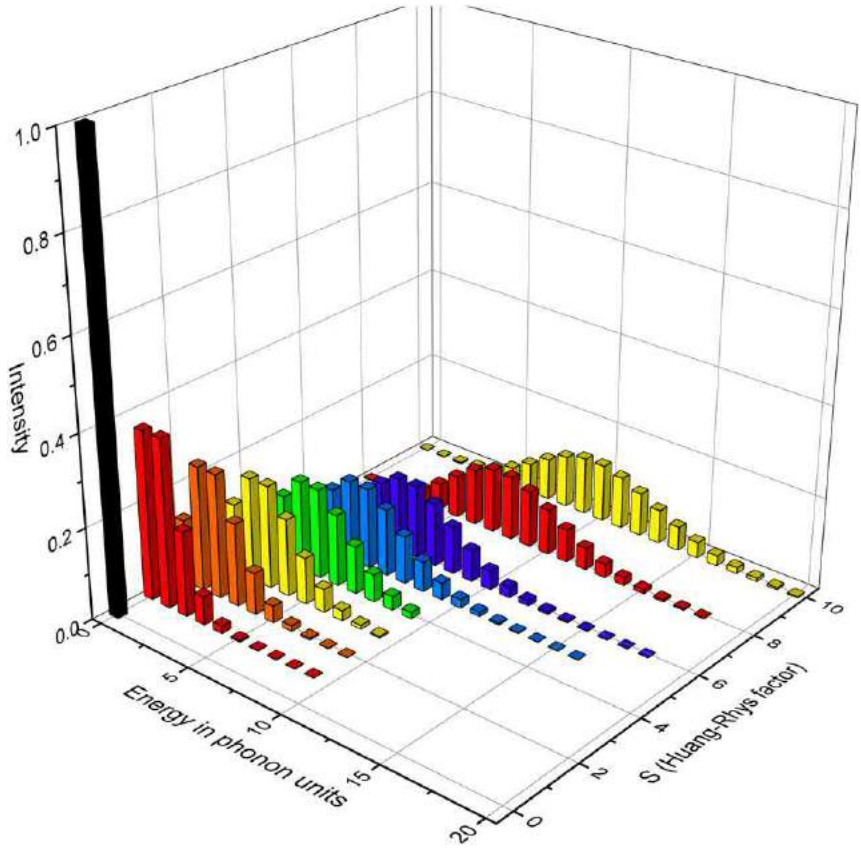

Fig. 6 Absorption line intensities for Huang–Rhys values in the $S=0$–10 range. The energy of the absorbed photon is given in phonon units ($\hbar\Omega$).

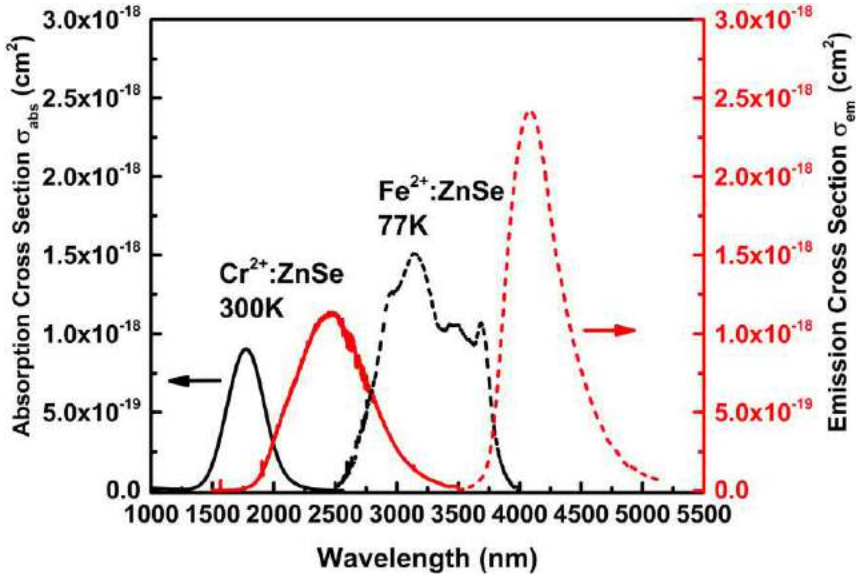

Fig. 7 Emission (red) and absorption (black) cross sections of Cr^{2+} (solid lines) and Fe^{2+} (dashed lines) doped into ZnSe.

The δ-function bandshape given above is not valid for a real system. A typical crystal lattice has many modes of vibration, not just one; this smoothes out the stick spectrum of Fig. 6 into a broad lineshape. There is one exception: all phonon modes have the same zero-phonon energy so the zero-phonon line remains a sharp Lorentzian line. However, increasing temperature further broadens the individual peaks into a single featureless peak hiding the zero-phonon line.

Note that we have finally come to the explanation of broadband tunable lasing of transition metal ions doped into solids. Vibronic (vibrational and electronic level) transitions combined with different vibrational center points for different electronic levels result in broad bands of absorption and emission. As an example Fig. 7 shows the measured absorption and emission spectra of Cr^{2+} and Fe^{2+} transmission metal ions doped into ZnSe.

The Cr^{2+} Revolution

While ruby was technically the first transition metal laser, alexandrite (Cr^{3+}:$BeAl_2O_4$) was the first (1978) tunable transition metal laser. Shortly afterwards, Ti^{3+}:Al_2O_3 was discovered (1982) with tunability in the 700–1000 nm region and a few years later < 50 fs Kerr-lens modelocking was demonstrated (1990) leading to the ubiquitous use of this material for ultrashort pulse research and applications. In 1996 (DeLoach *et al.*, 1996) Cr^{2+} doped II–VI semiconductors were introduced as a new class of laser materials. (II Refers to 2^+ ions in the Zn column of the periodic table (Group 12) and VI refers to 2^- ions in the oxygen column (Group 16).) In the past 20 years Cr^{2+} and more recently Fe^{2+} lasers have become the Ti:sapphire lasers of the mid-IR (2–6 μm) spectral region. A large variety of II–VI semiconductor materials have been shown to work as transition metal hosts. These materials have a 2^+ Group 12 cation (such as Zn or Cd) surrounded by 4 Group 16 anions arranged at the corners of a tetrahedron. The tetrahedral crystal field, like the octahedral field discussed above, splits the d orbitals into the same E and T_2 sets of orbitals but for a single d electron the E levels are now lower than the T_2 levels. For tetrahedral symmetry the E orbitals avoid the surrounding anions, while the T_2 orbitals point toward the surrounding anions. However, Cr^{2+} is a d^4 configuration (four electrons in d orbitals acting as a single hole since five electrons would be a half-filled shell) so we must perform one more sign change bringing us back to the original case where the T_2 levels are lower than the E levels. Because there are fewer ligands (four instead of six) for the tetrahedral symmetry case, the crystal field splitting is also smaller, i.e., $\Delta_t = (4/9)\Delta_o$. This smaller crystal field splitting shifts optical absorption and emission transitions into the infrared region of the spectrum in contrast to the shorter wavelength transitions of Ti^{3+}:sapphire which has octahedral crystal field symmetry. A number of TM:II–VI lasing ion – host combinations have been investigated including Cr^{2+}, Ni^{2+}, Co^{2+}, and Fe^{2+} as TM lasing ions and ZnSe, ZnS, ZnTe, CdSe, and alloys, such as $Cd_xMn_{1-x}Te$ as host materials. ZnSe and ZnS have proven to be the best host materials and Cr^{2+} and Fe^{2+} the best transition metal lasing ions. Reasons for this down-selection are discussed in the Transition Metal Laser Advantages and Issues section.

Cr^{2+} Laser Advances

Table 1 shows examples of demonstrated Cr^{2+} laser performance. Modes of operation include continuous wave (CW), gain switched (GS), and ML. Q-switched operation is not practical because of the short E-level radiative lifetime of a few microseconds but GS pulses over 1 J have been demonstrated. Tuning over the 2–3.3 μm range and modelocking of ultrashort pulses as short as 26 fs have

Table 1 Cr^{2+} laser performance

Gain medium	*Repetition rate*	*Energy*	*Wavelength/ range (μm)*	*Pulse width*	*Power (ave.) (W)*	*Temperature (K)*	*Regime*	*References*
Cr:ZnS (PC)	80 MHz	90 nJ	2.3	26 fs	2	RT	ML	Vasilyev *et al.* (2016)
Cr:ZnSe (PC)			2.4		140	RT	CW	Moskalev *et al.* (2016)
Cr:ZnSe (PC)	0.7 kHz linewidth		2.3		5.5	RT	CW	Mirov *et al.* (2015b)
Cr:ZnSe (PC)	0.1 Hz	1.1 J	2.65	7 ms	0.110	RT	GS	Fedorov *et al.* (2012)
Cr:ZnS (PC)			1.96–3.20		0.01–0.70	RT	CW	Sorokin *et al.* (2010)
Cr:ZnSe (PC)			1.97–3.35		0.01–0.55	RT	CW	Sorokin *et al.* (2010)
Cr:ZnSe (WG)			2.08–2.78		0.01–0.12	RT	CW	MacDonald *et al.* (2015)
Cr:ZnSe (WG)			2.5		1.7	RT	CW	Berry *et al.* (2013)
Cr:CdSe (SC)	1 kHz	0.5 mJ	2.6/2.3–3.4	40 ns	0.50/.35	RT	GS	McKay *et al.* (1999, 2002)

Abbreviations: CW, continuous wave; GS, gain switched; ML, modelocked; PC, polycrystalline; RT, room temperature; SC, single crystal; WG, waveguide.

Table 2 Fe^{2+} laser advances

Gain medium	*Repetition rate (Hz)*	*Energy*	*Wavelength/ range (μm)*	*Pulse width*	*Power (ave.) (W)*	*Temperature (K)*	*Regime*	*References*
Fe:ZnSe	Single pulse	0.3 mJ	3.95–5.05	60 ns		RT	GS	Fedorov *et al.* (2006)
Fe:ZnTe	Single pulse	0.18 mJ	4.35–5.45	40 ns		RT	GS	Frolov *et al.* (2011)
Fe:ZnSe			4.1		1.6	77	CW	Mirov *et al.* (2015,b)
Fe:ZnSe	850,000	600 nJ	4.04	64 ns	0.515	77	QS	Evans *et al.* (2014)
Fe:ZnSe	100	350 mJ	4.15	150 μs	35	77	GS	Mirov *et al.* (2015a)
Fe:ZnSe	Single pulse	42 mJ	4.5	750 μs		RT	GS	Frolov *et al.* (2013)
Fe:ZnS	Single pulse	3.2 J	3.6/3.4–4.2	750 μs		85	GS	Frolov *et al.* (2015)
Fe:ZnSe (WG)			4.1		0.049	77	CW	Lancaster *et al.* (2015)

Abbreviations: CW, continuous wave; GS, gain switched; QS, Q-switched; RT, room temperature; WG, waveguide.

been demonstrated. The current CW power record is 140 W. Cr^{2+} lasers are typically optically pumped in the 1.6–2.0 μm absorption band with Er^{3+} or Tm^{2+} fiber lasers although Raman-shifted emission from 1 μm Nd lasers can also be used. Energy transfer from the ZnSe conduction band to the Cr^{2+} ^5E excited level has been demonstrated but transfer efficiency is poor.

Fe^{2+} Laser Advances

Fe^{2+} ions can also be doped into II–VI semiconductors. Again the crystal symmetry at the cation substitution site is tetrahedral. However, Fe^{2+} has a d^6 configuration, i.e., a half-filled d shell plus one orbital occupied by a single electron of opposite spin. Thus we can treat Fe^{2+} as having a single electron in a tetrahedral field, resulting in a Δ_t splitting with the E levels lower in energy than the T_2 levels. However, the actual situation is more complicated. We have neither considered spin–orbit coupling up to this point nor have we considered the Jahn–Teller effect (degeneracy of a ground state will be removed by crystal field distortion). These effects are not required (because they are small) to describe the observed spectra of Ti^{3+} and Cr^{2+} ions. However, in Fe^{2+} ions the relative strengths of crystal field and spin–orbit effects are not well known and details of observed spectra are not clearly explained. Still the general concept of an SCC model is quite adequate to explain laser operation of this ion. Fe^{2+} demonstrated laser performance is shown in **Table 2**.

Fe^{2+} lasers have two important differences from Cr^{2+} lasers. First the energy splitting between the upper and lower levels is less for Fe^{2+} ions so the absorption peak is shifted from 1.8 to 3.2 μm and the emission peak is shifted from 2.4 to 4.1 μm (3.7–5.0 μm tuning). Optical pumping of Fe^{2+} ions is typically accomplished using Er^{3+} lasers operating at 3 μm. Cr^{2+} lasers tuned to the 3 μm region can also be used. Secondly, the lifetime of the Fe^{2+} upper lasing level is much more affected by temperature than the lifetime of the Cr^{2+} upper lasing level. The result is that Fe^{2+} room temperature (RT) CW lasing operation is not practical. However, it is possible to achieve GS operation at RT but at low efficiency since the lifetime has been reduced from 50 μs at low temperatures to 360 ns at RT.

Transition Metal Laser Advantages and Issues

The vibronic nature of absorption and emission energy transitions allow for broadband absorption and emission spectra as already shown in **Fig. 7**. The broadband emission provides the broadband tuning and ultrashort pulses characteristic of most

transition metal lasers. However, transition metal lasers do have some important disadvantages that must be mitigated. Ti^{3+} is an example. Ti^{3+} doped into alumina (Al_2O_3) works well as a RT, broadly tunable laser material. However, Ti^{3+} doped into $YAlO_3$ has not shown laser operation even though it exhibits bright yellow broadband fluorescence. The reason is thought to be ESA from the ^2E metastable level to the conduction band of $YAlO_3$ (band gap 7.1 eV). Stimulated emission gain cannot overcome ESA loss. Ti^{3+} ESA does not take place in alumina because the crystal field splitting is smaller and the alumina band gap (8.8 eV) is larger. Ti^{3+} ions have also been doped into lower crystal field hosts, such as YAG and Gd-Sc-Al garnet (GSAG). The lower crystal field results in significant nonradiative relaxation of the excited ^2E level back to the $^2T_2$ ground state. Thus efficient lasing at RT for these materials is not likely other than gain-switch operation by sub-microsecond pump pulses.

ESA: Co^{2+} Yes, Cr^{2+} No

Co^{2+} [d^7] and Cr^{2+} [d^4] are illustrative examples of transition metal ions that do and do not exhibit ESA. ESA directly competes with lasing gain so an ESA cross section similar in magnitude to the emission cross section is the death knell for efficient lasing.

Tanabe–Sugano diagrams are useful for describing the spectroscopy of transition metal ions. The horizontal-axis of a Tanabe–Sugano diagram is expressed in terms of the splitting parameter $10Dq$, or Δ, divided by the Racah parameter B. The vertical-axis is in terms of energy, E, also scaled by B. The Tanabe–Sugano diagrams for [d7] (Fig. 8) and [d4] (Fig. 9) energy levels in tetrahedral symmetry are shown below. Co^{2+} is a [d7] transition metal ion and has a $^4T_2$ first excited state with broadband emission suitable for mid-IR lasing. However, the $^4T_1$ next higher level is at a location which approximately matches the emission energy so absorption from the $^4T_2$ level to the $^4T_1$ could be a problem. Details would depend upon how spin–orbit coupling and Jahn–Teller effects would split these levels but the experimental fact is that no Co^{2+}:II–VI laser with tetrahedral site symmetry has yet been demonstrated.

The first excited state for Cr^{2+} as shown in Fig. 9 is a ^5E state (spin multiplicity of 5). There are numerous higher lying levels so initially one might think that ESA could be a problem for Cr^{2+} in tetrahedral symmetry as well. However, all of the upper lying levels have different spin multiplicities (3 and 1) and since transitions between different total spins are forbidden, ESA should not occur and it has not been observed experimentally. The result is a robust lasing ion.

Nonradiative Relaxation

A good laser material should only be able to relax from the excited state back to the ground state through radiative emission (stimulated or spontaneous). Any nonradiative relaxation mechanism competes with the stimulated emission needed for lasing. In the case of transition metal ions, level crossing between upper and lower electronic level parabolas can occur due to offset of the different Q_o equilibrium points. The diagram in (Fig. 10) shows a hypothetical case where this occurs. At low temperature only the lowest vibronic level of the upper electronic level is occupied after optical excitation. But at higher temperatures when $kT \approx \Delta E$, higher vibronic levels will be occupied. The probability of crossover from the upper electronic level to the lower electronic level

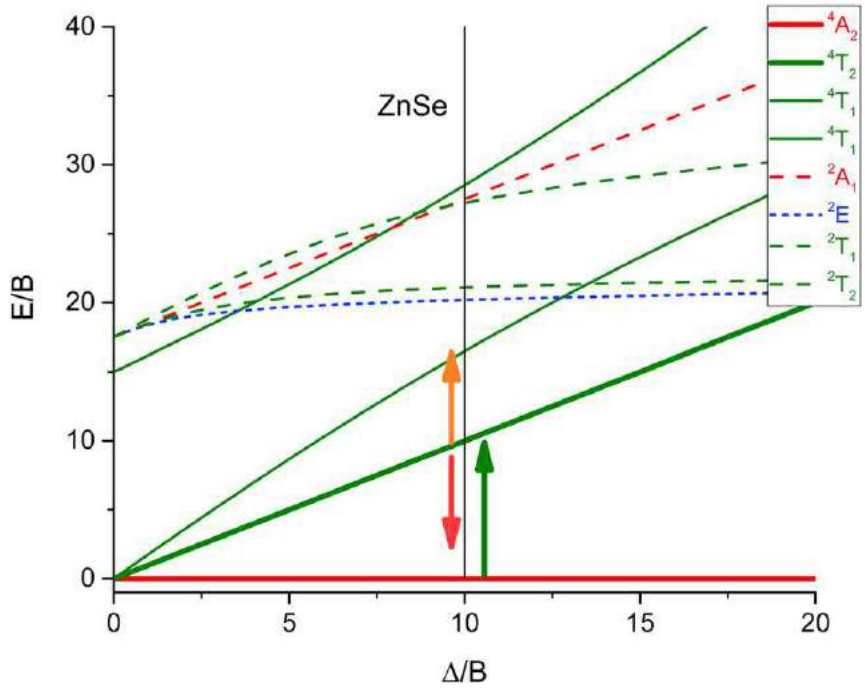

Fig. 8 Tanabe–Sugano diagram for Co^{2+} in tetrahedral symmetry. Solid-line curves have same spin as the ground state; dashed curves do not.

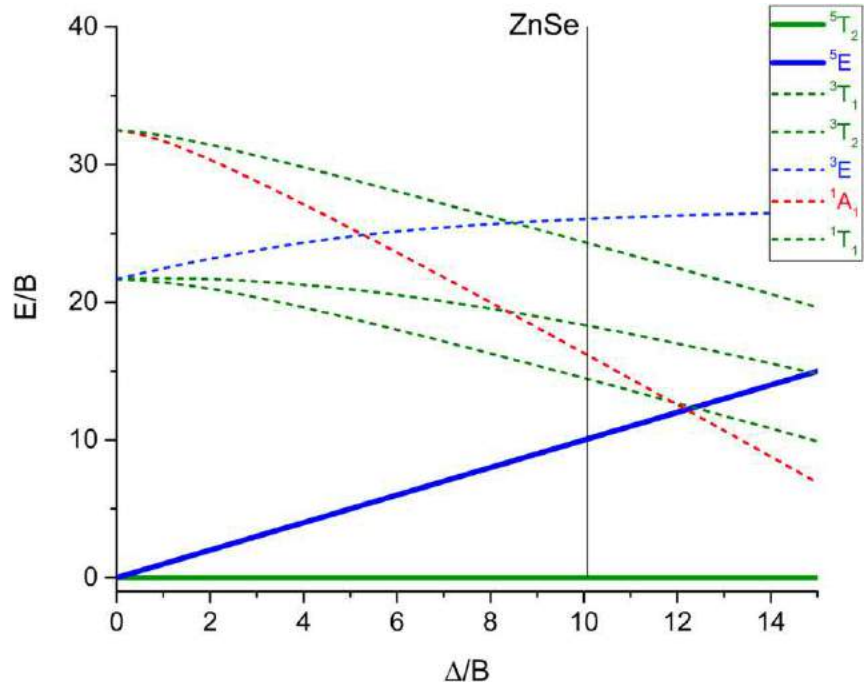

Fig. 9 Tanabe–Sugano diagram for Cr^{2+} in tetrahedral symmetry. Solid-line curves have same spin as the ground state; dashed curves do not.

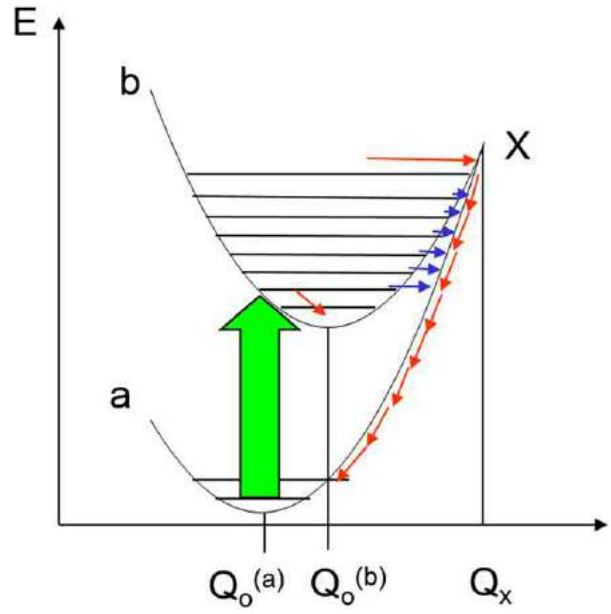

Fig. 10 Single configuration coordinate diagram of nonradiative relaxation. *X* is the classical energy crossover point. The blue lines represent tunneling from the *b* electronic level to the *a* electronic level. Phonon emission relaxes the *a* state back to thermal equilibrium.

without emission of a photon increases with temperature. Once the crossover has occurred, the vibronic excitation can relax quickly by emitting phonons, essentially sliding down the lower electronic parabola. Thus as temperature increases nonradiative relaxation becomes stronger to the point that a useful population inversion and lasing are not possible. The thermal relaxation path can be described by a rate constant N and an activation energy ΔE.

$$W_{\mathrm{nr}} = N \exp\left[-\frac{\Delta E}{kT}\right]$$

Usually, the calculated activation energy based upon our SCC model is too high compared to the measured effect. The reason is that we have not allowed for quantum mechanical tunneling. Because the wavefunctions extend beyond the edge of the parabola there is a finite probability that tunneling from one parabola to another can take place. This makes the classical activation energy ΔE too high. A more correct way is to calculate the transition probability, which involves the overlap integral for the upper and lower vibrational wavefunctions and add the results for all overlapping pairs. This calculation has been done and yields a nonradiative relaxation rate $W_{\text{tunneling}} = N$ (the electronic part of the wavefunction overlap) $\times$ a vibrational exponential decay term that increases with temperature and has an S (Huang–Rhys) factor in the exponential like the radiative relaxation rate.

$$W_{\text{tunnel}} = N\exp[-S\langle 2m+1\rangle]\sum_{j=0}^{\infty}\frac{(S\langle m\rangle)^{j}(S\langle m+1\rangle)^{p+j}}{j!(p+j)!}$$

N is typically $\sim 10^{13}\ \text{s}^{-1}$. p is the number of phonons it takes to equal the zero-phonon energy gap between the electronic energy levels involved and $<m>$ is the *m*ean thermal occupancy.

$$\langle m\rangle = \left[\exp\left(\frac{\hbar\omega}{kT}\right) - 1\right]^{-1}$$

This equation agrees quite well with the magnitude and temperature dependence of transition metal ion lifetimes. Experimentally one measures the excited state lifetime $\tau = 1/W$ and total relaxation rate $W = W_{\text{rad}} + W_{\text{nonrad}}$ or $1/\tau = 1/\tau_{\text{rad}} + 1/\tau_{\text{nonrad}}$. At low temperatures the relaxation is primarily radiative so we can determine the radiative lifetime by cooling our sample to the point where lifetime no longer depends upon temperature. As temperature increases we will reach a point where nonradiative relaxation becomes dominant and quenches emission. As emission becomes more quenched, lasing threshold increases and we reach a point where laser operation is no longer practical.

The details of transition metal ion excited level lifetimes versus temperature depend on both the ion and the host material. For example, Ti^{3+} has only a small amount of nonradiative relaxation at RT in sapphire ($W_{\text{nonrad}} = W_{\text{rad}}$ at 350K) but in the lower crystal field GdScGa-garnet, the nonradiative component is completely dominant at RT and $W_{\text{nonrad}} = W_{\text{rad}}$ already at 125K. The Cr^{2+} $W_{\text{nonrad}} = W_{\text{rad}}$ temperature is 350K in ZnSe but 300K in CdSe. Fe^{2+} is interesting as a transition metal laser ion because of its operation in the 3.5–5 μm regions. However, it has a $W_{\text{nonrad}} = W_{\text{rad}}$ temperature of 175K which requires cryogenic cooling for CW laser operation.

Nonradiative relaxation also becomes an issue for laser power scaling. As the laser power generated increases so does the heat load in the laser material. Active cooling is used to mitigate heating of the laser material but temperatures inside the laser crystal will increase as pump power increases regardless of how "good" the heat sink is. Rising temperature increases the nonradiative relaxation rate making lasing even less efficient, in turn leading to even more heating. This can result in a runaway situation where lasing will suddenly stop altogether.

Thermal Lensing

In addition to thermal quenching of emission, laser materials can, in general, experience thermal lensing. The refractive index of a material changes with temperature. The thermo-optic coefficient (dn/dT) of YAG at 1064 nm is $+9.1 \times 10^{-6}\ \text{K}^{-1}$. But dn/dT is much higher for II–VI semiconductors, for example, ZnSe $dn/dT = 62 \times 10^{-6}\ \text{K}^{-1}$ at 3.39 μm, 7x higher. Extracting heat from the sides of the laser material sets up a thermal gradient, which in turn establishes an index gradient, i.e., a lens in the material. The lensing power gets stronger as the pumping get stronger. Once the thermal lens becomes strong enough to destabilize the laser resonator, lasing power will fail to increase with increasing pump power or completely stop. Techniques to mitigate this effect, such as thin disk cooling have been investigated. In thin disk cooling a thin slab of laser material is mounted directly to a heat sink. Heat then flows parallel to the resonating laser beam and the radial index gradient should be much less. The technique works to some degree. However, this requires a complex multi-pass pump beam setup to achieve efficient pump absorption in such a thin material. Also, intense pumping of the small volume of the disk leads to a temperature rise in the disk resulting in runaway thermal quenching of the emission.

Concentration Effects

Increasing the Cr^{2+} doping concentration to reduce the number of pump passes required for a thin disk to efficiently absorb the pump power is not an option due to concentration quenching of the emission. Doping higher than a particular level causes rapid nonradiative relaxation.

Many laser materials have an optimum lasing ion concentration. This effect is seen in Nd:YAG at concentrations of a few percent, a level where random substitution of Nd ions results in a significant fraction of them being at nearest neighbor locations. Neighboring ions in a crystal lattice can, in many cases, quench emission. However, in Cr-doped II–VI materials concentration quenching of emission is seen at much lower levels; the fluorescence lifetime is reduced by 50% at the 0.1% level in ZnSe. Interestingly, Fe^{2+} ions do not display strong quenching at such low concentration levels (Myoung *et al.*, 2012). The reason for Cr concentration quenching at low levels is unknown but may be a tendency of Cr ions to preferentially diffuse into a crystal at nearest neighbor lattice points.

Mitigating Thermal Effects

A spinning disk has recently been used to scale Cr:ZnSe laser power from ~20 to 140 W CW (Moskalev *et al.*, 2016). Rotation of the disk effectively increases the heated volume thereby reducing the temperature and the thermal gradient in the gain region. By the time the material has rotated back into the resonator beam, the heat has diffused away and convection has cooled the material back to its starting temperature. This allows considerable scaling of laser power but with the complication of using moving parts that must be very flat over the entire excitation ring to prevent beam pointing wobble.

Another technique that has been used to mitigate thermal effects is lasing in a fiber or waveguide (WG) configuration. Waveguiding can be designed to be resistant to thermal gradients in the material. Also, because the gain can be distributed over a long distance the temperature rise in the material can be reduced. Finally, heat sinking can be brought quite close to the heated region along the entire length of the WG fiber including total immersion in a flowing coolant. The technical issue here is fabrication of WGs or fibers with low loss. At this time WGs have been fabricated in Cr:ZnSe and Fe:ZnSe using laser inscription, a technique that uses focused ultrashort laser pulses to write refractive index changes into the laser material in a pattern that provides guiding. Another technique uses high pressure chemical deposition of Cr:ZnSe on the walls of a hollow silica fiber core to form a high index, laser active guiding region. Lasers have been demonstrated using both of these techniques but powers thus far have been limited to a few watts. Passive losses need to be reduced to achieve high-power operation. Current devices using both techniques have losses of ~1 dB/cm. Waveguiding devices may actually be of more practical use where flexibility of design and compact robust packaging are needed.

Summary

Transition metals have played a key role in the development of tunable solid-state lasers. The broadband emission spectra of transition metals enabled by coupling to lattice vibrations provides both broadband tunability and ultrashort pulse generation. Transition metal based laser devices operating in the visible, near-IR, and more recently the mid-IR are now commercially available. The emphasis in the future is likely to be the incorporation of these lasers into infrared applications where broad tunability, high pulse energy, high average power or ultrashort pulses are needed.

References

Berry, P.A., MacDonald, J.R., Beecher, S.J., *et al.*, 2013. Fabrication and power scaling of a 1.7 W Cr:ZnSe waveguide laser. Optical Materials Express 3, 1250–1258.

DeLoach, L.D., Page, R.H., Wilke, G.D., Payne, S.A., Krupke, W.F., 1996. Transition metal-doped zinc chalcogenides: Spectroscopy and laser demonstration of a new class of gain media. IEEE Journal of Quantum Electronics 32, 885–895.

Evans, J.W., Berry, P.A., Schepler, K.L., 2014. A passively Q-switched, CW-pumped Fe:ZnSe laser. IEEE Journal of Quantum Electronics 50, 204–209.

Frolov, M.P., Korostelin, Y.V., Kozlovsky, V.I., *et al.*, 2011. Laser radiation tunable within the range of 4.35-5.45 μm in a ZnTe crystal doped with Fe^{2+} ions. Journal of Russian Laser Research 32, 528–536.

Frolov, M.P., Korostelin, Y.V., Kozlovsky, V.I., *et al.*, 2013. Study of a 2-J pulsed Fe:ZnSe 4-μm laser. Laser Physics Letters 10, 125001.

Frolov, M.P., Korostelin, Y.V., Kozlovsky, V.I., *et al.*, 2015. 3J Pulsed Fe:ZnS laser tunable from 3.44 to 4.19 μm. Laser Physics Letters 12, 055001.

Fedorov, V., Mirov, M.S., Mirov, S., *et al.*, 2012. Compact 1J mid-IR Cr:ZnSe laser. In: Frontiers in Optics 2012/Laser Science XXVIII, Paper FW6B.9.

Fedorov, V.V., Mirov, S.B., Gallian, A., *et al.*, 2006. 3.77-5.05-mm Tunable solid-state lasers based on Fe^{2+}-doped ZnSe crystals operating at low and room temperatures. IEEE Journal of Quantum Electronics 42, 907–917.

Henderson, B., Bartram, R.H., 2000. Crystal Field Engineering of Solid State Laser Materials. Cambridge: Cambridge University Press.

Lancaster, A., Cook, G., McDaniel, S.A., *et al.*, 2015. Fe:ZnSe channel waveguide laser operating at 4122 nm. In: CLEO, OSA Technical Digest, San Jose, CA, Paper SM2F.5.

MacDonald, J.R., Beecher, S.J., Lancaster, A., *et al.*, 2015. Ultrabroad mid-infrared tunable Cr:ZnSe channel waveguide laser. IEEE Journal of Selected Topics in Quantum Electronics 21, 375–379.

McKay, J., Schepler, K.L., Catella, G.C., 1999. Efficient grating-tuned mid-infrared Cr^{2+}: CdSe laser. Optics Letters 24, 1575–1577.

McKay, J.B., Roh, W.B., Schepler, K.L., 2002. Extended mid-IR tuning of a Cr^{2+}: CdSe laser. In: Fermann, M.E., Marshall, L.R. (Eds.), Proceedings of Advanced Solid-State Lasers, Paper WA7.

Mirov, S., Fedorov, V., Martyshkin, D., *et al.*, 2015a. High average power Fe:ZnSe and Cr:ZnSe mid-IR solid state lasers. In: Advanced Solid-State Lasers, Optical Society of America, Berlin, Paper AW4A.1.

Mirov, S.B., Fedorov, V.V., Martyshkin, D., *et al.*, 2015b. Progress in mid-IR lasers based on Cr and Fe-doped II–VI chalcogenides. IEEE Journal of Selected Topics in Quantum Electronics 21, 1601719.

Moskalev, I., Mirov, S., Mirov, M., *et al.*, 2016. 140 W Cr:ZnSe laser system. Optics Express 24, 21090–21104.

Myoung, N., Fedorov, V.V., Mirov, S.B., Wenger, L.E., 2012. Temperature and concentration quenching of mid-IR photoluminescence in iron doped ZnSe and ZnS laser crystals. Journal of Luminescence 132, 600–606.

Sorokin, E., Sorokina, I.T., Mirov, M.S., *et al.*, 2010. Ultrabroad continuous-wave tuning of ceramic Cr:ZnSe and Cr:ZnS lasers. In: Advanced Solid-State Photonics, Optical Society of America, Paper AMC2.

Vasilyev, S., Moskalev, I., Mirov, M., *et al.*, 2016. Recent breakthroughs in solid-state mid-IR laser technology. Laser Technik Journal 13, 24–27.

UV Lasers

Yushi Kaneda, The University of Arizona, Tucson, AZ, United States

Ultraviolet (UV) lasers, lasers with emission wavelengths shorter than 400 nm, have industrial as well as scientific applications including biochemical sensing, material processing, photolithography, and metrology. Because of their short wavelengths, UV light allows small features to be observed or patterned. The state-of-art semiconductor manufacturing process employs ArF excimer lasers at 193.4 nm to pattern (and fabricate) transistors as small as 10 nm scale (as of 2016) on silicon wafers, for example.

The energy of the photon is inversely proportional to its wavelength, therefore UV photons has higher energies compared to those in visible region. This makes their interaction with materials stronger, often leading to optical damage relatively easily. Materials transparent to shorter wavelength are limited as larger bandgap is needed to transmit photons with higher energy. One of the aspects of UV lasers is that they are required to have relatively narrow spectral width from chromatic aberration's point of view; most glass materials, which are limited in choices for transparency in the UV region, have large wavelength dependence at short wavelengths, as it gets close to the absorption edge of the material.

In this article, different types of UV laser sources are reviewed, and the techniques to generate UV light from longer wavelengths are explained below.

There are several types of UV lasers:

1. Excimer lasers such as XeCl at 308 nm, KrF at 248 nm, and ArF at 193.4 nm, which are currently in industrial use. Other excimer lasers include F_2 at 157 nm and Ar_2 at 126 nm.
2. Gas lasers, such as He–Cd at 325 nm, Ar^+ at 364 nm or 351 nm. They are usually relatively large and bulky, inefficient, often need water cooling, and require fairly frequent maintenance (replacement of tubes). He–Cd lasers at 325 nm may be used in the manufacturing process of volume Bragg grating (VBG) devices by interferometric exposure, due to its matching sensitivity of photo-thermal refractive (PTR) glasses to that wavelength. It should be noted that the coherence length of such lasers to be used for interferometric exposure should be fairly long (narrow spectral width). With recent development of new types of lasers such as harmonic generation, these lasers are beginning to be replaced. Harmonically converted Ar^+ lasers at either 257 nm (second harmonic generation (SHG) of 514.5 nm) as well as 244 nm (SHG of 488 nm), by placing the nonlinear crystal inside the laser resonator, has been used in industry; semiconductor inspection as well as fiber Bragg grating fabrication.
3. Semiconductor lasers using AlGaN below 400 nm. Diode lasers are attractive from application point of view, for their potential low cost, efficiency, and compactness. With the bandgap of AlN being around 6 eV ($\lambda \sim 200$ nm), AlGaN laser diodes (LDs) have the potential of emitting in the deep ultraviolet (DUV) region, which usually mean wavelengths shorter than 300 nm. It was in the late 1990s when the first blue-violet GaN diode laser was realized, and with the commercial demand in the optical disk, the display, as well as the lighting applications, the development of GaN LDs and light emitting diodes (LEDs) had been intensive to date. The demand for UV diode lasers are emerging, but not as strong as those for consumer applications, besides it is more challenging to fabricate reliable devices. As of this writing in 2016, the shortest wavelength that is commercially available from diode laser is 370 nm from Nichia with 70 mW output power. The process development of the material growth as well as the fabrication toward UV LDs at shorter wavelength has been a challenge, and only a few demonstrations with shorter emission wavelengths were made to date. With the electrical current injection pumping, laser diode grown on sapphire substrate with output wavelength as short as 336 nm had been reported (Yoshida *et al.*, 2008). With optical pumping, which requires another laser at a shorter wavelength, AlGaN lasers grown on bulk AlN substrate have been demonstrated at wavelength down to 237 nm under ArF laser pumping (Kneissl and Rass, 2015). For the detailed principles of semiconductor lasers in general.
4. Solid-state lasers, using transitions in cerium ions in 280–290 nm. Cerium-doped laser host crystals, such as $LiSrAlF_6$ (Ce:LiSAF) or $LiCaalF_6$ (Ce:LiCAF), had been studied since mid-1990s (Marshall *et al.*, 1994; Dubinskii *et al.*, 1993). Cerium ions in these host crystals can be pumped with the forth harmonic of neodymium lasers near 266 nm, and emit around 290 nm, tunable between approximately 280 nm and 320 nm. Tunable emission in this spectral range can be useful in chemical and biological sensing. Emission cross section is as high as of the order of 10^{-18} cm^2, so the optical gain can be high. Its lifetime is rather short, of the order of tens of nanoseconds, so they are better suited for pulsed pumping. Nevertheless, chopped-CW operation was demonstrated with a Ce:LiCAF laser.
5. Nonlinear frequency conversion from visible and/or infrared lasers. With the exception of excimer lasers in industrial applications, harmonically converted infrared lasers are the most common UV lasers. Common infrared lasers at 1064 nm including Nd:YAG, Nd:YVO_4, or fiber lasers can be converted to 355 nm by their third harmonic generation (THG), 266 nm by the fourth harmonic generation (4HG), or 213 nm by the fifth harmonic generation (5HG). Nonlinear materials that are transparent for these wavelengths and allows for phasematching are fairly limited; LBO is transparent to 160 nm but phasematches for THG, but not direct SHG at 266 nm. It can, however, phasematches for 266 nm generation by the sum-frequency mixing (SFM) between the fundamental and THG. Beta-barium-borate, β-BaB_2O_4, BBO is transparent to 189 nm, phasematches for the direct SHG down to 205 nm. It phasematches for 213 nm generation via both SFM between 1064 and 266 nm and SFM between

 doi:10.1016/B978-0-12-803581-8.09370-X

Table 1 Some of the nonlinear optical crystals and their properties

	LBO	*BBO*	*CLBO*	*KDP*	*KBBF*
Transparency	160 nm	180 nm	189 nm	200 nm	155 nm
Phasematched second harmonic generation (SHG)	278 nm	205 nm	238 nm	259 nm	162 nm
d_{eff} for SHG at 266 nm	N/A	1.54 pm/V	0.78 pm/v	0.46 pm/v	0.36 pm/v

532 and 355 nm. Cesium–lithium borate, $CsLiB_6O_{10}$, CLBO is transparent to 180 nm, and phasematches for direct SHG down to 238 nm as well as SFM at 213 nm via 1064 nm and 266 nm but not 532 and 355 nm. Potassium dihydrogen phosphate, KH_2PO_4, KDP is transparent to 200 nm, can phasematch down to 259 nm for direct SHG. Potassium difluo-diberryllo-borate, $KBe_2BO_3F_2$, KBBF is transparent to 155 nm, phasematches for direct SHG at 162 nm, but is difficult to grow and fabricate into polished devices and not commercially available as of 2016. Some of the properties of these materials are summarized in **Table 1**. They have effective nonlinear optical coefficients of just a few pm/V or less, so with realistic size of the crystal of a few centimeter or less, the normalized conversion efficiency is of the order of 10^{-4} W^{-1}, meaning 1 W at the harmonic is generated with the fundamental of 100 W. When one needs very large pulse energies, crystals of large aperture are required to avoid optical damage caused by the high peak power. For that purpose, KDP crystals are developed, and are grown in aqueous solution in the size of multiple tens of kilogram of single crystals to yield SHG/SFM devices with a clear aperture of more than 30 cm. These are being used in institutes including National Ignition Facility at Lawrence Livermore National Laboratory with the goal to generate the THG pulses 1.8 MJ of pulse energy and 500 TW of peak power.

The range of wavelengths that can be phasematched can be extended by way of SFM. Besides the harmonics of common 1064 nm lasers, the eighth harmonic of 1547 nm from erbium-doped fiber amplifier was generated by the SFM between the fundamental at 1547 nm and the seventh harmonic at 221 nm to generate 193.4 nm (Ohtsuki *et al.*, 2000), corresponding to the ArF wavelength, for the application of semiconductor inspection.

Since the efficiency of nonlinear frequency conversion is higher with higher input powers, pulsed lasers are well suited for efficient generation by taking advantage of high peak powers, opening opportunities for pulsed UV lasers for material processing. Via-hole drilling is one of the emerging application for THG sources in 340–360 nm, with a few tens of watts of average power with >10 kHz repetition rates, beginning to replace CO_2 lasers. Motivated by the industrial applications with more energy-efficient processes, harmonic conversion, such as THG at 355 nm, 40 W at 266 nm (Nishioka *et al.*, 2003) or 5HG at 213 nm are being developed and some noteworthy results were reported; 103 W at 355 nm (Rajesh *et al.*, 2008) using CBO, 27.9 W at 266 nm (Katsura *et al.*, 2007) using CLBO, and 10.2 W at 213 nm (Katsura *et al.*, 2007) at 10 kHz using CLBO.

As the single-pass harmonic conversion efficiency will not be high for continuous-wave (CW) lasers, the input power at the fundamental must be enhanced. Waveguide devices would maintain high intensity at the fundamental, but such devices for UV generation have not been hugely successful so far. Another technique to enhance the power at the fundamental to incident on the nonlinear crystal is the use of optical resonator. Like intracavity-doubled Ar^+ laser, a nonlinear crystal can be placed in a laser oscillator cavity of a visible laser, generating the UV output. Solid-state counterpart, recently mostly pumped with near-infrared semiconductor lasers, emits in the infrared, so there is not many visible solid-state lasers. For one, Pr:YLF lasers had been pumped with semiconductor lasers to oscillate at 522 nm, and a nonlinear crystal in the laser cavity had generated the DUV output at 261 nm (Ostroumov and Seelert, 2008).

6. Harmonic conversion in external resonators. When a nonlinear crystal cannot be placed in the CW laser resonator, such as frequency doubled source or a laser diode, or a MOPA structure, the nonlinear conversion efficiency can still be enhanced by placing the nonlinear crystal in an external cavity which is in resonance with the incoming fundamental input. The first concept of external resonant frequency doubler was introduced by Ashkin *et al.* (1966). In this first investigation, the resonance of fundamental and that of harmonic are considered. In practice, fundamental-resonant doublers are more commonly used, as the enhancement in the efficiency can be square of the enhancement factor of the resonator, and in 1987, an efficient external resonant doubler was demonstrated to generate 532 nm using a monolithic $MgO:LiNbO_3$ external resonator with 13% conversion efficiency (Kozlovsky *et al.*, 1987).

Fig. 1 shows the conceptual schematic of such external resonator. The input should typically be single frequency, although there were demonstrations of externally resonant doubler with CW multi-longitudinal mode lasers or modelocked lasers. In this section, the principle of external doubling resonator is explained with the assumption of input fundamental being single-frequency. We further assume that the external resonator is a unidirectional ring as schematically depicted in **Fig. 1**.

Let us consider the electric field of the circulating power just inside the input coupler. Following Siegman's convention (Siegman, 1986), we write the phase of the reflection as in phase and the transmission from partial mirrors as shifted by 90 degrees. For the sake of simplicity, the magnitude of electric field is represented by the square root of the optical power. Noting that the SHG acts as a loss to the fundamental in the resonator,

$$it_i\sqrt{P_{\text{in}}} + r_i\sqrt{(1-\delta)(1-\gamma\,P_{\text{circ}})}\,e^{i\phi}\sqrt{P_{\text{circ}}} = \sqrt{P_{\text{circ}}}$$

where r_i and t_i are the field reflectivity and the field transmission of the input coupler, ϕ is the phase shift for one round-trip, δ is the passive loss of the resonator, which is the sum of all the losses such as residual transmission of the high reflector or

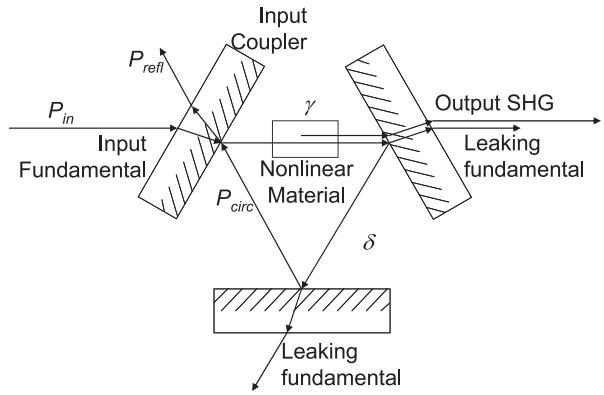

Fig. 1 Schematic of an optical resonator with a nonlinear crystal inside.

scattering/absorption of any optical components/surfaces, and g is the normalized nonlinear conversion efficiency which relates the generated power at the harmonic and the incident power at the fundamental by

$$P_{\mathrm{SH}} = \gamma {P_{\mathrm{circ}}}^2$$

Therefore,

$$\sqrt{P_{\mathrm{circ}}} = \frac{i t_i \sqrt{P_{\mathrm{in}}}}{1 - r_i \sqrt{(1-\delta)(1-\gamma P_{\mathrm{circ}})} e^{i\phi}}$$

For all practical purposes, the term $\delta\ \gamma\ P_{\mathrm{circ}}$ in the square root in the denominator is much smaller than 1, and can be ignored. The circulating power at the fundamental in the resonator is written as, assuming the resonant condition ($\phi = 0$),

$$P_{\mathrm{circ}} = \frac{(1-R_i)P_{\mathrm{in}}}{\left(1 - \sqrt{R_i\ (1-\delta-\gamma P_{\mathrm{circ}})}\right)^2}$$

where r_i^2 is rewritten as R_i as the power reflectivity. It is necessary to numerically solve the equation above to find the circulating power in the resonator. When we look at the reflection from the resonator at the fundamental, it can be written as

$$P_{\mathrm{refl}} = P_{\mathrm{in}} \left(\frac{\sqrt{R_i} - \sqrt{(1-\delta-\gamma\ P_{\mathrm{circ}})}}{1 - \sqrt{(1-\delta-\gamma\ P_{\mathrm{circ}})}} \right)^2$$

This indicates that the reflection becomes zero when $R_i = (1 - \delta - \gamma\ P_{\mathrm{circ}})$, or $1 - R_i = \delta + \gamma\ P_{\mathrm{circ}}$. It means the entire power incident onto the resonator is coupled into it when the transmission of the input coupler equals the total loss of the resonator. This condition is called impedance matching. When the transmission of the input coupler is higher/lower than the total loss of the resonator (including the depletion of the fundamental by the harmonic generation, which depends on the power at the fundamental), it is called "overcoupling/undercoupling."

For example, 1 W of fundamental incident on the resonator with a loss of 0.5%, normalized conversion efficiency of $1 \times 10^{-4}\ \mathrm{W}^{-1}$, input coupler with 2% transmission would generate 0.569 W of harmonic with 75.5 W of fundamental circulating in the resonator. In this condition, only 53 mW is reflected from the resonator, meaning nearly 95% of the input power is coupled into the resonator. The optimum input coupling in this condition is found to be $R_i = 98.7\%$, therefore the input coupler $R_i = 98\%$ is slightly overcoupling. Single-pass interaction with the same fundamental power would yield 0.1 mW, therefore this resonator is providing more than 5000 times enhancement of the harmonic power. The enhancement factor, the ratio between the circulating fundamental inside the resonator and the incident power onto the resonator, is a good indication of the performance of the resonant doublers. Some examples are shown in **Fig. 2** for different values of δ and R_i. Evidently the lower the resonator loss, the higher the enhancement factor, therefore the low-loss optical components are critical to the performance of the resonant doublers.

Examples of impedance matching are shown in **Fig. 3**. In this example, normalized efficiency of $1 \times 10^{-4}\ \mathrm{W}^{-1}$ and 1 W of input power is assumed. For different values of resonator loss, the impedance matching occurs at different input couplings.

The resulting output powers at the harmonic are shown in **Fig. 4** for different resonator optical losses. The optimum output is obtained at the impedance-matched coupling. As was indicated in the previous example, hundreds of milliwatt at the harmonic can be obtained from 1 W of fundamental.

As one can see in **Fig. 4** as well as 2, it is evident that small optical loss of the resonator is critical for the efficient conversion in the CW mode, in which single-pass conversion efficiency is not high. Such optical losses come from resonator mirrors, the bulk scattering or absorption of nonlinear crystal, its surfaces, or any components in the resonator. Proper quality control of these components and characterization is important for practical devices.

As shown, external resonant doublers are powerful tool to efficiently generate second harmonic in CW in which the efficiency would be very low in single-pass interactions. This approach had been used in the UV generation for industrial as well as scientific applications. Here are some of the examples: 80 mW of 257 nm output was demonstrated by the second harmonic of argon ion

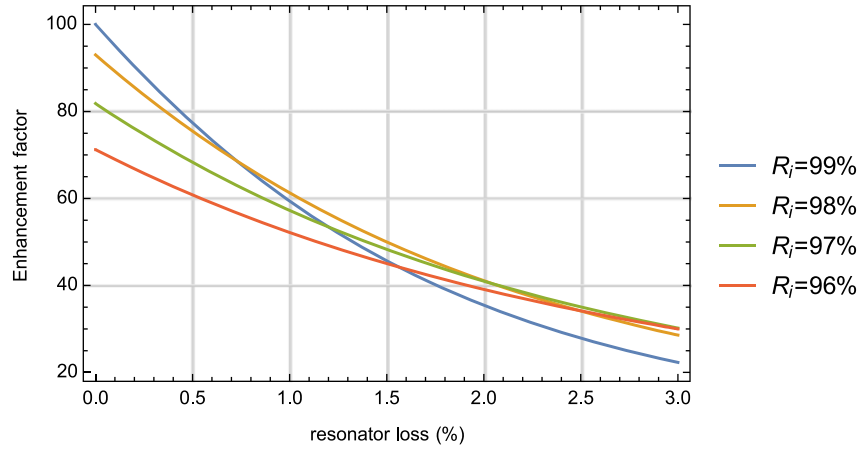

Fig. 2 Enhancement factors as the function of resonator loss.

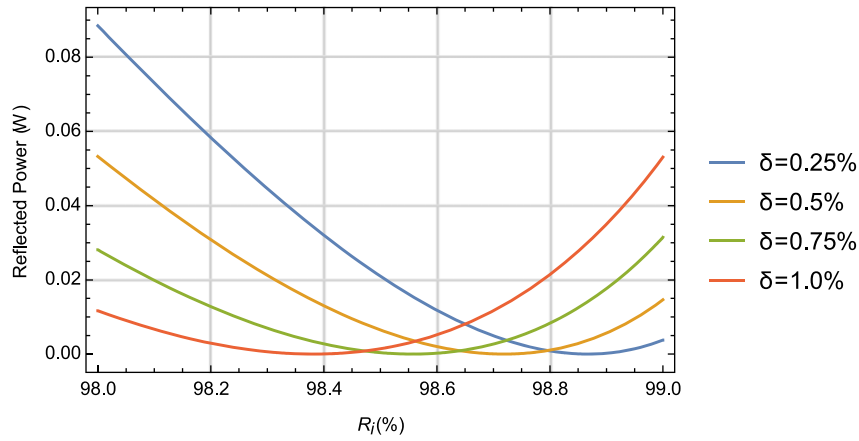

Fig. 3 Reflected power at the fundamental as the function of input coupler.

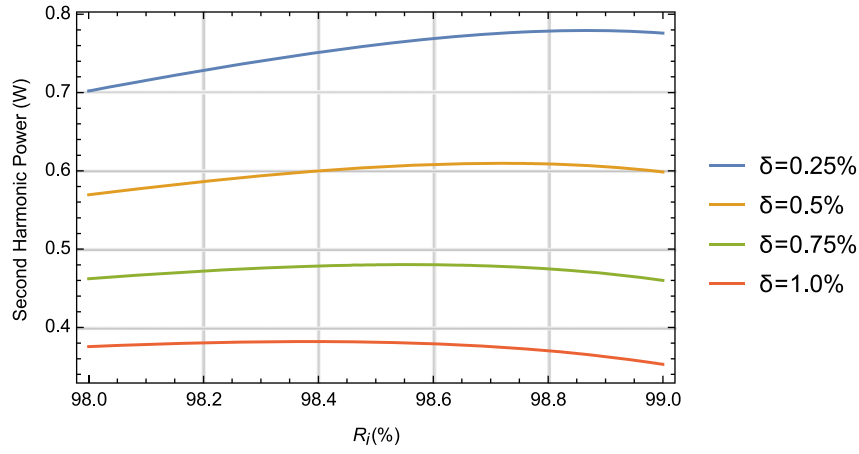

Fig. 4 Second harmonic output power as the function of input coupler.

laser in an external cavity for spectroscopy applications (Bergquist *et al.*, 1982). Fourth harmonic of 1064 nm near infrared laser for industrial applications was demonstrated with up to 12 W of CW output at 266 nm with 24 W of infrared power (Sudmeyer *et al.*, 2008), giving 50% efficiency from IR to DUV conversion efficiency. Owing to its continuous nature, and being single frequency giving very long coherence length, these lasers can be used for metrology or for material processing including interferometric lithography. UV output at tailored wavelengths were used for spectroscopic applications, such as optical lattice clock

(Katori *et al.*, 2003). Trapping/cooling laser for mercury atoms were demonstrated the fourth harmonic of an Yb:YAG disk laser with >750 mW of output power at 254 nm (Scheid *et al.*, 2007), or by the fourth harmonic of an optically-pumped semiconductor laser (OPSL) with more than 100 mW at the same wavelength (Paul *et al.*, 2011) were demonstrated. Trapping/cooling laser for cadmium atoms at 229 nm was demonstrated by the fourth harmonic of an OPSL with 0.56 W of output, which was used to resolve all the stable isotopes of cadmium atoms for the first time (Kaneda *et al.*, 2016).

In this section, we went over only harmonic conversion, but SFM technique is also available to extend the phasematching limit of a material toward the shorter wavelength. One of the examples is the SFM between 234 and 1110 nm to generate 193.4 nm in CW (Sakuma *et al.*, 2015), which is compatible with ArF lasers, using CLBO crystals, which phasematches for 238 nm or longer for direct SHG.

Conclusion

In this article, different types of UV lasers were reviewed, and practical aspects of nonlinear frequency conversion into the UV region were explained. A few borate crystals that are capable of harmonic conversion into the UV were raised and reviewed. Application of external resonator for harmonic conversion is reviewed and its principle is explained. As described, UV lasers carry its usefulness in applications including scientific as well as manufacturing, especially semiconductors with small feature sizes. With the demand driven by Moore's law, development of UV lasers seems to continue toward shorter wavelengths and higher powers. Laser sources near 200 nm, some in sub-200 nm region, are actively developed and their powers are of the order of watts for the applications in semiconductor inspection, >tens of watts for the exposure tools. Material processing also seems to be moving toward shorter wavelengths, and multiple watts to tens of watts of UV lasers are developed and deployed in the field.

References

Ashkin, A., Boyd, G.D., Dziedzic, J.M., 1966. Resonant optical second harmonic generation and mixing. IEEE Journal of Quantum Electronics 2 (6), 109–124.

Bergquist, J.C., Hemmati, H., Itano, W.M., 1982. High power second harmonic generation of 257 nm radiation in an external ring cavity. Optics Communications 43 (6), 437–442.

Dubinskii, M.A., Semashko, V.V., Naumov, A.K., *et al.*, 1993. Ce^{3+}-doped Colquiriite. Journal of Modern Optics 40 (1), 1–5.

Kaneda, Y., Yarborough, J.M., Merzlyak, Y., *et al.*, 2016. Continuous-wave, single-frequency 229 nm laser source for laser cooling of cadmium atoms. Optics Letters 41 (4), 705–708.

Katori, H., Takamoto, M., Pal'chikov, V.G., *et al.*, 2003. Ultrastable optical clock with neutral atoms in an engineered light shift trap. Physical Review Letters 91 (17), 173005.

Katsura, T., Kojima, T., Kurosawa, M., *et al.*, 2007. High-power, high-repetition UV beam generation with an all-solid-state laser. In: CLEO/Europe and IQEC 2007 Conference Digest. Munich: Optical Society of America, Munich, p. CA5_3.

Kneissl, M., Rass, J., 2015. III-Nitride Ultraviolet Emitters: Technology and Applications. Springer, p. 208.

Kozlovsky, W.J., Nabors, C.D., Byer, R.L., 1987. Second-harmonic generation of a continuous-wave diode-pumped Nd:YAG laser using an externally resonant cavity. Optics Letters 12 (12), 1014–1016.

Marshall, C.D., Speth, J.A., Payne, S.A., *et al.*, 1994. Ultraviolet laser emission properties of Ce^{3+}-doped $LiSrAlF_6$ and $LiCaAlF_6$. Journal of the Optical Society of America B 11 (10), 2054–2065.

Nishioka, M., Fukumoto, S., Kawamura, F., *et al.*, 2003. Improvement of laser-induced damage tolerance in CsLiB6O10 for high-power UV laser source. In: Conference on Lasers and Electro-Optics/Quantum Electronics and Laser Science Conference. Baltimore, MA: Optical Society of America.

Ostroumov, V., Seelert, W., 2008. 1 W of 261 nm CW generation in a Pr3+: LiYF4 laser pumped by an optically pumped semiconductor laser at 479 nm. In: Proceedings of SPIE, Solid State Lasers XVII: Technology and Devices, San Jose, CA, pp. 68711K-687114.

Ohtsuki, T., Kitano, H., Kawai, H., *et al.*, 2000. Efficient 193 nm generation by eighth harmonic of Er^{3+}-doped fiber amplifier. In: Conference on Lasers and Electro-Optics, paper CPD9.

Paul, J., Kaneda, Y., Wang, T.-L., *et al.*, 2011. Doppler-free spectroscopy of mercury at 253.7 nm using a high-power, frequency-quadrupled, optically pumped external-cavity semiconductor laser. Optics Letters 36 (1), 61–63.

Rajesh, D., Yoshimura, M., Eiro, T., *et al.*, 2008. UV laser-induced damage tolerance measurements of CsB_3O_5 crystals and its application for UV light generation. Optical Materials 31, 461–463.

Sakuma, J., Kaneda, Y., Oka, N., *et al.*, 2015. Continuous-wave 193.4 nm laser with 120 mW output power. Optics Letters 40 (23), 5590–5593.

Scheid, M., Markert, F., Walz, J., *et al.*, 2007. 750 mW Continuous-wave solid-state deep ultraviolet laser source at the 253.7 nm transition in mercury. Optics Letters 32 (8), 955–957.

Siegman, A., 1986. Lasers. California: University Science Books, Mill Valley.

Sudmeyer, T., Imai, Y., Masuda, H., *et al.*, 2008. Efficient 2nd and 4th harmonic generation of a single-frequency, continuous-wave fiber amplifier. Optics Express 16 (3), 1546–1551.

Yoshida, H., Yamashita, Y., Kuwabara, M., *et al.*, 2008. Demonstration of an ultraviolet 336 nm AlGaN multiple-quantum-well laser diode. Applied Physics Letters 93 (24), 241106.

Further Reading

Wernicke, T., Martens, M., Kuhn, C. *et al.*, 2015. Challenges for AlGaN Based UV Laser Diodes. DM2D.4.

Single-Frequency Lasers

Xiushan Zhu, University of Arizona, Tucson AZ, United States

Introduction

Single-frequency lasers are a category of lasers operating on only one longitudinal mode. Single longitudinal mode operation of a laser is achieved by inserting filters into the cavity to suppress the oscillation of all the other mode except the mode with the lowest threshold. The cavity of a single-frequency laser usually consists of a gain medium, an output coupler, one or two filters, and a cavity mirror, as shown in **Fig. 1**. Since a single frequency laser is susceptible to destabilization from backward reflection, Faraday isolator is always used after the output coupler to ensure high performance operation of single longitudinal mode. When a laser operates with only one longitudinal mode, it offers excellent features including ultra-narrow spectral linewidth, extreme stability in intensity and phase, and ultra-long coherence length, which are highly required for a variety of applications. A single-frequency laser can usually provide an amount of wavelength tunability, which depends on the bandwidth of the gain medium and the cavity design. Continuous-wave (CW) and Q-switched single-frequency lasers have been achieved with almost all types of current lasers. The average output power of single-frequency lasers can be elevated to hundreds of watts or even kW level by use of current laser amplifier technologies. Single-frequency lasers, both in CW and pulsed operation, have found a variety of applications in interferometric sensing, coherent light detection and ranging (LIDAR), coherent optical communication, high resolution spectroscopy, optical data storage, optical cooling and trapping, microwave photonics, and laser nonlinear frequency conversions.

Background

In their classical paper proposing an "optical maser", single frequency operation of a laser was predicted by Schawlow and Townes (1958). However, single longitudinal mode operation was not observed in the first laser action in ruby (Maiman, 1960). Nevertheless, owing to the narrow gain bandwidth of gas gain medium, single-frequency operation of He-Ne laser was demonstrated by Kogelnik and Patel (1962) by using a resonant reflector acting as a spectral filter as well at one end of the resonator. From then on, different types of longitudinal mode selection schemes were used to achieve single longitudinal mode operation of a laser. Fabry-Perot etalons, for example, have been extensively used as interferometric filters in single-frequency solid-state lasers since Collins and White were the first to place tilted etalons inside the cavity (Collins and White, 1963). A single frequency laser can emit quasi-monochromatic radiation with very narrow linewidth and low noises, which are advantageous for diverse scientific investigations and engineering applications. For instance, intense research on single-frequency lasers resulted in the great discovery of the "Lamb dip" in 1963 (McFarlane *et al.*, 1963).

Spatial Hole Burning and Multimode Operation

Generally, the line-shape of stimulated emission cross-section of an ideal homogeneously broadened laser transition is fixed and the same for all atoms in the laser medium. Thus, the macroscopic gain of the laser medium at a given frequency depends only on the population inversion and the gain profile is just the constant atomic line-shape multiplied with a scalar factor, which is proportional to the population inversion. Therefore, in theory, a laser with an ideal gain profile always oscillates with one single longitudinal mode, which has a net gain of unity, i.e. under steady-state conditions the gain of this lasing mode exactly equals its losses, and all other modes experience a net gain of less than unity and will not lase. As shown in **Fig. 2**, when the population inversion is small, the net gains of all modes are less than the losses and all of the modes are below the threshold. As the population inversion increases, the net gains of all modes increase correspondingly and that of mode m equals the losses giving a net gain of unity. Therefore, only mode m reaches the threshold and starts to lase and all other modes are still below threshold. In this case, the laser operates on single longitudinal mode.

However, different from the prediction of Schawlow and Townes, people found that a laser usually tends to operate on multiple longitudinal modes when there is no mode selection element inside the laser cavity. The most significant effect leading to multi-longitudinal-mode operation in homogeneously broadened lasers is spatial inhomogeneity of the gain, especially "spatial

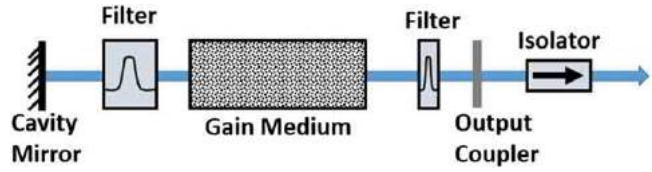

Fig. 1 Basic configuration of a single-frequency laser.

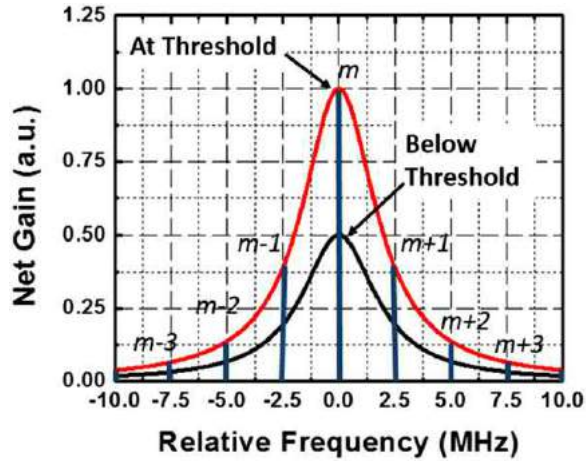

Fig. 2 Gain profile of an ideal homogeneously broadened laser transition and the net gain of corresponding longitudinal modes.

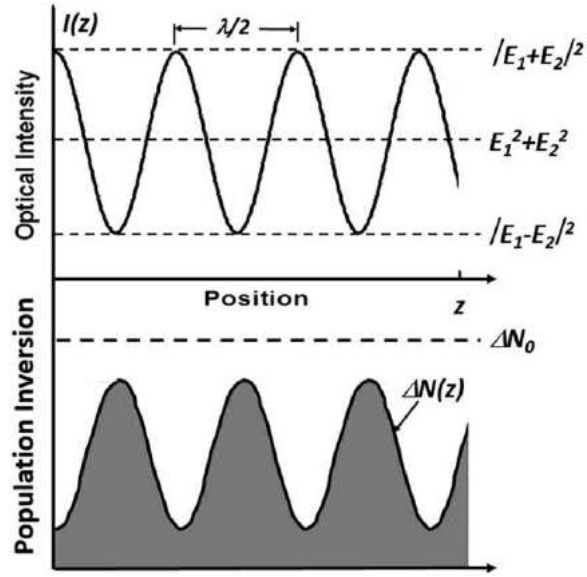

Fig. 3 Standing wave and the corresponding population inversion versus position in a laser gain medium.

hole burning" caused by the standing wave of the laser in the gain medium (Tang *et al.*, 1963). As shown in **Fig. 3**, in a linear cavity resonator, the forward (E_1) and the backward (E_2) propagating laser beams form a standing wave which depletes the gain much more strongly at the peaks of the standing wave than at the nodes. As a result, the population inversion becomes much larger at the nodes than at the peaks of the standing wave, thus leading to a spatial imhomogeneity of the gain. Since the bandwidth of the laser transition is normally much broader than the mode spacing, some adjacent modes will start to lase when these modes have their peaks near the peak positions of the population inversion and obtain a net gain equal to the losses as well. Therefore, due to the spatial hole burning, in a typical laser a few modes always oscillate simultaneously even if the gain of the adjacent modes is slightly smaller than that of the center mode. When a laser operates on multiple longitudinal modes, on the time scale of the round trip time (inverse of the mode spacing), the intensities of these modes change chaotically due to the beating between these lasing modes. On a time scale much longer than the cavity round trip time, the laser output power becomes fairly stable because it is the sum of the intensities of these lasing modes. Therefore, the intensity noise of a single-frequency laser is much lower than that of a laser operating on multiple longitudinal modes.

Single Mode Operation with Eliminated Spatial Hole Burning

There are basically two fundamental routes to achieve single-frequency operation of a laser. Because spatial hole burning is identified as the essential reason for multi-longitudinal-mode operation of a laser, the straightforward solution is to avoid spatial hole burning by preventing standing waves in the resonator. The alternative solution is to suppress the adjacent modes by

employing different mode-selection schemes to compensate or overwhelm the gain difference between the center mode and the adjacent modes caused by spatial hole burning.

The two most widely used methods to avoid the formation of standing waves are: the travelling wave laser and the "twisted-mode" laser. In the first paper explaining the role of "spatial hole burning", Tang *et al.* (1963) suggested to develop travelling wave laser, such as unidirectional ring laser, to achieve the single frequency operation of a laser without using any mode-selection elements. In a unidirectional ring laser, as shown in **Fig. 4(a)**, the light is forced to travel in one direction only by a Faraday isolator or an acoustic optical modulator and consequently standing waves are not formed. In 1965, Evtuhov and Siegman (1965) demonstrated that standing waves could be avoided by a "twisted-mode" technique, in which the laser medium is placed between two quarter-wave plates and the two counter-propagation waves are transformed into circular polarized light before entering the laser medium as shown in **Fig. 4(b)**. The two circularly polarized waves travelling in the opposite direction do not interfere with each other if they have the same sense and therefore standing waves are not formed inside the gain medium and the gain is depleted uniformly. However, "twisted-mode" technique is sensitive to stress-induced birefringence caused by thermal load and mechanical stress. In addition, it cannot be used with birefringent laser materials.

Single Mode Operation with Mode Suppression

Although the adjacent modes and the center mode may experience the same gain of unity because of the spatial hole burning, single longitudinal mode operation can still be achieved if all longitudinal modes except the lasing mode are sufficiently suppressed by using different mode selection mechanisms.

The common approach is to incorporate an interferometric filter into the cavity to introduce much higher loss to all other modes than the lasing mode. Etalons acting as Fabry-perot filters have been widely used to achieve single-frequency operation of a laser. However, this approach requires the lasing mode to be at the minimum of the loss curve. Active locking techniques have to be employed to ensure stable single frequency operation. An alternative approach is to increase the mode spacing and reduce the gains of adjacent modes to values that are still smaller than unity even the effect of spatial hole burning is added. In a short cavity, the mode spacing becomes larger than the spectral width of a gain profile and thus single-frequency laser can be achieved. Sometimes, when the gain profile of a laser transition is broad or the bandwidth of the interferometric filter is larger than the mode spacing, both approaches have to be used to achieve single frequency operation as shown in **Fig. 5**.

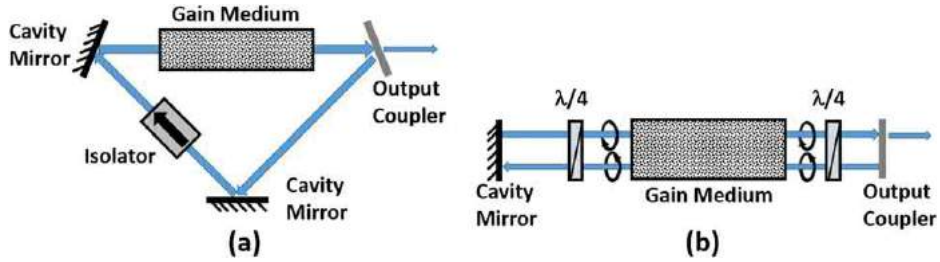

Fig. 4 (a) Configuration of a unidirectional ring cavity laser and (b) configuration of a twisted-mode laser.

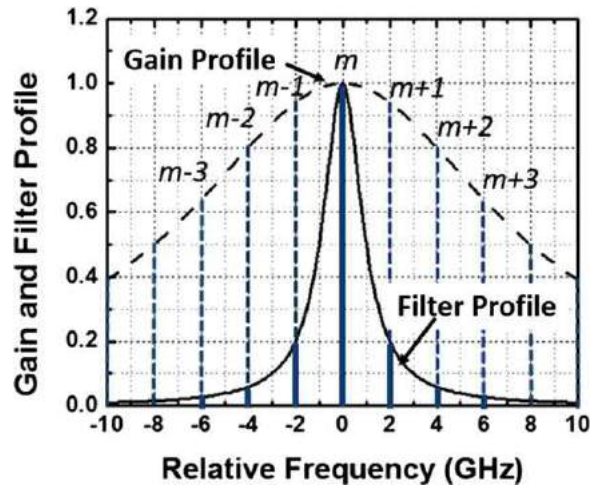

Fig. 5 Single-frequency operation of a laser achieved by using an interferometric filter and a short cavity with mode spacing of 2 GHz.

It should be noted that unidirectional ring laser and twisted-mode laser can effectively eliminate spatial hole burning and achieve single frequency operation with a homogeneously broadened gain medium. The two techniques, however, cannot be independently used to achieve single-frequency laser operation in some laser materials which have broad gain profiles or have both homogeneously and inhomogeneously broadened laser transitions. In this case, the two techniques have to be combined with the techniques of interferometric filters and short cavity to achieve single-frequency laser.

It should be noted that, due to the Guoy phase shift, a multi-transverse-mode laser doesn't operate with single-frequency even if the laser design meets the conditions for single longitudinal mode operation. Guoy phase shift depends on the order of the transverse modes and thus different transverse modes in a stable laser cavity have different oscillation frequencies. Therefore, a single-frequency laser has to be operated on single transverse mode as well.

Overview

Single longitudinal mode operation has been achieved in almost all types of current lasers by using different techniques discussed above or their combinations. The single frequency laser sources most available today are short cavity gas (i.e. He-Ne) lasers, ring cavity dye lasers, ring cavity or short cavity or twisted-mode solid-state lasers, ring cavity fiber lasers, semiconductor or fiber distributed feedback (DFB) lasers, semiconductor or fiber distributed Bragg reflector (DBR) lasers, and external cavity diode lasers.

Owing to the narrow gain bandwidth of gas, it is easy to develop a single frequency gas laser with a short length cavity ($\sim$20 cm or less). Ring cavity dye lasers can also easily operate on single longitude mode because of their homogenously broadened laser transitions. Single-frequency dye lasers have the advantage of widely tunable wavelength ranges. However, gas and dye lasers are not widely used today because they are not efficient, reliable, robust, and compact as current solid-state lasers, semiconductor lasers, and fiber lasers.

Single frequency operation in solid-state lasers have been achieved with various techniques, such as microchip lasers, lasers inserted with etalons, twisted-mode lasers, and nonplanar ring oscillators (NPRO). A michrochip laser usually has such a short cavity (less than 1 mm) that the free spectral range is much larger than the gain bandwidth and thus robust single-longitudinal-mode operation can be easily achieved (Kane and Byer, 1985). The techniques of using etalon filters and twisted-mode have already been discussed in detailed in Section 2. A NPRO is a unidirectional ring laser with special monolithic design in which the laser beam circulates in a single laser crystal (Zayhowski and Mooradian, 1989).

Nonplanar Ring Oscillator

As shown in **Fig. 6**, the resonator of an NPRO laser is a laser crystal cut at specific shapes enabling the unidirectional circulation of the laser beam via partial reflection of the dielectric mirror coated on the front face and total internal reflection on all other internal faces. The dielectric mirror coated on the front face is highly transmissive for the pump light and partially transmissive for the laser light and thus serves as the output coupler mirror of the resonator.

Unidirectional propagation of the laser light is obtained because the cavity loss of one oscillation direction is larger than that of the other one. Since the resonator is nonplanar, the polarization direction of the laser light is rotated slightly after each round trip. The two polarization rotations of one oscillation direction cancel partly, resulting in a smaller cavity loss when the laser light hits the front face of the crystal because the reflectivity of dielectric coating is polarization dependent. The other oscillation direction, however, experiences a larger cavity loss and is thus firmly suppressed. The discrimination between the two oscillation directions can be enhanced by attaching a small magnet to the laser crystal to increase the polarization rotation via the Faraday effect. Because a NPRO operates with travelling wave and the round trip length is short ($<$10 cm), no spatial hole burning and large free spectral range make it very easy to obtain stable single-frequency operation. The relative large free spectral range also allows mode-hop free wavelength tuning over several gigahertz. The wavelength tuning can be achieved by applying a piezoelectric transducer on the laser crystal, incorporating an electro-optic crystal in the resonator, changing the crystal temperature, or adjusting the pump power.

NPROs have been developed with Yb: YAG at 1.03 μm, Nd: YAG at 1.06 μm, Er: YAG at 1.645 μm, Tm: YAG at 2.01 μm, and Ho: YAG at 2.09 μm. Because NPROs are monolithic and robust and have low optical losses, their laser noises are very small and

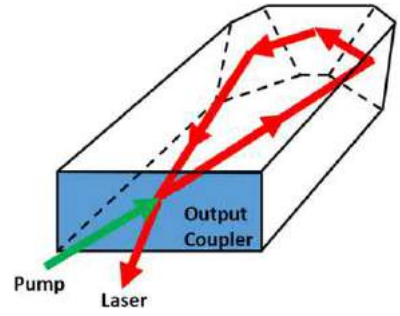

Fig. 6 Configuration of a nonplanar ring oscillator.

their typical linewidths can be a few kilohertz. Most importantly, the output power of a single-frequency NPRO can be up to several watts. Therefore, NPROs are often used as master oscillator for high-power single-frequency master oscillator and power amplifier laser system.

Because one of the prerequisites of single frequency laser is that its transverse mode must be single mode, complex resonator cavity design and fabrication and careful alignment are required to obtain single frequency operation in solid-state lasers, gas lasers, and dye lasers. Using single-mode waveguide, such as fiber and semiconductor, is therefore much easier to achieve stable and robust single-frequency operation of a laser. Fiber lasers and semiconductors lasers have become key workhorses for single frequency laser sources due to their compact sizes and robust single-longitudinal mode operation.

Ring Cavity Fiber Lasers

Fiber lasers have the advantages of compactness, reliability, cost-effective, alignment-free, and maintenance-free. Single-frequency fiber lasers combining the advantages of fiber lasers and single-frequency lasers have been thoroughly investigated and extensively used for various applications. However, a conventional fiber laser generally has meters of fiber length with linear cavity configuration and thus cannot generate single-frequency laser output due to the spatial hole burning in the gain medium as discussed above. This problem is usually solved by use of either a unidirectional ring cavity combined with a narrow-band filter or a very short (a few centimeters long) linear cavity combined with narrow-band fiber Bragg gratings capable of selecting a single longitudinal mode despite the presence of spatial hole burning.

Because the long fiber cavity can provide high gain and the potential for wide wavelength tunability by incorporating a wavelength tunable component into the ring cavity, the unidirectional fiber ring laser is very attractive for single-frequency laser development with high output power and wide wavelength tunable range. The lack of a spatial hole-burning effect in a travelling-wave field renders a ring-cavity design more suitable for single-frequency operation of a fiber laser with lower phase noise and a correspondingly smaller linewidth compared to linear cavity configuration. The configuration of a conventional single-frequency ring cavity fiber laser is shown in **Fig. 7**. The gain fiber is pumped by a pump laser source via a wavelength division multiplexer (WDM). Unidirectional propagation of the laser in the fiber ring is achieved with an isolator and the laser is coupled out from the fiber ring by a fiber coupler. Single-frequency laser with wide wavelength tunable range is realized by a tunable thin-film filter (TTF). Stable single-frequency operation of a fiber ring laser can also be achieved by other techniques such as using a passive multiple ring cavity or a compound ring resonator (Yeh *et al.*, 2006), using two cascaded Fabry-Perot filters (Park *et al.*, 1991), incorporating two or three fiber Bragg gratings (Guy *et al.*, 1995), and utilizing a saturable-absorber-based auto-tracking filter (Yeh and Chi, 2005). However, unidirectional ring lasers need expensive components such as Faraday isolators and TTFs and still have a tendency to mode hop that may have to be eliminated with an additional narrow-band optical filter in the fiber ring cavity.

Distributed Feed-back Lasers

As discussed above, single-frequency operation of a linear-cavity laser can be achieved with very short length cavity and narrow spectral filters. Distributed feedback (DFB) laser is a typical type of single-frequency laser in which the active region is periodically structured as a diffraction grating (Kogelnik and Shank, 1972). The active region thus not only provides optical gain for the laser but also optical feedback and mode selection for the laser.

DFB lasers either semiconductor lasers or fiber lasers have a typical configuration as shown in **Fig. 8(a)**. The whole resonator consists of a periodic structure, which acts as a diffraction grating and contains a gain medium, and a highly reflective mirror.

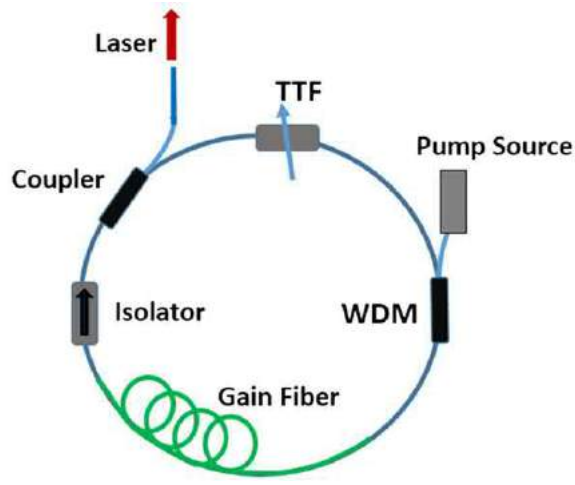

Fig. 7 Configuration of a single-frequency ring-cavity fiber laser.

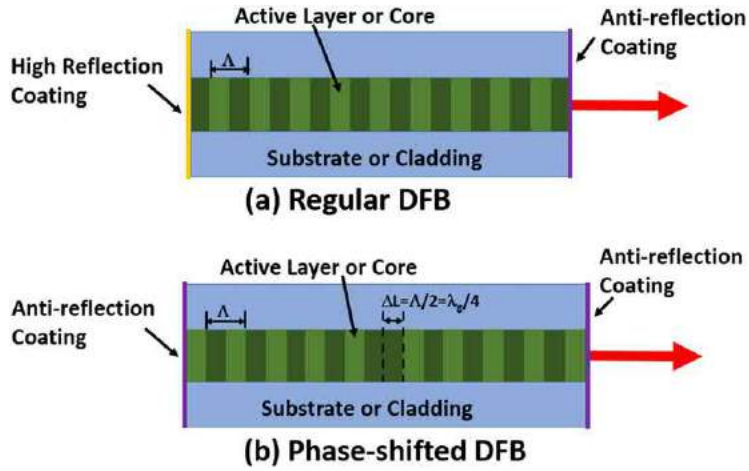

Fig. 8 Configuration of (a) regular DFB and (b) phase-shifted DFB single-frequency lasers.

The highly reflective mirror and the diffraction grating form the optical cavity. The diffraction grating is inscribed in the gain medium so as to reflect only a narrow band of wavelengths, and thus allow single longitudinal mode operation of the laser. The lasing wavelength for a DFB laser is approximately equal to the Bragg wavelength, which is defined by:

$$\lambda = 2n_{eff}\Lambda \tag{1}$$

where λ is the wavelength, n_{eff} is the effective refractive index of the guided laser mode, and Λ is the grating period. Because the effective refractive index and grating period are generally temperature dependent, altering the temperature of a DFB laser can change the wavelength selection of the grating and thus the wavelength of the laser output, producing a wavelength tunable laser. For a DFB laser either fiber or diode-based, the temperature tunability is about 0.01 nm/k. Altering the pump current of a DFB semiconductor laser will also lead to wavelength tuning.

An alternative approach is a phase-shifted DFB laser with a configuration as shown in **Fig. 8(b)**. Typically, the periodic structure is made with a phase shift in its middle. This could be a single quarter-wave shift at the center of the cavity, or multiple smaller shifts distributed in the cavity. Due to the large free spectral range and ultra-narrow spectral filtering of a phase-shifted DFB laser, robust single-frequency operation is often easily achieved with excellent wavelength reproducibility and narrow linewidth.

For a DFB fiber laser, the distributed reflection occurs in a fiber Bragg grating, typically with a length of a few millimeters or centimeters. Efficient pump absorption can be achieved with a highly doped fiber such as phosphate glass fiber and the output power can be tens of milliwatts or over 100 mW. DFB fiber laser has excellent compactness and robustness leading to a low intensity and phase noise level, i.e., also a narrow linewidth. The linewidth of a DFB fiber laser is typically less than 100kHz.

Semiconductor DFB lasers can be built with an integrated grating structure, e.g. a corrugated waveguide, in a very compact package. Semiconductor DFB lasers are available for emission in different spectral regions (0.8–2.8 μm). Typical output powers are some tens of mW. The linewidth is typically a few MHz to several hundred MHz.

Distributed Bragg Reflector Lasers

Distributed Bragg reflector (DBR) laser is another type of single-frequency laser based on linear cavity design. DBR lasers either semiconductor lasers or fiber lasers have typical configurations as shown in **Fig. 9**. Different from a DFB laser, where the whole active medium is embedded in a single distributed reflector structure, a DBR laser has at least one DBR outside the gain medium. A resonant cavity is defined by a highly reflective DBR or thin-film mirror on one end, and a low reflectivity cleaved exit facet or DBR on the other end. The DBR mirror is designed to reflect only a single longitudinal mode. As a result, single-frequency operation of the laser is obtained.

DBR and DFB lasers have distinct operational characteristics caused by their different locations of the feedback elements. Because the grating of a DFB is distributed all along the gain region, the grating and gain region experience similar conditions as the device is tuned with current and temperature. The DFB can exhibit a continuous tuning range of 2 nm or more. However, over a sufficiently long current or temperature range, the emission wavelength will suddenly jump to a longer wavelength, leaving a gap in the tuning range. Because the DBR laser has a passive grating region, its tuning characteristic of the DBR laser is different from that of the DFB laser. Increasing current in the gain region causes a red shift in laser output due to heating. However, the reflectivity curve of the passive grating does not change and the grating will experience loss of reflectivity at the longer wavelengths. As a result, the wavelength characteristic of a DBR laser will repeat itself with increasing temperature or current, and no gaps will occur in the tuning.

The DBR diode laser is fabricated with a monolithic, single mode ridge waveguide running the entire length of the device (Dupuis and Dapkus, 1978). The wavelength tunable range of a DBR diode laser is approximately a 2 nm range by changing

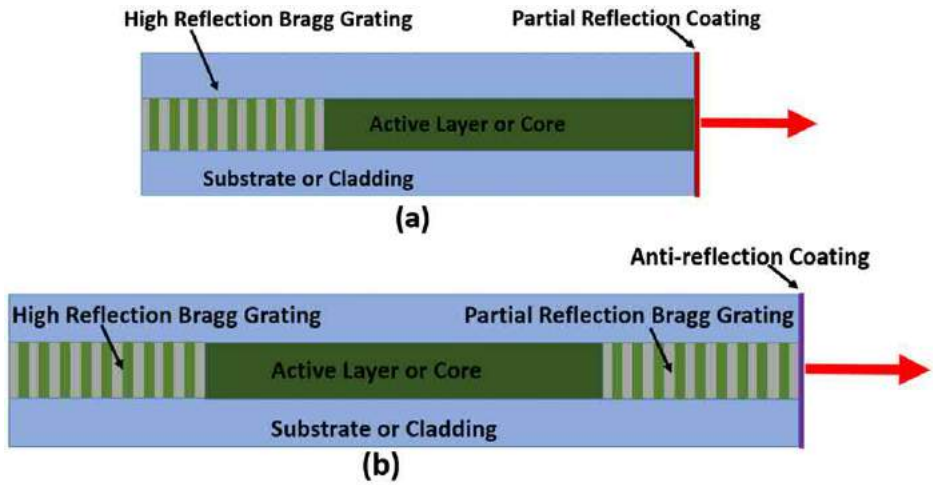

Fig. 9 Configuration of single-frequency DBR lasers with (a) one Bragg grating and (b) two Bragg gratings.

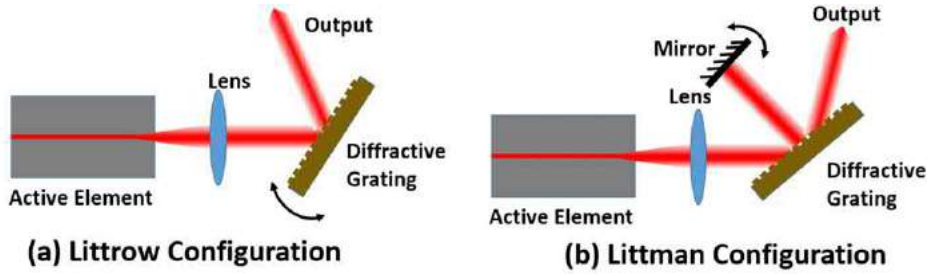

Fig. 10 Configuration of external cavity single-frequency lasers.

current or temperature. The temperature coefficient is approximately 0.07 nm/K, and the current coefficient is approximately 0.003 nm/mA. The typical linewidth of a DBR laser is less than 10 MHz. The output power typically can run up to several hundred milliwatts.

DBR fiber lasers have the advantages of a compact all-fiber design, easy and accurate wavelength selection during production, a large potential wavelength tuning range without mode-hopping through strain or temperature tuning, and narrow linewidth. The temperature coefficient is approximately 0.01 nm/K. A typical linewidth of a DBR fiber laser is less than 10 kHz. The output power typically can run up to several watts with a double-clad phosphate fiber laser (Polynkin *et al.*, 2005; Schulzgen *et al.*, 2006).

External Cavity Lasers

The wavelength tunable range of a DFB or DBR diode laser is typically a few nm. Broad wavelength tunable range (>100 nm) can be achieved with external cavity diode lasers (ECDLs), which usually use bulky grating and free-space optics that is compatible with most standard diode lasers. As shown in **Fig. 10**, a lens collimates the output beam of the diode, which is then incident upon a grating. The grating provides optical feedback for a very narrow spectral light and is used to select the stabilized output wavelength. Single-frequency operation of the ECDLs with high side mode suppression ratio (SMSR >45 dB) and very broad wavelength tunability can be obtained with proper optical design allowing only a single longitudinal mode to lase. Fiber Bragg grating (FBG) can be used to replace the free-space optics and grating to make ECDLs very compact and robust. However, the wavelength tunability of FBG-based ECDLs is usually not as broad as that of free-space ECDLs due to the limited wavelength tunable range of fiber Bragg grating. In addition to excellent wavelength tunability, compared to DFB and DBR diode lasers, ECDLs have much narrower linewidth (<1 MHz) due to the relatively long cavity. The ECDL can have a large tuning range but is often prone to mode hops, which are very dependent on the mechanical design as well as the quality of the antireflection (AR) coating on the laser diode.

FBG-based ECDLs have a very simple configuration that the light from the diode is coupled into the fiber of FBG directly and the wavelength tuning is achieved by heating or mechanically stretching and compressing the FBG. Free-space ECDLs have two different configurations: Littrow (Fleming and Mooradian, 1981) and Littman (Liu and Littman, 1981). The common Littrow configuration is shown in **Fig. 10(a)**. The first-order diffracted beam provides optical feedback to the laser diode and the wavelength tuning is realized by rotating the diffraction grating. A disadvantage of Littrow configuration is that the direction of the

output beam is also changed. The Littman configuration is shown in Fig. 10(b). The grating orientation is fixed, and an additional mirror is used to reflect the first-order beam back to the laser diode. The wavelength tuning is achieved by rotating that mirror. This configuration offers a fixed direction of the output beam. In addition, the Littman configuration intends to exhibit a smaller linewidth than Littrow configuration because the wavelength-dependent diffraction occurs twice per resonator round trip in the Littman configuration. However, the output power of a Littman laser is less than that of a Littrow laser because the zero-order reflection of the beam reflected by the tuning mirror is lost.

Characterization

Single-frequency lasers exhibit excellent features of low noises and narrow spectral linewidth. Relative intensity noise (RIN), phase noise or frequency noise, and spectral linewidth are key parameters defining the performance of a single-frequency laser.

Relative Intensity Noise

RIN is a typical parameter to characterize the optical power fluctuation of a laser and can be expressed as

$$RIN(t) = \frac{\langle \delta P(t)^2 \rangle}{\langle P(t) \rangle^2} \tag{2}$$

where $\delta P(t)$ is the power fluctuation of the laser and $\langle P(t) \rangle$ is the average power. RIN can also be described with a power spectral density depending on the noise frequency f and be calculated as the Fourier transform of the autocorrelation function of the normalized power fluctuations:

$$RIN(f) = \frac{2}{\langle P(t) \rangle^2} \int_{-\infty}^{+\infty} \langle \delta P(t) \delta P(t+\tau) \rangle e^{i2\pi f \tau} d\tau \tag{3}$$

RIN of a laser is often measured in the electrical domain with a photodiode and analyzed with an electronic spectrum analyzer (ESA) using a setup as shown in Fig. 11(a). The optical power of the laser is directly detected by an optical detector. The RIN is simply the ratio of the DC photocurrent to the AC noise on the detector output and then displayed on the ESA. The electric AC noise is proportional to optical power squared and hence RIN is usually presented as relative fluctuation in the square of the optical power specified in dB/Hz.

Intensity noise of a laser results partially from quantum noise associated with laser gain and resonator losses and partially from vibrations and thermal issues of the cavity, fluctuations in the laser gain medium, and the excess noise of the pump source. The RIN of a single-frequency laser is typically independent of laser power and is shown in Fig. 11(b). RIN typically has a peak at the relaxation oscillation frequency of the laser and then decreases with the increased frequency until it converges to the shot noise level. Since the relaxation oscillation is related to the emission transition of the gain medium, the relaxation oscillation frequency of DFB or DBR diode lasers is usually larger than 1 GHz while that of solid-state lasers and fiber lasers is typically in a frequency range between a few tens kHz and a few MHz. The relaxation frequency peak of RIN can usually be suppressed with a close-loop servo system. The RIN with suppressed relaxation oscillation frequency peak is shown in the inset of Fig. 11(b).

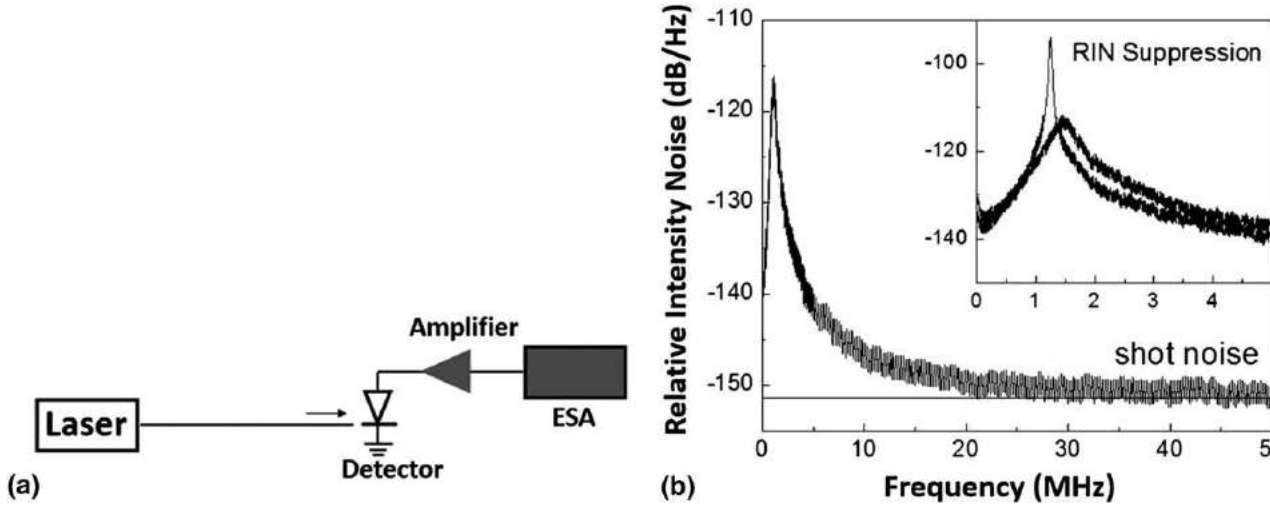

Fig. 11 (a) Configuration of setup for RIN measurement; (b) typical RIN of a single-frequency DBR fiber laser at 1.55 μm. Inset: relaxation oscillation frequency peak is suppressed.

Phase or Frequency Noise

Phase noise is the random deviation of the oscillation phase from a purely linear phase evolution, which results in random fluctuation of the instantaneous frequency of the single-frequency laser, namely frequency noise. Hence the output of a single-frequency laser is not a perfectly monochromatic but with a finite linewidth. Phase noise is mainly caused by quantum noise, cavity noise and intensity noise. The power spectral density of single-sided phase noise includes three contributions: white, flicker and random walk noises, and can be written as

$$S_\phi(f) = \frac{\delta(v)}{\pi f^2} + \frac{k_f}{f^3} + \frac{k_r}{f^4} \tag{4}$$

where $\delta(v)$ is the Lorentzian spectral linewidth of the laser, k_f and k_r are constants that give the strength of flicker frequency noise and random walk frequency noise, respectively. The power spectral density of frequency noise is directly related to that of the phase noise by the equation

$$S_{(v)}(f) = f^2 S_\phi(f) \tag{5}$$

Phase noise or frequency noise is generally directly measured by frequency discriminators, which convert frequency fluctuation to intensity fluctuation and then measure the power spectral density, $S_v(f)$, of the optical frequency fluctuations by detecting the intensity variations. Frequency discriminator is usually constructed with a configuration of interferometer, such as a Michelson interferometer, a Mach-Zehnder interferometer or a Fabry-Perot interferometer. A typical frequency discriminator based on an all-fiber Michelson interferometer is shown in **Fig. 12(a)**. The single-frequency laser is launched into one left port of a 2×2 50/50 fiber coupler and split into the two arms of the fiber coupler. At the two right ports of the fiber coupler, two mirrors are used to reflect the laser back into the fiber coupler. The output at the other left port is detected by an optical detector and analyzed by a dynamic spectrum analyzer (DSA). The measured frequency noise of single-frequency DBR fiber lasers at 1550 and 1060 nm are shown in **Fig. 12(b)**. The frequency noise typically decreases with the increased frequency and exhibits some spikes related to various noises. Generally, a single-frequency laser with a lower frequency noise has a narrower linewidth. However, frequency discriminator does not directly yield linewidth. The fundamental laser linewidth can be determined from the power spectral density of the optical frequency. Direct measurement of the spectral linewidth of a single-frequency laser should be accomplished with heterodyne methods.

Spectral Linewidth

Due to quantum mechanical fluctuations, cavity vibrations and fluctuations, gain medium fluctuations, and other laser noises, a single-frequency laser is not perfectly monochromatic and has finite spectral linewidths. The spectral linewidth of a single frequency laser is the full width at half maximum (FWHM) of the optical spectrum. When a laser only has quantum fluctuations, it has the Schawlow-Townes linewidth,

$$\Delta v = \frac{4\pi h v (\Delta v)^2}{P_{out}} \tag{6}$$

where $h(v)$ is the photon energy, $\Delta(v)$ is the resonator bandwidth, and P_{out} is the output power. Spontaneous emission of excited atoms and ions is a major quantum noise and makes the laser output has a finite spectral width. Nevertheless, the spectral linewidth of a single-frequency cannot be measured with typical grating-based optical spectrum analyzers or scanning filter methods. Frequency discriminator, heterodyne detection, and self-heterodyne detection have to be used for high resolution linewidth measurement. The spectral linewidth of a single-frequency laser cannot be directly obtained with frequency

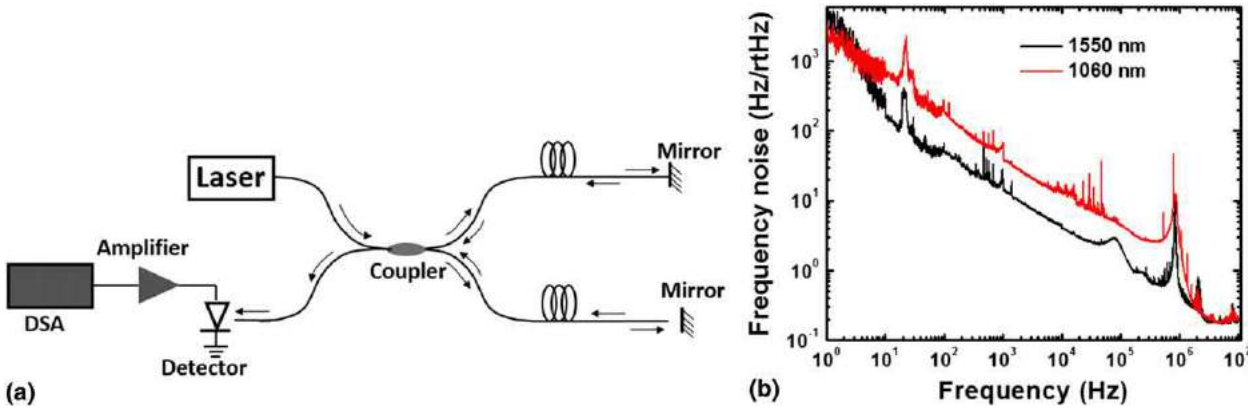

Fig. 12 (a) Configuration of the setup for frequency noise measurement and (b) typical frequency noise of single-frequency DBR fiber lasers at 1550 and 1060 nm.

discriminator but can be deduced from the data of frequency noise. Heterodyne and delayed self-heterodyne detection methods capable of measuring the spectral linewidth of laser directly have been extensively used.

Heterodyne detection

The basic configuration of heterodyne detection is shown in Fig. 13. A very stable single-frequency laser is used as the reference, namely local oscillator laser. The single-frequency laser under test is the signal laser. The central frequency of the local oscillator laser must be tuned close to that of the signal laser to allow the beat frequency to fall within the bandwidth of typical detection electronics. The local oscillator laser and the signal laser under test are launched into the two input ports of a 2×2 50/50 fiber coupler, respectively. An optical spectrum analyzer (OSA) and a photodetector are connected to the two output ports of the fiber coupler, respectively. The OSA is used for course wavelength tuning. The photodetector is used to detect the interference beat note and converts it to an electrical tone displaying on an electrical spectrum analyzer.

Heterodyne detection method can not only detect the spectral linewidth of a single-frequency laser, but also measure its optical power spectrum. This method is the only technique that is capable of characterizing non symmetrical spectral lineshape. The key component required for this method is a stable, narrow linewidth reference laser. The linewidth of the reference laser must be narrower than or at least comparable to that of the laser source to be measured in order to achieve reasonable measurement accuracy. However, this method becomes inaccurate for extremely narrow linewidth measurements since the characterization of the reference laser itself is very difficult. The applicability of this method is also limited by the bandwidth of the photodetector limiting the measurement frequency range to at most tens of gigahertz vicinity of the reference laser central frequency.

Delayed self-heterodyne detection

Delayed self-heterodyne detection doesn't need a separate local oscillator and the spectral linewidth measurement is based on the interference between the laser under test and a delayed replica of itself. The basic configuration of delayed self-heterodyne detection technique for spectral linewidth measurement is shown in Fig. 14(a). The basic idea of this technique is to convert the optical phase or frequency fluctuations of the laser into variations of light intensity in a Mach-Zehnder type interferometer. The light of the laser under test is split into the two arms of the interferometer by a 50/50 fiber coupler. An acousto-optic modulator is used on one arm to shift the spectrum up in frequency by Ω in order to reject the detected DC signal in the photodiode and to allow the use of a standard RF spectrum analyzer to measure the spectrum of the photocurrent fluctuations. On the other arm a long piece of optical fiber is used to achieve the optical delay that is longer than the coherence length of the laser under test. The optical fields from the two arms are mixed by another 50/50 fiber coupler and the interference signal is detected with a fast photodiode. The laser linewidth is then deduced from the recorded power spectrum of the fluctuations of the photocurrent.

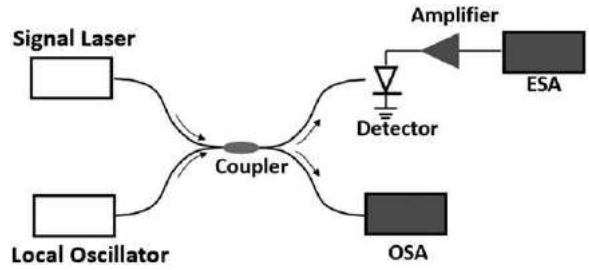

Fig. 13 Configuration of the setup for heterodyne detection.

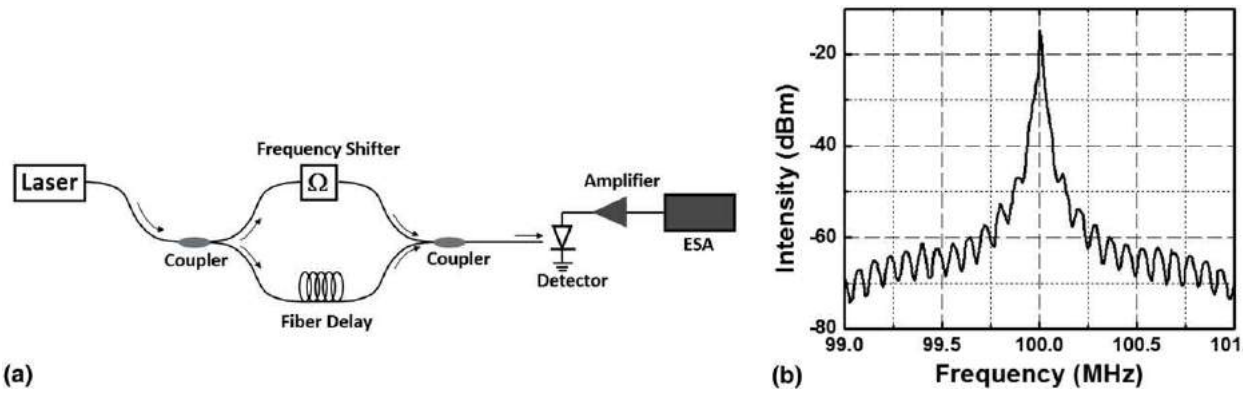

Fig. 14 (a) Configuration of the setup for self-delayed heterodyne detection; (b) typical spectral linewidth measurement result of self-delayed heterodyne detection for a single-frequency DBR fiber laser at 1550 nm.

A typical linewidth measurement result of a single-frequency DBR fiber laser is shown in Fig. 14(b). Because the linewidth of this laser is very narrow and the optical delay fiber is comparable to its coherent length, the measurement result has some ripples there. Generally, an optical delay length of at least 5 times longer than the laser coherence length was suggested for a direct laser linewidth measurement from the spectrum. For very narrow linewidths, however, the required length for the optical delay line may become unpractically long. Therefore self-delayed heterodyne method is only suited for direct measurement of laser linewidths on the order of or above 10 kHz. Nevertheless, the linewidth of a single-frequency laser with a coherence length longer than the optical delay length can be estimated from -20 dB bandwidth of the power spectrum.

Compared with heterodyne detection, the delayed self-heterodyne method doesn't need a local oscillator. Since the local oscillator signal in these measurements is provided by the laser under test, slow drift in wavelength is usually tolerable. Hence delayed self-heterodyne detection is inherently self-calibrated and capable of measuring a large range of laser frequencies where the fiber loss is tolerable.

Applications

Single-frequency lasers have found a variety of applications where long coherence length or narrow spectrum linewidth is essential. Those applications include high resolution spectroscopy, coherent light detection and ranging (LIDAR), remote sensing, laser nonlinear frequency conversions, coherent optical communication, optical data storage, optical cooling and trapping, and microwave photonics.

High resolution spectroscopy highly depends on the linewidth of the laser source. For instance, the absorption linewidth of a gas at standard temperature and pressure is in the range of a few GHz, so the linewidth and stability of the detection laser should be less than 10 MHz for high precision measurements of the absorption lineshape and strength. Single-frequency lasers with inherent narrow linewidth and high intensity stability are excellent laser sources for high resolution spectroscopy.

Single-frequency lasers are in great demand for LIDARs, which can be used to remotely sense properties of targets including the distance, the speed, the particle size and the concentration. Due to the long coherence length of the laser, these properties can be detected by measuring the intensity or frequency shift of returned pulses scattered by the target. LIDARs have been used for scientific meteorological applications and for commercial applications such as aviation safety control and weather forecasting. LIDARs can also be used to remotely monitor the physiological status of vegetation, the decline of forest, and the density of fish population.

Because the beating between longitudinal modes doesn't occur in a single-frequency laser, its output power is extremely stable, which is especially advantageous for nonlinear processes such as harmonic generation or optical parametric oscillator and optical data storage where a low intensity noise is required.

Single-frequency laser sources are also attractive for resonant enhancement cavities for high efficiency nonlinear frequency conversions and for coherent beam combining. Both applications require narrow-linewidth lasers with high stability.

References

Collins, S.A., White, G.R., 1963. Interferometer laser mode selector. Applied Optics 2, 448–449.

Dupuis, R.D., Dapkus, E.P., 1978. Room-temperature operation of distributed-Bragg-confinement $Ga_{1-x}Al_xAs$-Gaas lasers grown by metalorganic chemical vapor deposition. Applies Physics Letters 33, 68–70.

Evtuhov, V., Siegman, A., 1965. A 'twisted-mode' technique for obtaining axially uniform energy density in a laser cavity. Applied Optics 4, 142–143.

Fleming, M., Mooradian, A., 1981. Spectral characteristics of external-cavity controlled semiconductor lasers. IEEE Journal of Quantum Electronics 17, 44–59.

Guy, M.J., Taylor, J.R., Kashyap, R., 1995. Single-frequency erbium fiber ring laser with intracavity phase-shifted fiber Bragg grating narrowband filter. Electronics Letterrs 31, 1924–1925.

Kane, T.J., Byer, R.L., 1985. Monolithic, unidirectional single-mode ring laser. Optics Letters 10, 65–67.

Kogelnik, H., Patel, C.K.N., 1962. Mode suppression and single frequency operation in gaseous optical masers. Proceedings of the IRE 50, 2365–2366.

Kogelnik, H., Shank, C.V., 1972. Coupled-wave theory of distributed feedback lasers. Journal of Applied Physics 43, 2327.

Liu, K., Littman, M.G., 1981. Novel geometry for single-mode scanning of tunable lasers. Optics Letters 6, 117–118.

Maiman, T.H., 1960. Stimulated optical radiation in Ruby. Nature 187, 493–494.

McFarlane, R.A., Bennett, W.R., Lamb, W.E., 1963. Single mode tuning dip in the power output of an He-Ne optical maser. Applied Physics Letters 2, 189–190.

Park, N., Dawson, J.W., Vahala, K.J., 1991. All fiber, low threshold, widely tunable single-frequency, erbium-doped fiber ring laser with a tandem fiber Fabry-Perot filter. Applied Physics Letter 59, 2369–2371.

Polynkin, A., Polynkin, P., Mansuripur, M., Peyghambarian, N., 2005. Single-frequency fiber ring laser with 1 W output power at 1.5 μm. Optics Express 13, 3179–3184.

Schawlow, A.L., Townes, C.H., 1958. Infrared and optical masers. Physical Review 112, 1940–1949.

Schulzgen, A., Li, L., Temyanko, V.L., Suzuki, S., Moloney, J.V., Peyghambarian, N., 2006. Single-frequency fiber oscillator with watt-level output power using photonic crystal phosphate glass fiber. Optics Express 14, 7087–7092.

Tang, C.L., Statz, H., deMars, C., 1963. Spectral output and spiking behavior of solid-state lasers. Journal of Applied Physics 34, 2289–2295.

Yeh, C.H., Huang, T.T., Chien, H.C., Ko, C.H., Chi, S., 2006. Tunable S-band erbium-doped triple-ring laser with single-longitudinal-mode operation. Optics Express 15, 382–386.

Yeh, C., Chi, S., 2005. A broadband fiber ring laser technique with stable and tunable signal-frequency operation. Optics Express 13, 5240–5244.

Zayhowski, J.J., Mooradian, A., 1989. Single-frequency microchip Nd laser. Optics Letters 14, 24–26.

Semiconductor Lasers

Stephan W Koch, Philipps University of Marburg, Marburg, Germany
Martin R Hofmann, Ruhr University Bochum, Bochum, Germany

Introduction

The optical emission from semiconductor lasers arises from the radiative recombination of charge carrier pairs, i.e., electrons and holes in the active area of the device. The conduction-band electron fills the valence-band hole by simultaneously transferring the energy difference to the light field, so-called radiative recombination. Hence, the frequency of the laser light is mainly determined by the bandgap of the semiconductor laser material. In order to achieve lasing, one needs carrier inversion, i.e., sufficiently many electrons in the conduction band such that the light absorption probability is more than compensated by the probability of emission.

A large variety of semiconductor materials has been shown to be suited for semiconductor lasers. The most stringent requirement for the material is that it has a direct bandgap. With other words, the maximum of the valence band and the minimum of the conduction band have to be at the same position in k-space (in the center of the Brillouin zone). This condition excludes the elemental semiconductors Silicon and Germanium from being used for semiconductor lasers.

In order to achieve carrier inversion, it is necessary to pump the semiconductor laser, i.e., to excite sufficiently many electrons from the valence band into the conduction band. Even though this pumping in semiconductors, as in other solid state laser materials, can be done optically, the large technological impact of semiconductor lasers is at least in part due to the possibility of electrical pumping with a few tens of milli-Amperes at voltages of a few Volts.

The original version of such a semiconductor laser device is shown in Fig. 1. The laser structure consists of a GaAs-based p–n junction which is biased in forward direction. Electrons are injected from the n-region and holes from the p-region, respectively. Due to the diode characteristics of the p–n junction, semiconductor lasers are often also called laser diodes. In the device shown in Fig. 1 the laser mirrors are formed by the cleaved facets of the semiconductor crystal.

Because of the high refractive index of GaAs, the semiconductor–air interface provides a reflectivity of about 32%, which is sufficient for laser emission. However, the efficiency of this early type of lasers, the so-called homojunction device, was very poor because of a low carrier density in the active area and because of a weak overlap between the inverted region and the optical mode. Considerable improvement toward room-temperature continuous wave (CW) operation was achieved by the introduction of the so-called double heterostructure. Its principle is shown in Fig. 2.

The basic idea behind the double-heterostructure laser is that the active material (e.g., GaAs) is embedded in another semiconductor with a slightly larger bandgap (e.g., (AlGa)As). This causes a potential well in which electrons and holes are confined in order to achieve high carrier densities in the active area. Moreover, the smaller refractive index of the surrounding high-bandgap material helps to guide the optical field within the region of highest inversion. Very often, additional ridge-shaped lateral structuring is used to form a complete optical waveguide.

Even more advanced laser structures can be fabricated since modern crystal growth techniques, such as molecular beam epitaxy (MBE) and metal organic chemical vapor deposition (MOCVD), became available. These techniques allow the grower not only to determine the composition of the semiconductors with remarkable precision, but also to define the shape virtually on an atomic scale. In particular, it is now possible to grow microstructures so small that their electronic and optical properties deviate substantially from those of the corresponding bulk materials.

So-called quantum-well (QW) structures, in particular, turned out to be superior to bulk materials in many aspects and are currently the active medium in most commercial semiconductor lasers. The gain medium in QW laser structures consists of one or more films with a thickness of several atomic monolayers. These are embedded in barrier material with a larger bandgap that serves to confine the carriers into the wells. QWs are examples of structures where quantum confinement restricts the free carrier motion to the two dimensions within the plane of the films since the well thickness is comparable to or less than the thermal wavelength of the carriers. Hence, one speaks of quasi-two-dimensional (2D) structures where no free motion perpendicular to the films is

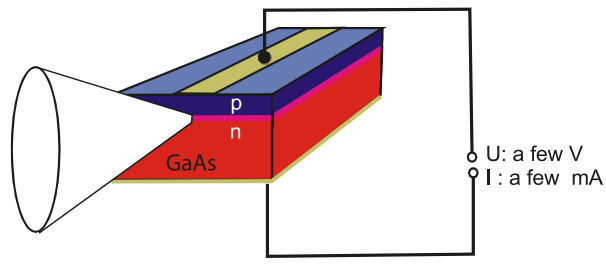

Fig. 1 Operation principle of semiconductor laser.

 doi:10.1016/B978-0-12-803581-8.10103-1

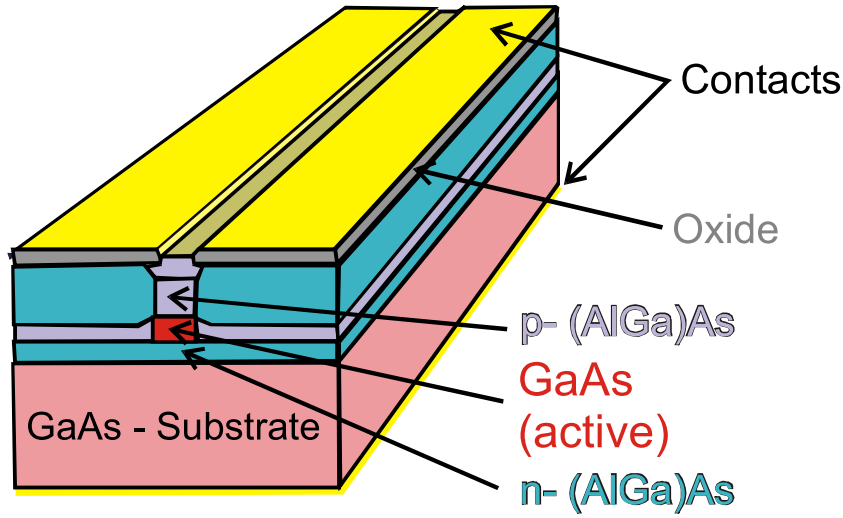

Fig. 2 Double-heterostructure laser diode.

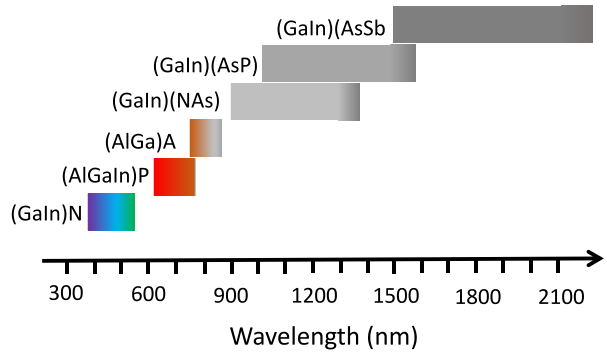

Fig. 3 Typical diode laser materials and their emission wavelength ranges.

possible. The corresponding energy states of the carriers are quantized, i.e., only discrete values are allowed for that part of the carrier kinetic energy which is related to the confined direction (Haug and Koch, 2009).

Such quantum confinement effects are not restricted to one dimension as in QWs, but can also be obtained in two and three dimensions, such as in quantum wires and quantum dots, respectively. Even though wire and dot structures can be grown with various degrees of precision, their use as semiconductor laser gain medium is still an emerging field. However, especially quantum dot structures hold great promise for the future, since they can be grown in large numbers using comparable techniques already used for QWs (Bimberg *et al.*, 1998). The possibility of layer growth under strained conditions, such as QWs, between a barrier material with a slightly different atomic lattice constant, also enriches the flexibility to design the laser gain medium with the desired emission wavelength.

In principle, the emission wavelength of a semiconductor laser can be roughly designed by the choice of the material. Different semiconductors have different gaps between valence band and conduction band and the bandgap energy determines the spectral range of the emission. By fabricating semiconductor alloys as, for example, (GaIn)As or (AlGa)As, one can tune the bandgap and thus the spectral range of the emission. Presently, a large variety of different materials is available, for example, covering the red ((AlGaIn)P), the near infrared ((AlGa)As, (GaIn)As), the telecom range ((GaIn)(AsP)), infrared (GaIn)(AsSb), and even the blue-ultraviolet range ((GaIn)N). In recent years, the progress in material growth has even enabled the realization of green laser diodes based on the (GaIn)N material system. **Fig. 3** gives an overview of the most common commercial diode laser materials and their emission wavelengths.

Great care has to be taken that the crystal used for a semiconductor laser has almost perfect quality. In detail, dislocations in the crystal have to be avoided as well as any kinds of defects since they act as nonradiative recombination centers. Nonradiative recombination competes with the radiative recombination that is essential for the laser process. Nonradiative recombination effects include Auger-recombination and recombination via defects. Auger-recombination is the process where an electron and a hole recombine and completely transfer the energy difference to another carrier (electron or hole), which is excited to a higher state. This is an intrinsic effect that depends mainly on the band structure and the relevant values of the transition probabilities

(transition matrix elements). The probability for Auger-recombination is temperature and carrier density dependent and can be minimized, for example, by properly cooling the device and by keeping the carrier densities as low as possible.

Nonradiative recombination via defects is an extrinsic effect that depends on the material quality and can thus be minimized by optimized growth. The goal to avoid dislocations, however, introduces severe limitations for the choice of materials for laser structures. The growth processes require "host" crystals, so-called substrates, on which the real structure can be grown. As standard substrates for optoelectronic applications, GaAs, InP, SiC, and sapphire wavers are available in sufficient quality. The layers grown by epitaxy assume the crystal structure of the substrate which implies that the lattice constants of substrate and of the layers grown on top should be very similar. If the lattice constants do not agree, i.e., if there is structural mismatch, a certain amount of strain develops that increases with increasing lattice mismatch. Too much strain leads to the formation of dislocations, which are detrimental for laser applications. The control of the dislocation density was, for example, one of the critical issues for the wide-gap III–V materials, such as GaN, because no substrate was available that matched the GaN lattice constant. A small value of strain, however, is often even desirable as long as it does not cause dislocations. Strain influences the band structure of the material in a predictable way and can therefore be used to tailor the laser emission properties.

Searching a laser material for a certain desired wavelength range thus implies two important requirements that in some cases may only be simultaneously met with quaternary semiconductor alloys (e.g., (GaIn)(AsP) on InP-substrates): first, the bandgap of the material has to fit the desired photon energy and second, and more severe, a substrate with the right lattice constant has to be found. This second condition excludes many material compositions from laser applications. In many cases, both conditions can. Presently, this problem introduces enormous challenges in the growing area of optoelectronic/electronic integrated circuits. The growing demand on higher data transfer rates in modern communication systems motivates the use of optical rather than electrical connections even on a microchip level. This requires the embedding of laser diodes into electronic circuits, which are based on silicon technology. Since silicon as an indirect semiconductor is not an appropriate material for semiconductor lasers, alternative technologies have to be implemented to include laser diodes into electronic circuits. This may either be realized with large technological effort by fusion of electronic and optoelectronic wafers or by developing new laser materials that have a lattice constant compatible with that of silicon as, for example, Ga(NAsP).

Physics of the Gain Medium

From a material physics point of view, the important structural aspects of a certain laser material are summarized in the band structure of the gain medium, i.e., in the quantum mechanically allowed energy states of the material electrons. The calculation of the band structure of a particular gain medium is therefore a crucial aspect of semiconductor laser modeling (Chuang, 1995; Chow and Koch, 1999). Additionally, however, one has to understand the relevant properties of the excited electron–hole plasma in the laser's active region.

In its ground state, i.e., without excitation and at very low temperatures, a perfect semiconductor is an insulator where no electrons are in the conduction-band states. The presence of a population inversion, i.e., occupied conduction-band and empty valence-band states in a certain range of frequencies is, however, a requirement for obtaining optical gain and light amplification from a semiconductor. Sufficiently strong pumping leads to the generation of an electron–hole plasma, where the term "electron" specifically refers to conduction-band electrons and "holes" is a short-hand term for the missing electrons in the originally full valence bands. Holes are also quasiparticles, just as the electrons in a solid are quasiparticles, which have an effective mass that is considerably different form the free-electron mass in vacuum and which is determined by the band structure, i.e., the interaction with all the ions of the solid. Thus the properties of the missing electrons, such as a positive charge for the missing negative electron charge, a spin opposite to that of the missing electron, etc., are attributed to the holes. As a consequence of their electrical charges, the excited electrons and holes interact via the Coulomb potential which is repulsive for equal charges (electron–electron, hole–hole) and attractive for opposite charges (electron–hole). Furthermore, electrons and holes obey the Fermi–Dirac statistics and the Pauli exclusion principle not allowing two Fermions to occupy the same quantum state. The consequences of the Pauli principle are often referred to as "band filling effects" in the physics of semiconductor lasers.

The theory therefore has to account not only for the band structure but also for this band filling, i.e., the detailed population of the electron and hole states. Taking into account only band structure and band filling leads to a so-called free-carrier theory. However, the carrier interactions and the exclusion principle render the electron–hole plasma in a semiconductor gain medium an interacting many-body system where the properties of any one carrier are influenced by all the other carriers in their respective quantum states. Generally, advanced many-body techniques are needed to systematically analyze such an interacting electron–hole plasma (Haug and Koch, 2009) and to compute the resulting optical properties, such as gain, absorption or refractive index, all of which are changing in a characteristic way with the changing electron–hole density.

It is often helpful to investigate the characteristic timescales and the most direct consequences of the various interaction effects even though the consequences of the different many-body effects are in general not additive. The fastest interaction under typical laser conditions is the carrier–carrier Coulomb scattering that acts on scales of femto- to picoseconds to establish Fermi–Dirac distributions of the carriers within their bands. These distributions are often referred to as "quasi-equilibrium distributions" since they describe electrons and holes in the conduction and valence bands, which may be characterized by an "electronic temperature" or "plasma temperature." The plasma temperature is a measure for the mean kinetic energy of the respective carrier ensemble and relaxes toward the lattice temperature as a consequence of electron–phonon (=quantized lattice vibration) coupling, i.e., the carriers can emit or absorb phonons. The coupling to longitudinal optical phonons in polar materials is especially strong with

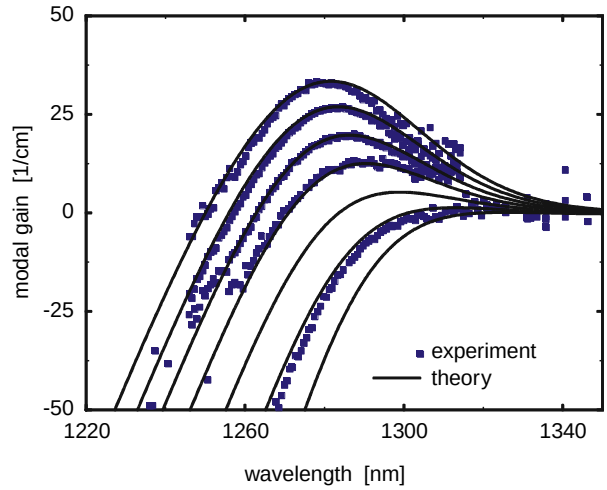

Fig. 4 Comparison of experimental and theoretical gain spectra of a (GaIn)(NAs)/Gaas laser.

scattering times in the picosecond range. The corresponding times for the electron-acoustic phonon scattering are in the nanosecond range. Generally, the carrier–phonon interaction provides an exchange of kinetic energy of the carriers with the crystal lattice in addition to the relaxation of the initial nonequilibrium distribution. Deviations between plasma and lattice temperature may occur, especially when lasers are subject to strong pumping or rapid dynamic changes.

In addition to leading to rapid carrier scattering the Coulomb interaction directly influences the laser gain spectrum. Gain, i.e., negative absorption, occurs in the spectral region where we have population inversion, in other words, where the net probability for a test photon to be absorbed is lower than the probability to be amplified by stimulated electron–hole recombination. Energetically, the gain region is bounded from below by the effective (or renormalized) semiconductor bandgap and from above by the electron–hole chemical potential. Here, the chemical potential is the energy at which neither absorption nor amplification occurs, i.e., where the semiconductor medium is effectively transparent.

The effective bandgap energy for elevated carrier densities is significantly below the fundamental absorption edge of an unexcited semiconductor. This "bandgap renormalization" is a consequence of the fact that the Pauli exclusion principle reduces the repulsive Coulomb interaction between carriers of equal charge. Furthermore, the Coulomb screening, i.e., the density dependent weakening of the effective Coulomb interaction potential contributes to the reduction of the bandgap.

Another many-body effect originates in the Coulomb attraction between an electron and a hole, and leads to an excitonic or Coulomb enhancement of the interband transition probability. All these effects contribute to the detailed characteristics of the optical spectra of a semicondcutor gain medium. Fig. 4 shows an example of calculated spectra.

The example demonstrates that the microscopic theory correctly accounts for the changes in the gain spectrum with changing carrier density by means of a consistent treatment of band structure, band filling, and many-body Coulomb effects.

The Resonator

Besides a gain medium, a laser needs feedback for the emitted photons in order to allow for the positive feedback process required for amplified-stimulated emission. Hence, the inverted medium has to be embedded into a resonator. One of the conceptually simplest versions is a so-called Fabry–Perot resonator, which consists of two parallel plane mirrors. A double-heterostructure Fabry–Perot laser is shown in Fig. 2. For semiconductor lasers, such Fabry–Perot resonators, are very easy to fabricate: by cleaving the semiconductor crystal perpendicular to the optical waveguide. The cleaved edges provide a reflectivity of about 32% due to the high refractive index of the semiconductor. This is sufficient to provide enough feedback for laser operation.

However, despite its conceptual simplicity, the cleaved edge Fabry–Perot concept also introduces a number of problems. First, the light output occurs on two sides of the laser, whereas for most applications all the output power is desired to be concentrated from one facet only. For practical applications, this problem can be solved by depositing coatings onto the cleaved facets of the structure. For example, deposition of a 10% reflectivity coating on one facet and of a 90% reflectivity coating on the other facet concentrates the laser emission almost completely to the facet of lower reflectivity. Antireflection coatings with reflectivities below 10^{-4} are used when laser diodes are coupled to external cavities, for example, in tunable lasers.

A second problem with Fabry–Perot lasers is that the emission wavelength is often not well defined. A Fabry–Perot resonator of length L (optical length, nL) has transmission maxima – so-called modes – at all multiples of $c/2\,nL$, where c is the speed of light. That means that for typical devices with a length of 500 μm, hundreds of modes are present even within the limited gain bandwidth of a semiconductor. These modes often compete with each other resulting in multimode emission, which may become particularly complex and uncontrollable when the laser is modulated for high-speed operation.

This problem can be overcome in so-called distributed feedback (DFB) lasers. In these devices, the refractive index is periodically modulated along the waveguide. In a simplified picture, this periodic pattern causes multiple reflections at different locations within the waveguide. For certain wavelengths, these multiple reflections interfere constructively to provide enough "distributed" feedback for laser operation. With the DFB concept single-mode emission can be realized even for high-speed modulation of the lasers. Alternatively, the region of DFB can be separated from the region of amplification in so-called distributed Bragg reflector (DBR) lasers. These are devices with multiple sections where one section is responsible for the optical gain and another section (the DBR-section) is responsible for the optical feedback.

So far, we have discussed so-called edge-emitting lasers, which means that the light emission is in the plane of the active area of the device. The lengths of these lasers are typically hundreds of microns and thus correspond to a few hundreds optical wavelengths. This causes one of the major disadvantages of edge emitters the multimode emission. Another disadvantage is the poor elliptic beam profile due to diffraction at the small rectangular output aperture. Finally, the fabrication procedure of edge emitters is complex since it requires multiple steps of cleaving before it is even possible to test the laser.

An alternative semiconductor laser concept was realized in the early 1990s: the vertical-cavity surface-emitting laser (VCSEL). A typical VCSEL structure is shown in **Fig. 5**. In VCSELs, the light emission is parallel to the growth direction, i.e., perpendicular to the epitaxial layers of the structure. The cavity length of a VCSEL is in the order of a few (1–5) multiples of half the wavelength of the emitted light. These small dimensions imply that the interaction length between active medium and light field in the resonator is very short. Therefore, the mirrors of a VCSEL have to be of very good quality: reflectivities of more than 99% are required in most configurations.

Very high reflectivities can be achieved with dielectric multilayer structures, so-called Bragg reflectors. Bragg reflectors consist of a series of pairs of layers of different refractive index, where each layer has the thickness of a quarter of the emission wavelength. Most preferably, the Bragg reflectors are epitaxially grown together with the active area of the device in only one growth process.

Further technological challenges arise from the need of electrical pumping of the VCSEL. Thus a p–n junction has to be implemented and sufficient strategies to effectively inject charge carriers into the active region have to be developed. Consequently practical electrically pumped VCSEL structures have an even more complex architecture than the schematic structure shown in **Fig. 5** (Michalzik, 2013).

The small VCSEL resonators with Bragg reflectors at both ends of the cavity are called microcavities. They are designed such that only one longitudinal mode is present in the region of the semiconductor gain spectrum. Accordingly, VCSELs are well suited to provide single-mode characteristics. Moreover, the aperture for light emission can be chosen round with a diameter of a few micrometers so that diffraction is weak and the beam profile is almost perfect.

VCSELs can be tested on wafer without further processing and can be arranged in 2D arrays. A major disadvantage of VCSELs is their temperature dependence. The bandgap of the active material and the cavity mode vary differently with temperature so that the cavity mode shifts away from the region of maximum gain when the temperature is varied. This severely limits the temperature range of operation for VCSELs.

In addition to VCSELs, current semiconductor laser development also focusses on *vertical external cavity surface emitting lasers* (VECSELs) also known as *semiconductor disk lasers*, schematically shown in **Fig. 6**. Here, the top Bragg mirror of the VECSEL is replaced by an external mirror (see **Fig. 5**) in order to obtain an extended external cavity. This novel class of microlasers is particularly attractive for several reasons. First of all, a wide variety of semiconductor materials can be used as the gain medium allowing for the realization of a broad range of emission wavelengths. Moreover, the external cavity allows for the inclusion of nonlinear elements. For example, by inserting a properly designed nonlinear crystal, one can reach new frequencies via intracavity frequency doubling (Kuznetsov *et al.*, 1999; Keller and Tropper, 2006).

As in VECSELs, also in VECSELs round-trip gain is provided by relatively few QWs such that the total material gain is rather small in comparison to a typical edge-emitting semiconductor laser. Therefore, the gain medium, i.e., the active mirror in **Fig. 6** and the lasing mode have to be optimally aligned under the desired VECSEL operational conditions.

One of the intriguing challenges in the VECSEL design is that one typically does not know a priori what the internal operation conditions will be. In particular, the effective temperature of the lasing medium is unknown since the QWs suffer from significant

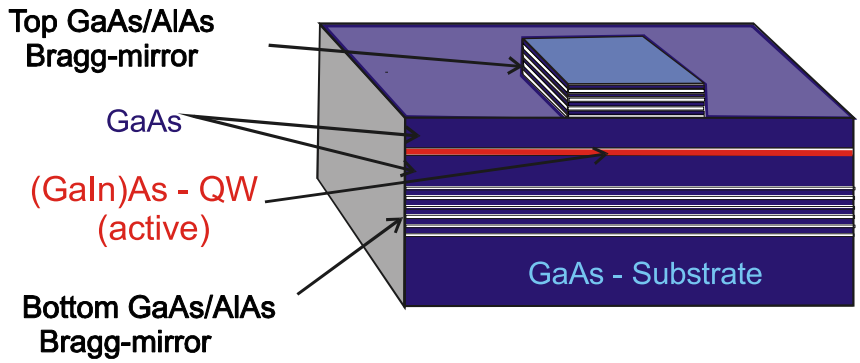

Fig. 5 Vertical-cavity surface-emitting laser (VCSEL).

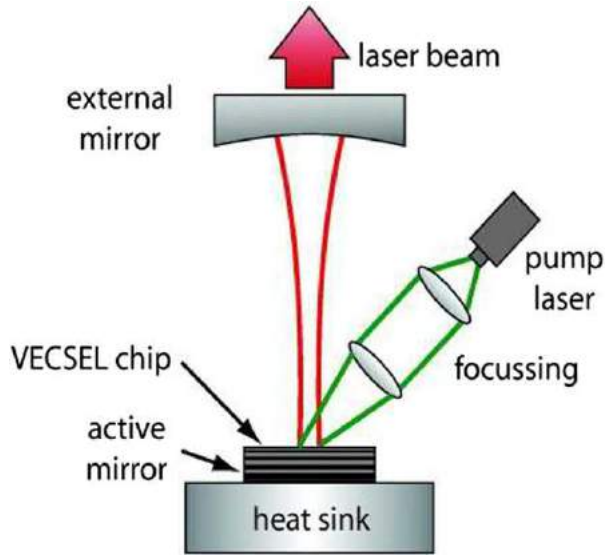

Fig. 6 Vertical external cavity surface emitting lasers (VECSELs) configuration showing the VECSEL chip (active mirror) consisting of a multiple quantum-well (QW) structure grown on a distributed Bragg reflector (DBR). The structure is excited with a pump laser which is focused on the QWs. Lasing in the VECSEL system is achieved by using an external high reflectivity mirror (output coupler) to form a closed cavity. The generated heat is removed by a heat sink typically consisting of a diamond heat-spreader on a copper block.

heating effects. These thermal problems result both from the fundamental impossibility to fully convert the optical pump power into lasing output and from the nonradiative carrier recombination processes. In particular, the intrinsic and thus unavoidable Auger losses lead to a significant QW heating. As countermeasure, one includes heat dissipating elements in the VECSEL structure. However, as we can see in **Fig. 6**, this heat sink typically provides only remote cooling of the QWs. Thus the effective gain medium temperature is determined by the detailed balance between heat generation and dissipation. To make matters worse, this temperature will change with the operating conditions.

To account for all these effects in the VECSEL design, detailed microscopic modeling is needed. A combined effort of modeling and dedicated growth of VECSEL structures has recently lead to the generation of record output powers in the regime of more than 100 W CW emission (Wang *et al.*, 2012).

Semiconductor Laser Dynamics

The dynamical behavior of semiconductor lasers is rather complex because of the coupled electron–hole and light dynamics in the active material. Like other laser materials, semiconductor lasers exhibit a typical threshold behavior when the pump intensity is increased. For weak pumping, the emission is low and spectrally broad: the device operates below threshold and emits (amplified) spontaneous emission. At threshold, the optical gain compensates the losses in the resonator (i.e., internal losses and mirror losses) and lasing action starts. In the optimum case the output power increases linearly with the injection current above threshold. The derivative of this curve is called differential quantum efficiency. The differential quantum efficiency can reach values close to unity and even the overall efficiency of the diode laser, i.e., the conversion rate from electrical power into optical power can be as high as 60%. Ideally, the laser emission is single mode above threshold. However, as discussed above, edge-emitting diode lasers tend to emit on multiple longitudinal modes and the competition of these modes can lead to a complex dynamical behavior, which is drastically enhanced as soon as the laser receives small amounts of optical feedback. Semiconductor lasers are extremely sensitive to feedback and tend to show dynamically unstable, often even chaotic behavior under certain feedback conditions.

As other lasers, also semiconductor lasers respond with relaxation oscillations when they are suddenly switched on or disturbed in stationary operation. Relaxation oscillations, i.e., the damped periodic response of the total carrier density and the output intensity as a dynamically coupled system, are important for many aspects of the semiconductor laser dynamics. The relaxation oscillation frequency determines the maximum modulation frequency of diode lasers. This frequency is strongly governed by the differential gain in the active material, i.e., by the variation of the optical gain with carrier density.

An important application of relaxation oscillations is short pulse generation by gain switching of laser diodes. The idea of gain switching is to inject a short and intense current pulse or optical pump pulse into the laser structure. This pulse initiates relaxation oscillations but it is so short that already the second oscillation maximum is not amplified such that only one intense pulse is emitted. The shortest pulse widths achieved with this scheme so far are in the few picoseconds range. For short pulse generation, a further complexity of the diode laser comes into play. The emitted pulses are observed to be strongly chirped, i.e, the center frequency changes during the pulse. The origin of this complex self phase modulation is the strong coupling of real and imaginary

part of the susceptibility in the semiconductor. When the gain (which is basically given by the imaginary part of the susceptibility) is changing during a pulse emission, the refractive index (basically the real part of the susceptibilty) is also changing considerably. This leads to a self phase modulation when a strong optical pulse is emitted. This strong phase-amplitude coupling is also responsible for the high feedback sensitivity of diode lasers, for complex spatiotemporal dynamics (e.g., self focussing, filamentation), and for a considerable broadening of the emission linewidth. The phase-amplitude coupling is often quantified with the so-called linewidth enhancement factor α. This parameter, however, is a function of carrier density, photon wavelength, and temperature (Chow and Koch, 1999; Osinski and Buus, 1987).

Diode lasers and VECSELs are also well suited for the generation or amplification of extremely short pulses with durations far below 1 ps. The concept used for such extremely short pulse generation is called modelocking. The idea of modelocking is to synchronize the longitudinal modes of the laser such that their superposition is a sequence of short pulses. The most successful concept to achieve modelocking of diode lasers is to introduce a saturable absorber into the laser cavity and to additionally modulate the injection current synchronously to the cavity round-trip frequency (Delfyett *et al.*, 1994). A saturable absorber exhibits a nonlinear absorption characteristics: the absorption is high for weak intensities and weak for high intensities so that high peak power pulses are supported. Such absorbers can be integrated as a separate section into multiple-section diode lasers, which have already been successfully used for sub-picosecond pulse generation. Furthermore, in such multiple-section devices electro-absorption modulators, passive waveguide sections, and tunable DBR-segments for wavelength tuning can be integrated, too.

However, when sub-picosecond pulses are generated or amplified in semiconductor laser amplifiers the dynamics of the pulse generation or pulse amplification is strongly governed by complex nonequilibrium carrier dynamics in the semiconductor (Hofmann, 1999). The most prominent effects are spectral hole burning and carrier heating. Spectral hole burning occurs when an intense laser field removes carriers from a certain area in k-space faster than carrier–carrier scattering can compensate for. This leads to nonthermal carrier distributions and a gain suppression in certain energy regimes. As mentioned above, a nonthermal carrier distribution relaxes quickly (100 fs) into a thermal (quasi-equilibrium) distribution when the intense laser field is switched off but this carrier distribution may have a plasma temperature much higher than the lattice temperature. This effect which is referred to as carrier heating also leads to a transient reduction of the gain on a few picoseconds timescale until the carrier distributions have cooled down to the lattice temperature by interactions with phonons. These ultrafast effects are of course particularly important for amplification and generation of short pulses but they generally influence the whole dynamical behavior of diode lasers and VECSELs (Kilen *et al.*, 2016).

Applications

Semiconductor laser diodes are already an essential part of our daily life as they are key components in numerous important applications. To name a few prominent examples, laser diodes act as light sources in glass fiber transmission systems, which are the basis of the internet. They are used to read out information in CD, DVD, and Blue ray DVD systems. Further, well-established applications include bar code scanners, laser pointers, laser printers, or the optical computer mouse. The recent development of high power laser diodes also enables applications in material processing. The high flexibility of the laser diode materials in terms of emission wavelengths makes them also suitable for high-end scientific applications in spectroscopy, for example. The rapid development of laser diodes with new and improved specifications will continuously open further application fields as, for example, compact laser displays with high brilliance making use of the recently established availability of efficient red, blue, and green laser diodes.

References

Bimberg, D., Grundmann, M., Ledentsov, N.N., 1998. Quantum Dot Heterostructures. New York, NY: Wiley.
Chow, W.W., Koch, S.W., 1999. Semiconductor Laser Fundamentals: Physics of the Gain Materials. Berlin: Springer-Verlag.
Chuang, S.L., 1995. Physics of Optoelectronic Devices. New York, NY: Wiley.
Delfyett, P.J., Dienes, A., Heritage, J.P., Hong, M.Y., Chang, Y.H., 1994. Femtosecond hybrid mode-locked semiconductor laser and amplifier dynamics. Applied Physics B 58, 183–195.
Haug, H., Koch, S.W., 2009. Quantum Theory of the Optical and Electronic Properties of Semiconductors, fifth ed. Singapore: World Scientific Publisher.
Hofmann, M., 1999. Gain and emission dynamics of semiconductor lasers. Recent Research Developments in Applied Physics 2, 269–290.
Keller, U., Tropper, A.C., 2006. Passively modelocked surface-emitting semiconductor lasers. Physics Reports 429, 67.
Kilen, I., Koch, S.W., Hader, J., Moloney, J.V., 2016. Fully microscopic modeling of mode locking in microcavity lasers. Journal of the Optical Society of America B 33, 75.
Kuznetsov, M., Hakimi, F., Sprague, S., Moodardian, A., 1999. Design and characteristics of high-power (> 0.5-W CW) diode-pumped vertical-external-cavity surface-emitting semiconductor lasers with circular TEM00 beams. IEEE Journal of Selected Topics in Quantum Electronics 5, 561.
Michalzik, R., 2013. VCSELs Fundamentals, Technology and Applications of Vertical-Cavity Surface-Emitting Lasers. Berlin: Springer-Verlag.
Osinski, M., Buus, J., 1987. Linewidth broadening factor in semiconductor lasers – An overview. IEEE Journal of Selected Topics in Quantum Electronics 23, 9–29.
Wang, T.L., Heinen, B., Hader, J., *et al.*, 2012. Quantum design strategy pushes high-power vertical-external-cavity surface-emitting lasers beyond 100 W. Laser & Photonics Reviews 6, L12–L14.

Further Reading

Koch, S.W., Jahnke, F., Chow, W.W., 1995. Physics of semiconductor microcavity lasers, review article. Semiconductor Science and Technology 10, 739–751.