TERAHERTZ SENSING TECHNOLOGY
SELECTED TOPICS IN ELECTRONICS AND SYSTEMS

Editor-in-Chief: M. S. Shur

Published

Vol. 12: Current Research on Optical Materials, Devices and Systems in Taiwan
eds. S. Chi and T. P. Lee

Vol. 13: High Speed Circuits for Lightwave Communications
ed. K.-C. Wang

Vol. 14: Quantum-Based Electronics and Devices
eds. M. Dutta and M. A. Stroscio

Vol. 15: Silicon and Beyond
eds. M. S. Shur and T. A. Fjeldly

Vol. 16: Advances in Semiconductor Lasers and Applications to Optoelectronics
eds. M. Dutta and M. A. Stroscio

Vol. 17: Frontiers in Electronics: From Materials to Systems
eds. Y. S. Park, S. Luryi, M. S. Shur, J. M. Xu and A. Zaslavsky

Vol. 18: Sensitive Skin
eds. V. Lumelsky, M. S. Shur and S. Wagner

Vol. 19: Advances in Surface Acoustic Wave Technology, Systems and Applications (Two volumes), volume 1
eds. C. C. W. Ruppel and T. A. Fjeldly

Vol. 20: Advances in Surface Acoustic Wave Technology, Systems and Applications (Two volumes), volume 2
eds. C. C. W. Ruppel and T. A. Fjeldly

Vol. 21: High Speed Integrated Circuit Technology, Towards 100 GHz Logic
ed. M. Rodwell

Vol. 22: Topics in High Field Transport in Semiconductors
eds. K. F. Brennan and P. P. Ruden

Vol. 23: Oxide Reliability: A Summary of Silicon Oxide Wearout, Breakdown, and Reliability
ed. D. J. Dumin

Vol. 24: CMOS RF Modeling, Characterization and Applications
eds. M. J. Deen and T. A. Fjeldly

Vol. 25: Quantum Dots
eds. E. Borovitskaya and M. S. Shur

Vol. 26: Frontiers in Electronics: Future Chips
eds. Y. S. Park, M. S. Shur and W. Tang

Vol. 27: Intersubband Infrared Photodetectors
ed. V. Ryzhii

Vol. 28: Advanced Semiconductor Heterostructures: Novel Devices, Potential Device Applications and Basic Properties
eds. M. Dutta and M. A. Stroscio

Vol. 29: Compound Semiconductor Integrated Circuits
ed. Tho T. Vu
TERAHERTZ SENSING TECHNOLOGY

Volume 1: Electronic Devices and Advanced Systems Technology

Editors

Dwight L Woolard
US Army Research Laboratory

William R Loerop
US Army Soldier Biological and Chemical Command

Michael S Shur
Rensselaer Polytechnic Institute

World Scientific
New Jersey • London • Singapore • Hong Kong
Preface

As we begin the new millennium, significant scientific and technical challenges remain within the terahertz (THz) frequency regime and they have recently motivated an array of new research activities. Indeed, the last research frontier in high-frequency electronics now lies in the so-called terahertz (or submillimeter-wave) regime between the traditional microwave and the infrared domains. Today, the terahertz or THz regime has been broadly defined as the portion of the submillimeter-wavelength electromagnetic (EM) spectrum between approximately 1 mm (300 GHz) and 100 μm (3 THz). While the THz frequency regime has long offered many important technical advantages (e.g., wider bandwidth, improved spatial resolution, component compactness) and while significant scientific interest in THz frequency science and technology has existed since the early 1900's, the solid-state electronics capability at THz frequencies today remains extremely limited from a basic signal source and systems perspective (i.e., the best output power is less than or on the order of milliwatts). The relatively limited development of semiconductor-based electronic circuits within the THz band may be attributed to the confluence of two fundamental factors. First, very challenging development and engineering problems are present in this quasi-optical regime where EM wavelength is on the order of component size. Second, the practical and scientific applications of this shorter-wavelength microwave region, where the atmospheric propagation paths are extremely attenuating, have been restricted to a few specialized fields (e.g., molecular spectroscopy for Earth, planetary and space science). Furthermore, engineering efforts to extend conventional three-terminal semiconductor devices upward from millimeter-wave as well as separate efforts to extrapolate traditional solid-state laser technology down from the far infrared have been prohibited due to fundamental physical factors associated with the respective device technologies. Indeed, electron velocities are not high enough to extend the operational of conventional transit-mode transistors (e.g., HFET's and HBT's) into the THz range even for devices with deep submicron dimensions. Alternatively, when the laser end of the spectrum is considered, one finds that the energies of photonic transitions corresponding to the terahertz range are small compared to the thermal energies at room or elevated temperatures (one terahertz corresponds to a 4.14 meV photon energy and to approximately 300 micron wavelength in free space) and effects, such as free carrier generation, plague these devices.

Fortunately, two-terminal semiconductor devices (e.g., Schottky and Heterostructure Barrier Varactor and Schottky mixers) utilize charging effects very near the contact interface and are inherently faster than transistor devices. Hence, two-terminal transport-based semiconductor devices long ago emerged as the key technology for the generation, amplification and detection of electrical signals at submillimeter-wave frequencies. However, the overall performance (e.g., power and efficiency) of even state-of-the-art two-terminal technologies suffers as they are extended for operation high into the THz band. This longstanding limitation in THz electronic technology, along with the excessive cost of instrumentation, has certainly been a major stumbling block to new scientific inquiries at THz frequencies and has most probably prevented the spread of science and technology related issues to the broader scientific and engineering communities. Recent advances in nanotechnology, molecular chemistry and biological science have already begun to chart the course for new and important applications of THz
electronics in the coming twenty-first century. In fact, the growing interest in the precise detection, identification and characterization of very small organic and inorganic systems has already begun to emphasize the future value of a robust THz-frequency sensing science and to establish it as an important driver for the rapid advancement of electronics technology at THz frequency. These new and exciting sensing applications only provide added motivation for realizing a practically useful THz electronics technology, which has long been recognized to offer much to conventional electronic application areas such as extended bandwidth for special scenario communications (i.e., short-range, networked and satellite) and significantly enhanced signal processing power.

During the last few years, major research programs have emerged within the U.S. Army and the Department of Defense (DoD) that have been focused on advancing the state-of-the-art in THz-frequency electronic technology and on investigating novel applications of THz-frequency sensing. These basic programs grew out of small seed efforts supported by a number of agencies including the U.S. Army Research Laboratory (ARL), the U.S. Soldier Biological and Chemical Command (SBCCOM) and the Air Force Office of Scientific Research (AFOSR). More recently, these efforts have been intensified and propagated primarily from the support of a Defense Advanced Research Project Agency (DARPA) Program on “Terahertz Technology for Sensing and Satellite Communications” and a Multidisciplinary University Research Initiative (MURI) Program on “Sensing Science and Electronic Technology at THz Frequencies” that is managed out of the Army Research Office (ARO) of ARL. One of the main catalysts for these programs is associated with the idea of using the fundamental interactions of THz radiation at the molecular level for sensing and characterizing chemical and biological (CB) agents. The science and technology emerging from these programs has potentially important ramifications to such areas as CB defense, biomedical and molecular science. Historically, the U.S. DoD has been instrumental in establishing the basic foundations for many important endeavors in science and engineering and recent DoD support for THz electronics is once again playing a major role in promoting an interest among the broader community. Indeed, there has been a steadily growing interest among the international scientific and technical communities in the unique challenges associated with developing a robust electronics technology and with developing a detailed understanding of THz-frequency sensing science.

This growing wave of research and development has already started to shrink the “THz gap” through recent advances in both photonic and electronic technology. On the photonic technology side, the most prominent recent development has been the emergence of relatively high power THz quantum cascade lasers developed by Ruedeger, Köhler and Alessandro Tredicucci of the Scuola Normale Superiore in Pisa and colleagues in Turin and Cambridge 1,2 (see also earlier work of the Swiss group 3). This

---

1 http://physicsweb.org/article/news/6/5/5
2 http://optics.org/articles/news/8/5/16/1
device used 1500 alternating layers of gallium arsenide and aluminum gallium arsenide illustrating dramatic improvements in sophistication of modern semiconductor materials growth technology. This device (1.5 mm long and 0.2 mm wide) emitted 2 mW at 4.4 THz (wavelength of 67 μm) in single mode operation at 50 K. On the electronic technology side, recent progress in planer integration of two-terminal multipliers and in engineering photomixer technology is defining the new state-of-the-art in solid-state electronic sources and new concepts such as ballistic tunneling devices and plasma waves electronics offer promise for considerable potentials improvements in the future.

The recent plethora of new research activities associated with THz-frequency science and technology has motivated the organization of this book. This book has been structured to stand as an up to date and detailed reference for the new THz-frequency technological advances that are emerging across a wide spectrum of sensing and technology areas. This two-volume book gives a comprehensive review of existing terahertz technology, innovative applications in sensing, and new emerging ideas, whose full significance remains to fully understood and established in many cases.

The first volume of the book focuses primarily on established electronic device concepts and on advanced systems technology. Chapter 1 begins with a general review of THz technology by Peter Siegel. Here, issues ranging from a historical background to applications are discussed and state-of-the-art THz sensors and sources are assessed. His conclusion (albeit quite optimistic) is that the THz technology "enters adulthood." In Chapter 2, Haddad et al address classical subterahertz sources, which are two-terminal semiconductor devices with static or dynamic negative differential resistance (such as Gunn diodes and IMPATT, BARITT, and TUNNETT diodes, respectively). A new concept for a ballistic tunneling transit-time device (BT3D) is also introduced which has the potential for circumventing earlier limitations to two-terminal device operation in the THz regime. In Chapter 3, Weikle et al summarize multiplier and harmonic generator technologies and identify the important challenges that remain. This chapter discusses the importance of advanced fabrication technology and the effective implementation of new circuit concepts in realizing the necessary functionality and performance of THz-frequency sources for future sensing applications. Chapter 4 is authored by Urteaga et al and presents results on the fastest transistors on earth, which are InP-based Heterojunction Bipolar Transistors. This impressive technology has already yielded amplifiers operating in the 140 to 220 GHz range and holds promise for effective operation at frequencies up to 400 GHz. Dr. Brown gives a detailed engineering presentation in Chapter 5 on the interesting and exciting approach of utilizing conductive photomixing for realizing terahertz sources. These devices have already achieved performance close to their theoretical limits. They operate at temperatures close to room temperature, albeit at low power levels. However, future developments involving two-dimensional device arrays offer a methodology for significantly extending the capabilities of this technology. In Chapter 6, Kelsall and Soref explore the prospects of Si-based SiGe/Si quantum cascade lasers in the THz regime. These devices should be able to operate at temperatures up to 77 K and have a number of potential advantages to their GaAs-based counterparts that recently have achieved a remarkable success as mentioned above. Shur and Ryzhii review the concepts of plasma wave electronics in Chapter 7 and evaluate it as a novel approach to the detection and generation of THz-frequency
radiation. Here, surface plasma waves are shown to propagate with a much higher velocity than electrons. This phenomenon is demonstrated as a new method for realizing tunable THz detectors and emitters operating in the entire temperature range from cryogenic temperatures to room temperature and above. The last two chapters in volume I consider terahertz-frequency sensing and imaging applications that have been facilitated through the use of advanced system technology. Here, Mickan and Zhang review the state-of-the-art technology for T-Ray sensing and imaging in Chapter 8 and consider it use in studying gases, liquids and solids. Finally, Dorney et al presents a new time-domain imaging method in Chapter 9 that can be used to emulate data collection in geophysical prospecting. These results are shown to broaden the capabilities of THz imaging systems and may define new uses of the technology.

The editors would like to thank all the authors for their fine contributions to this outstanding review of “Terahertz Sensing Technology”. Both Volume I on *Electronic Devices & Advanced Systems Technology*, and Volume II on *Emerging Scientific Applications & Novel Device Concepts*, will be useful for technologists, scientists, engineers, and graduate students who are interested in the development of terahertz technology for sensing applications. The book can also be used as a textbook for graduate and senior undergraduate courses on terahertz electronics and as an additional reference text for courses in semiconductor physics, materials science, and electronic device design.
CONTENTS

Preface v

THz Technology: An Overview
P. H. Siegel 1

Two-Terminal Active Devices for Terahertz Sources
G. I. Haddad, J. R. East, and H. Eisele 45

Multiplier and Harmonic Generator Technologies for Terahertz Applications
R. M. Weikle, II, T. W. Crowe, and E. L. Kollberg 79

Submicron InP-based HBTs for Ultra-high Frequency Amplifiers
M. Urteaga, S. Krishnan, D. Scott, Y. Wei, M. Dahlstrom, S. Lee, and M. J. W. Rodwell 107

THz Generation by Photomixing in Ultrafast Photoconductors
E. R. Brown 147

Silicon-Germanium Quantum-Cascade Lasers
R. W. Kelsall and R. A. Soref 197

Plasma Wave Electronics
M. S. Shur and V. Ryzhii 225

T-Ray Sensing and Imaging
S. P. Mickan and X.-C. Zhang 251

Multistatic Reflection Imaging with Terahertz Pulses
T. D. Dorney, W. W. Symes, and D. M. Mittleman 327
THz Technology: An Overview

P. H. Siegel

California Institute of Technology, Division of Biology
and the Jet Propulsion Laboratory, Pasadena, CA 91109, USA

THz technology applications, sensors and sources are briefly reviewed. Emphasis is placed on the less familiar components, instruments or subsystems. Science drivers, some historic background and future trends are discussed.

Key Words: THz, submillimeter, technology, applications

1. Introduction

THz technology and applications have long been the province of molecular astronomers and chemical spectroscopists. However, recent advances in THz detectors and sources have started to open the field to new applications. Not surprisingly, there have been many attempts to utilize some of the unique attributes of the submillimeter wavelength bands in applications as diverse as detecting the water content of bulk paper to measuring the radar cross section of enormous ships and aircraft on small scale models. In this chapter we will take a brief look at the infancy of modern THz technology, pass through early childhood and end up at adolescence. The field is perched on adulthood and perhaps in another quarter century a more complete history can be written, hopefully by someone reading this article today!

2. Background

It is interesting to note that the term terahertz did not come into popular use until the mid 1970's [1,2,3,4] where it was employed by spectroscopists to describe emission or absorption frequencies that fell below the far infrared (IR). Before this, THz frequencies were officially designated by the term MMc or mega-megacycles [5]. Today terahertz or THz is broadly applied to submillimeter-wave energy that fills the wavelength range between 1000 and 100 microns (300 GHz to 3 THz). Below 300 GHz we cross into the millimeter-wave bands (best delimited in the author's opinion by the upper operating frequency of WR-3 waveguide — 330 GHz). Beyond 3 THz, and out to 30 microns (10 THz) is more or less unclaimed territory as few if any components exist. The border between far IR and submillimeter is still rather blurry and the designation is likely to follow the methodology (bulk or modal — photon or wave), which is dominant in the particular instrument.

Despite great scientific interest since at least the 1920's [6], the THz frequency range remains one of the least tapped regions of the electromagnetic spectrum. Sandwiched between traditional microwave and optical technologies where there is a limited atmospheric propagation path [7] (Fig. 1), little commercial emphasis has been placed on

1 This text is reprinted in large measure from: Peter H. Siegel, "Terahertz Technology," IEEE Transactions on Microwave Theory and Techniques, March 2002, pp. 910-928, © 2002 IEEE.
THz systems. This has, perhaps fortunately, preserved some unique science and applications for tomorrow’s technologists. For over 25 years the sole niche for THz technology has been in the high resolution spectroscopy and remote sensing areas where heterodyne and Fourier transform techniques have allowed astronomers, chemists, Earth, planetary and space scientists to measure, catalog and map thermal emission lines for a wide variety of lightweight molecules. As it turns out, nowhere else in the electromagnetic spectrum do we receive so much information about these chemical species. In fact, the universe is bathed in THz energy; most of it going unnoticed and undetected.

![ATMOSPHERIC TRANSMISSION](image)

Fig. 1. Atmospheric transmission at various locations and altitudes for given precipitable water vapor pressure (mm) in the THz region. Top: 0-500 GHz, Bottom: 500-2000 GHz. (Data calculated by Eric Grossman using his “Airhead Software”[7]).

This short review chapter will examine THz technology today with emphasis on frequencies above 500 GHz and on applications that may not be familiar to every reader. We will also try to do justice to molecular spectroscopy for Earth, planetary and space science, the chief drivers of THz technology to-date. An excellent overview of lower frequency millimeter and submillimeter-wave technology can still be found in the review papers of Paul Coleman [8,9] and James Wiltse [10]. Commercial uses for THz sensors and sources are just beginning to emerge as the technology enables new instrumentation and measurement systems. So called T-Ray imaging is tantalizing the inter-
ests of the medical community and promises to open the field up to the general public for the first time. Other less pervasive applications have been proposed, all of which would benefit from broader-based interest in the field. We will try to cover some of these and anticipate others in the course of this review.

We begin with a survey of THz Applications (III) and follow this with some selected information on THz Components (IV). More detailed discussions on THz components, materials and techniques cannot be covered. We conclude with some fanciful applications and predictions (V).

3. THz Applications

The wavelength range from 1mm to 100 µm corresponds to an approximate photon energy between 1.2 and 12.4 meV or to an equivalent black body temperature between 14 and 140 Kelvin, well below the ambient background on Earth. A quick look at the spectral signature of an interstellar dust cloud (Fig. 2) however, explains why astronomers are so interested in THz sensor technology. An excellent science review can be found in Phillips and Keene [11]. Fig. 2 (from that paper) shows the radiated power vs. wavelength for interstellar dust, light and heavy molecules, a 30K blackbody radiation curve and the 2.7K cosmic background signature.

Fig. 2. Radiated energy vs. wavelength showing 30K blackbody, typical interstellar dust and key molecular line emissions in the submillimeter (Reprinted from [11]).
Besides the continuum, interstellar dust clouds likely emit some forty thousand individual spectral lines, only a few thousand of which have been resolved and many of these have not been identified. Much of the THz bands have yet to be mapped with sufficient resolution to avoid signal masking from spectral line clutter or obscuration from atmospheric absorption. Results from the NASA Cosmic Background Explorer (COBE) Diffuse Infrared Background Experiment (DIRBE) and examination of the spectral energy distributions in observable galaxies, indicate that approximately half of the total luminosity and 98% of the photons emitted since the Big Bang fall into the submillimeter and far infrared [12]. Much of this energy is being radiated by cool interstellar dust. Older galaxies, like our Milky Way, have a much greater abundance of dust (Figs. 3 and 4) [13], making submillimeter detectors true probes into the early universe.


Fig. 3. Energy output vs. wavelength for galaxies of ascending ages (E-galaxy=youngest, Arp220=oldest) showing the advantages of THz detection systems for probing the early universe (Reprinted with permission from W. Langer [13]).
In addition, red shifted spectral lines from the early universe appear strongly in the far-IR where they are less obscured by intervening dust, that often hides our view of galactic centers. Individual emission lines such as C$^+$ at 158 microns (1.9 THz), the brightest line in the Milky Way submillimeter-wave spectrum, provide a detailed look at star forming regions where surrounding dust is illuminated by hot young ultraviolet emitting stars. Many other abundant molecules: water, oxygen, carbon monoxide, nitrogen, to name a few, can be probed in the THz regime. Since these signals are obscured from most Earth based observations (except from a very few high altitude observatories, aircraft or balloon platforms), they provide strong motivation for a number of existing or upcoming space astrophysics instruments, most notably SWAS [14] (Submillimeter Wave Astronomy Satellite), launched in December 1998 and currently sending back data on water, oxygen, neutral carbon and carbon monoxide in interstellar space; and in the near future the European Space Agency’s Herschel [15] (formerly FIRST – Far Infrared and Submillimeter space Telescope) scheduled for 2007 and NASA’s proposed SPECS (Submillimeter Probe of the Evolution of Cosmic Structure) [16], SPIRIT [12] (SPace InfraRed Interferometric Telescope) and FAIR (Filled Aperture InfraRed telescope) [17], which will examine this spectral region in great detail in the decade beyond 2010. For interstellar and intragalactic observations both high resolving power (large apertures) and high spectral resolution (1-100 MHz) are generally required. In the lower THz bands heterodyne detectors are generally preferred (although this is very application dependent). For the shorter wavelengths direct detectors offer significant sensitivity advantages. Probing inside star systems or galaxies requires extremely high angular resolution, obtainable only with untenably large diameter telescopes or from phase co-
herent interferometric techniques such as planned for SPIRIT and SPECS. In an apt comparison from [16], even a large submillimeter telescope like the James Clerk Maxwell 15m diameter telescope on Mauna Kea operating at 300 GHz has an angular resolution equivalent only to the human eye at 5000 Å. A ground-based (mountain top) interferometer ALMA (Atacama Large Millimeter Array) [18] based in Chile, which may have a baseline of 10 km or more and angular resolution better than 0.01 arc seconds is now being planned by NRAO (National Radio Astronomy Observatory) and several international partners and may well contain heterodyne spectrometers at frequencies as high as 1500 GHz [19].

Many of the same spectral signatures that are so abundant in interstellar and intragalactic space are also present in planetary atmospheres where background temperatures range from tens of Kelvin to several hundred K. Particularly important are thermal emission lines from gases that appear in the Earth’s stratosphere and upper troposphere; water, oxygen, chlorine and nitrogen compounds, etc. that serve as pointers to the abundances, distributions and reaction rates of species involved in ozone destruction, global warming, total radiation balance and pollution monitoring. Many key species either have thermal emission line peaks or their first rotational or vibrational line emissions in the submillimeter, especially between 300 and 2500 GHz (Fig. 5) [20,21].

![Fig. 5. Spectra of some important molecules in the Earth’s upper atmosphere and measurements being addressed by NASA heterodyne instruments. The peak power or minimum frequency for many emission lines occurs in the THz region. Figure based on J. Waters, EOS Atmospheres presentation [21].](image-url)
Again, these emission lines are best observed from platforms above the Earth's atmosphere. Several recent space instruments in particular, Earth Observing System Microwave Limb Sounder (EOS-MLS) launching on Aura in 2003 with high resolution heterodyne receivers at 118, 190, 240, 640 and 2520 GHz [22], have been designed to take advantage of the information content available through high resolution spectroscopic measurements of these gases at submillimeter-wave frequencies. Unlike the astrophysical sources, even modest diameter collecting surfaces are fully filled by the signal beam in atmospheric observations. Resolution requirements are set by the orbital path and speed or by the atmospheric processes themselves. In both limb sounding (scanning through the atmospheric limb along the tangent line) or nadir sounding (looking straight down through the atmosphere) precise spectral line shape information is required to separate out the effects of pressure and Doppler broadening at each altitude along the emission path. Spectral resolution of better than one part in a million is typically needed for line widths that range from tens of kHz in the upper stratosphere to 10 MHz or more lower down. In the lower stratosphere water and oxygen absorption makes the atmosphere optically thick in the THz bands and longer millimeter wavelengths must be used for chemical probing. Following a very successful deployment in September 1991 and many years of continuous data collection from UARS Microwave Limb Sounder [23], which carried heterodyne limb scanning spectrometers at 63, 183 and 205 GHz, several additional Earth sounders are planned for millimeter and submillimeter-wave observations over the next decade. Besides NASA's EOS-MLS, these include Sweden's Odin (astrophysics and Earth sounding at 118, 490 and 557 GHz) launched in February 2001 [24], the Japanese SMILES, Superconducting sub-Millimeter wave Limb Emission Sounder on the Japanese Experimental Module of the International Space Station (JEMS) planned for 2005 with a superconducting receiver at 640 GHz [25,26]; and still in study phase, the European Space Agency's SOPRANO (Submillimeter Observation of PROCesses in the Atmosphere Noteworthy for Ozone) [27] and NASA's follow on to EOS-MLS, Advanced Microwave Limb Sounder (AMLS) [28].

A last major space application for THz sensors is in planetary and small body (asteroids, moons and comets) observations. Understanding the atmospheric dynamics and composition of these Earth companion bodies allows us to refine models of our own atmosphere as well as gaining insight into the formation and evolution of the solar system [29]. Surface-based (landers) or orbital remote sensing observations of gaseous species in the Venutian, Martian and Jovian atmospheres as well as around Europa and Titan have all been proposed [30]. A soon to be launched (2003) ESA cornerstone observatory mission, Rosetta [31], contains a submillimeter-wave radiometer on MIRO (Microwave Imager on the Rosetta Orbiter) at 562 GHz which will look at water vapor and carbon monoxide in the head and tail of comet Wirtanen [32]. As with the Earth sounders, short wavelengths allow for small antennas (and therefore smaller instruments) and still provide adequate spatial resolution for many atmospheric processes. The products of high resolution submillimeter-wave remote sensing, such as composition, temperature, pressure, and gas velocity (winds) offer the planetologist a wealth of information on a global scale. It is not unreasonable to suppose that the first detection of planets containing atmospheric conditions (temperature, pressure, composition) suitable for extraterrestrial life forms will be confirmed by THz spectroscopy. Such a discovery would surely justify the technology investment in this field. An excellent, but short, overview of THz applications for space can be found in [33].
Back on Earth, the two most pervasive applications for THz technology have been in the areas of plasma fusion diagnostics and gas spectroscopy. An excellent review of THz techniques in the fusion field with close to 400 references can be found in Luhmann and Peebles [34]. Most of the measurements involve determination of the electron density profile as a function of position and time in the plasma core [35]. Identification of the power spectrum can be through Thomson scattering or detection of synchrotron radiation (spiraling electrons emitted from plasma discharges in a confined magnetic field via electron cyclotron emission, ECE). The temperature of the plasma can be inferred from the equivalent blackbody intensity recorded in a narrowband radiometer pointing along a radial line of sight into the plasma core [36]. Since the magnetic field intensity in a toroidal plasma varies linearly along a radial path, the ECE frequency changes correspondingly ($\omega B$). Using either a scanned LO or a wide IF bandwidth, one can obtain a profile of the temperature distribution along the plasma radius [37]. Another phenomenon associated with fusion plasmas that has a large effect on power balance is electron temperature fluctuations in the core. These appear in the output signal as white noise riding on top of the simple thermal electron noise. However, this additional output noise is correlated with position within the plasma and can therefore be separated out using interferometric techniques [38]. Since this involves a minimum of two radiometers and benefits from many more, it has been a major driver for the development of heterodyne imaging systems at millimeter and submillimeter-wave frequencies [39]. 2D and now 3D snapshot systems are being developed [37]. Typical tokamak systems have fundamental ECE resonances in the millimeter bands (100-150 GHz) however much can be gained from THz measurements of the ECE harmonics and in regions where strong magnetic fields are found. A schematic of an operating 2D imaging array using a 1D line detector arrangement is shown in Fig. 6 [39]. The 2D image is obtained by varying frequency (and hence penetration into the plasma).

Fig. 6. Diagram of 2D plasma ECE heterodyne imager using linear array of receive antennas for ~100 GHz. LO power is incident from the rear of the array and distributed to each element by a cylindrical mirror. Reprinted with permission from [39]. ©2001, American Institute of Physics.
The title of grandfather of THz technology belongs to the molecular spectroscopists, especially early pioneers like W. King, W. Gordy, C.M. Johnson and C. H. Townes, to name a select few. Although many of the measurements were (and still are) performed with broadband Fourier transform spectrometers using thermal sources and bolometric detectors, much of the later heterodyne instrumentation (sources and detectors) as well as modern ultrasensitive direct detector technology owes its origins to this field. The history of submillimeter-wave spectroscopy is well covered in [10] and the general field fills several texts [40] as well as more than 50 years of annual conferences at Ohio State [41]. The draw of submillimeter-wave spectroscopy as opposed to more readily realized microwave spectroscopy, is in the strengths of the emission or absorption lines for the rotational and vibrational excitations of the lighter molecules. Since these lines tend to increase in strength as $\tilde{J}$ or even $\tilde{J}^2$ and often peak in the submillimeter, there is a strong natural sensitivity advantage in working at THz frequencies. Modern applications foreseen for THz spectroscopy (besides categorizing and compiling specific spectral line emissions) involve rapid scan and gas identification systems such as targeted molecule radar's for detecting and identifying noxious plumes [42] or very versatile systems like FASSST (FAst Scan Submillimeter Spectroscopic Technique) developed at Ohio State [43,44] and optical pulse terahertz time domain spectroscopy instruments [45] (the T-Ray imagers described shortly). Such systems could conceivably measure and rapidly identify such diverse spectral signatures as simple thermal absorption from an intervening gas to dangling molecular bonds on the surface of a solid. A schematic and sample plots from FASSST appear in Figs. 7(a) and (b). The resolution of this system is in the 100 kHz range, it can scan and record $\sim$10,000 lines/second and has a bandwidth of over 100 GHz (based around available voltage tunable backward wave oscillators). The identification of signature gases can be done in milliseconds through look up tables once an exact determination of the spectral line frequency has been made. Following on this concept, proposals have even been made for THz detection of DNA signatures through dielectric resonances (phonon absorption) [46,47 and Fig. 8]. It is too early to tell what unique applications such systems will have, but the potential for interesting science as well as deployable instruments is certainly there.

Fig. 7. (a) FASSST schematic.
Fig. 7. (b) FASSST spectral plots at 500 GHz in increasing frequency resolution sweeps (top to bottom) for a combination of pyrrole, pyridine and sulfur dioxide at 10mTorr each. Reprinted with permission from [44].

Fig. 8. Comparison of Herring and Salmon DNA transmission spectra using an FTS system. Reprinted with permission from [47].
Since so little instrumentation is commercially available for THz measurements, and what does exist is generally too costly for any but the most well funded institutions (if any still exist!), other drivers for the technology have been very slow to take hold. Strong 183 and 557 GHz water lines have been proposed for many planetary and space sensor passive emission measurements including potential life detection, but these same spectral lines (or many others) can be used to determine water content of materials through transmission measurements. At least one application for characterizing the water content of newspaper print has been proposed [48] and a patent filed [49]. Another application that was demonstrated and actually made into a commercial system, came out of United Technologies Research Center in the early 1980’s and involved using optically pumped far IR lasers to detect small voids in electric power cable [50]. Mie scattering from a focused far IR laser running methanol at 118 μm was used to detect voids with radii on the order of λ in a polyethylene-covered coax. The prototype seemed to work nicely as shown by the data in Fig. 9, and could detect both the size and position of the voids as well as defects in the inner conductor and particulate scattering. Unfortunately, before the application could take hold, the cable manufacturing process was changed to one in which a semiconducting outer coating was applied to the polyethylene sheath making the cable impenetrable at THz frequencies [51]!

The atmospheric opacity severely limits radar and communications applications at THz frequencies, however some close-in systems have been proposed and studied [52]. Secure communications (through high attenuation outside the targeted receiver area) or secure intersatellite systems, benefit from the small antenna sizes needed to produce

Fig. 9. 2.5 THz cable void inspection system and plots of scattered power from void filled (left) and solid (right) cable. Peaks are indicative of void size. Reprinted with permission from [50].
highly directional beams as well as the large information bandwidth allowed by THz carriers. Operation in the stratosphere (air-to-air links) is particularly advantageous for THz communications or radar systems because of the low scattering compared to IR and optical wavelengths (proportional to $\lambda^2$ rather than $\lambda^1$) and the much greater penetration through aerosols and clouds. As can be seen from Fig. 1, stratospheric windows abound. A novel scheme for markedly increasing the bandwidth, number of discrete channels and even adding track and scan capability to a THz communications system was proposed by Elliott Brown [53], taking advantage of optical keying techniques and THz power generation via photomixer arrays (Fig. 10). Airways data rates in the tens of gigabits per second are possible in such a system. Although concepts for ultrawide bandwidth “pocket” communications transceivers have been floated for years, the problems inherent in producing small and efficient THz transmitters or local oscillator sources to drive heterodyne systems have so far precluded any commercial development in this area, however new photoconductor components may soon change all of this (see section IV).

Fig. 10. Top: Optically diplexed THz transmitter source based on 1.5 micron fiber technology and THz power generation through photomixing. Bottom: Gigabit data rate TR module based on this technology. Reprinted with permission from [53].
Another rather clever application for the small spot size associated with THz wavelengths has been to use THz sources to illuminate scale models of large objects thereby simulating the radar scattering signatures (RCS) that would be obtained at much lower frequencies on actual equipment such as planes, tanks and battleships [54,55]. The savings in anechoic test chamber dimensions alone make the high cost of THz test systems attractive in comparison! An example of such a system, pioneered at MIT and University of Lowell, is shown as Fig. 11. Solid-state sources have been used up to 660 GHz [55] and in earlier systems far IR lasers were employed at 1.2, 2.5, and 3.1 and 8 THz [54]. Complete 3D synthetic aperture radar images can be processed with this system (Fig 11 images), by using dual polarization heterodyne transceivers and a special stepped CW scheme which gates out the effects of unwanted signals or chamber reflections.

Fig. 11. (a) THz radar cross-section (RCS) system. Compact range layout and submillimeter transmit/receive arrangement. Reprinted with permission from Coulombe [55].

Fig. 11. (b) Left: Optical image of scale model tank. (c) Right: Processed Submillimeter-wave radar image—cross section of 3D image. Reprinted with permission from Coulombe [55].
Perhaps the most intriguing application for commercializing THz technology at this time is in the area of THz time domain spectroscopy or T-Ray imaging [56,57]. In this technique, pioneered by Martin Nuss and others at Bell Laboratories in the mid 1990’s [58,59] and recently picked up by at least two commercial companies, Picometrix in the US [60] and Teraview, a spin-off of Toshiba Research Europe Ltd. in the UK [61], in-situ measurements of the transmitted or reflected THz energy incident upon a small sample are processed to reveal spectral content (broad signatures only), time of flight data (refractive index determination, amplitude and phase, and sample thickness), and direct signal strength imaging. The principle involves generating and then detecting THz electromagnetic transients that are produced in a photoconductor or a crystal by intense femtosecond optical laser pulses. The laser pulses are beam split and synchronized through a scanning optical delay line and made to strike the THz generator and detector in known phase coherence. By scanning the delay line and simultaneously gating or sampling the THz signals incident on the detector a time dependent waveform proportional to the THz field amplitude and containing the frequency response of the sample is produced (Fig. 12). Scanning either the THz generator or the sample itself allows a 2D image to be built up over time. Recent innovations are leading to both rapid scanning [56] and true 2D sampling using CCD arrays [62]. In the Picometrix and Lucent technologies systems the photoconductive effect in low temperature grown GaAs or radiation damaged silicon on sapphire is used for both the generator and detector. The Teraview system uses THz generation via difference frequency mixing in a nonlinear crystal (ZnTe) and detection via the electro-optical Pockels effect (measuring
the change in birefringence of ZnTe induced by THz fields in the presence of an optical pulse) as first demonstrated by X.C. Zhang at RPI [63]. The femtosecond optical pulses are currently derived from expensive Ti:Sapphire lasers but much effort is being placed on longer wavelength, especially 1.5 micron, solid-state systems that can take better advantage of fiber technology [59]. The RF signals produced by the optical pulses typically peak in the 0.5 to 2 THz range and have average power levels in the microwatt range and peak energies around a femtojoule. This makes T-Ray imaging a very attractive tool for the medical community (non-invasive sampling) as well as for nondestructive probing of biological materials or electronic parts. The technique is rapidly gaining an enormous following and is purged to be an exploding commercial success once the system can be made less costly (replacement of the Ti:Sapphire laser with solid-state devices), faster (through 2D imaging techniques) and somewhat more sensitive (with better sources and detectors). A wide range of applications already exist [57,57,64] and many more will likely appear as commercial systems begin to disseminate.

Fig. 12. (a). T-Ray imager schematic. Reprinted with permission from [59].
Fig. 12. (b). Time domain and frequency domain images of THz transmission through piece of wood. Reprinted with permission from [56].

Visible image of human tooth

Terahertz image of cavity in human tooth

Cavity image composed from absorption data

Fig. 12. (c). Time domain, time of flight and processed 2D data of transmission through extracted tooth. Reprinted with permission from [57].
Finally, there are several new and untested applications that might evolve from advances in THz technology. A room temperature THz heterodyne camera has been proposed [65] that would have many times the sensitivity of the T-Ray imager, as well as the frequency resolution of the scanning spectrometer or single pixel receivers now in common use for Earth science applications. Similarly cryogenic direct detector cameras have been proposed for the submillimeter [66] with extremely low NEP (noise equivalent power) capability. These next generation instruments should enhance the T-Ray applications as well as opening up new opportunities. Leaving the sensors world, microminiature THz power converters (nanorectifiers) that might be incorporated onto nanorobots and operate in-vivo or in hostile environments have been suggested [67]. These nanorectennas could give new meaning to the term miniature power supply! These applications as well as many others might be enabled by THz vacuum nanotube sources [68]. A search for other applications has turned up a couple of curious articles that should be of interest to the adventurous reader [69,70]!

In the following sections we will briefly highlight specific components that are employed for the applications mentioned here and some of the instrumentation that has been constructed from these components for measurement and test. In some cases, particularly in the astronomy and spectroscopy areas, a wealth of published information already exists, including several special issues of MTT and Proceedings of the IEEE [71], and we cannot hope to cover all the technologies that are being brought to bear. Instead we will select a few of the more recent innovations and techniques and leave the remainder to the references. For keeping abreast of advances in this rapidly changing field the interested reader should be sure to scan a few of the now many annual THz conference proceedings, especially: NASA's International Symposium on Space THz Technology, the IEEE's International Conference on Terahertz Electronics, SPIE's International Conference on Millimeter and Submillimeter Waves and the original Ken Button MIT Magnet Lab conference and associated journal, International Conference on Infrared and Millimeter Waves, now in its 26th year. It is also worthwhile keeping an eye on Applied Physics Letters, where much of the solid-state device and component work appears.

4. THz Components

In this section we highlight a few of the major component technologies that have been developed for THz applications. They broadly fall into two categories: sensors and sources. Space does not permit us to examine other THz component building blocks such as guiding structures, quasi-optics, antennas, filters or submillimeter-wave materials. THz component fabrication techniques and device processing are also fields in and of themselves and will not be covered. Many of these topics are addressed in other chapters of this text.

4.1 Sensors

THz sensors have progressed faster than any other submillimeter-wave technology. Today there are near-quantum-limited detectors that can measure both broad band or extremely narrow band signals up to or exceeding 1 THz. There soon promises to be available individual photon counters that should do for the submillimeter what pho-
tomultipliers have done for optical wavelengths. The critical differences between detection at THz frequencies and detection at shorter wavelengths lies in the low photon energies (1-10 milli-electron volts) and in the rather large Airy disk diameter (hundreds of microns). The former condition means that ambient background thermal noise almost always dominates naturally emitted narrow band signals requiring either cryogenic cooling of the detector elements or long-integration-time radiometric techniques or both. The latter condition almost always imposes a mode converter or matched director (antenna) between the signal and the sensor element. Note that the crossover frequency at which an ideal thermal noise limited detector (such as a room temperature Schottky barrier diode) surpasses the sensitivity (power spectral density) of an ideal quantum detector (like a photodiode), occurs between 1 and 10 THz, a sometimes under appreciated advantage of RF over optical detection [72]. In comparison to longer wavelength radio techniques on the other hand, THz sensors suffer from a lack of available electronic components – lumped resistors, capacitors and inductors as well as amplifiers and low loss transmission media. To date the most common THz sensors have been heterodyne detectors, since applications have been centered on high-resolution spectroscopy. This may be changing however and more and more emphasis is now being focused towards direct detection techniques and components. For this quick overview we will look at heterodyne and direct detector technologies, each of which has both room temperature and cryogenic realizations. Since the field is vast, only a surface skim can be accommodated. The interested reader will have to track down additional details through the references.

4.1.1 Heterodyne Semiconductor

For applications involving Earth science and planetary or some in-situ measurements (plasma diagnostics for example) the sensitivity offered by semiconductor sensors is generally adequate for reasonable science return. For passive systems heterodyning is used to increase signal-to-noise by reducing bandwidth. Since low noise amplifiers are not yet available above about 150 GHz (this may soon be changing [73]) signal acquisition is accomplished through frequency downconversion (crystal rectification) and post or intermediate frequency (IF) amplification. The figure of merit (ignoring antenna noise) is the receiver noise figure or equivalent input noise temperature: 

\[ T_r = T_m + LT_{IF} \]

where \( T_m \) represents the downconverter or mixer added noise (shot and thermal contributions), \( L \) is the difference frequency conversion loss and \( T_{IF} \) is the noise added by the first stage amplifier. This determines the minimum detectable temperature difference \( \delta T \), for a given predetection bandwidth \( B \), and post-detection integration time \( \tau \):

\[ \delta T = T_r / \sqrt{(B \tau)} \]  

(Noise Power \( P_n = kT_rB \)). Radiometric techniques in use since the 1940's [74] are still employed for acquiring weak signals embedded in background noise.

For applications where the sensitivity of room temperature detectors is adequate, the basic single Schottky diode mixer is the preferred downconverter in the THz frequency range. Over the past 50 years this component has transitioned from the “cat-whisker” point contact crystal technology prevalent since the beginning of MTT to the “honeycomb diode [75]” of Fig. 13a to a fully integrated planar geometry with much greater flexibility and reliability, as well as superior performance (Fig. 13b). Planar diode mixers have been constructed and space qualified at frequencies as high as 2500 GHz [76] with noise performance below 5000K double sideband. Local oscillator (LO) power is
still a serious issue as semiconductor downconverters that rely on the nonlinear resistance of an exponential diode, require \( \approx 0.5 \) mW of RF drive level at frequencies close to that of the observed signal. Multiple diode configurations, such as balanced and sub-harmonically pumped circuits require even more LO power (\( \approx 3-5\) mW) at submillimeter wavelengths [77]. Receivers based on room temperature Schottky diode mixers typically have radiometric sensitivities (\( \delta T \)) near 0.05K at 500 GHz and 0.5K at 2500 GHz for a one second integration time and a 1 GHz predetection bandwidth. This is sufficient for detecting many naturally occurring thermal emission lines. Some improvement (typically 2-4X) can be obtained with cooling of the detector as was demonstrated three decades ago by Weinreb and Kerr [78]. A wealth of papers have been published over the years describing many variations on the Schottky diode downconverter, including traditional waveguide based and planar antenna/diode combinations (“mixtennas” [79]). The interested reader is referred to the many MTT Transactions and MTT conference articles for detailed designs and performance as well as the texts by Kollberg [80] and Maas [81] and two nice but older overview papers by Blaney [82] and Clifton [83].

![Image of Schottky diode](image)

**Fig. 13.** Left: Whisker contacted Schottky honeycomb diode circa 1980 for 200 GHz operation in waveguide downconverter. Chip width is 125 \( \mu \)m. Right: Planar integrated submicron area Schottky diode on 3\( \mu \)m thick x 30 \( \mu \)m wide GaAs membrane for operation at 2500 GHz, also coupled to a waveguide mount.

### 4.1.2 Heterodyne Superconductor

High sensitivity detectors must rely on cryogenic cooling for operation in the THz range. Several superconducting heterodyne detectors have been developed including those based on the Josephson effect [84], superconductor-semiconductor barriers (super-Schottky [85]) and bolometric devices [86]. However the superconducting equivalent of the Schottky diode downconverter (in terms of widespread use below 1 THz) is the superconductor-insulator-superconductor (SIS) tunnel junction mixer. The current flow mechanism is based on the photon assisted tunneling process discovered by Dayem and Martin [87] in the early 1960’s. The first receivers using this effect were developed by groups at Bell Laboratories, Holmdel, N.J., and the California Institute of Technology, Pasadena, [88] and at the University of California at Berkeley [89] and analyzed by Tucker [90] in the late 1970’s. SIS mixers are widely used at frequencies from 100 GHz to 700 GHz and very recently up to 1200 GHz [91] at observatories around the world and will soon be flown in space [15]. Like the Schottky diode down-
converter, the SIS mixer (Fig. 14) relies on an extremely nonlinear I-V characteristic, in this case created by the sharp onset of tunneling between the single-electron quasiparticles on either side of a thin superconducting gap. The tunneling process is non-classical and the sensitivity limit (equivalent mixer noise temperature) is governed by the Heisenberg uncertainty principle \( T_m = \frac{h}{2k} \) \[^92\]. In fact the mixer conversion efficiency can actually become a gain under certain circumstances \[^93,94,95\]. Sensitivities for the SIS mixers fall close to the quantum limit \((h\nu/2kT)\) for frequencies up to several hundred GHz and are within a factor of 10 of this limit up to 1 THz. Receiver noise temperatures (dominated by antenna, optics and mixer mount losses) vary from less than 50K at 100 GHz to over 500K above 1 THz (Fig. 15). An additional advantage of the SIS mixers is the very modest LO power requirement compared to Schottky diodes; on the order of microwatts, rather than milliwatts, for a single device. SIS devices reach a natural frequency limit, \( f_{\text{cutoff}} \), at approximately twice the superconducting energy gap, \( 2\Delta = 3.5kT_c \) for ambient temperatures well below the critical temperature, \( T_c \) \[^92\]. This upper operating frequency is dependent upon the tunnel junction material composition: \( f_{\text{cutoff}} = 146 T_c \). For the most common SIS tunnel junctions, niobium-aluminum oxide-niobium with \( T_c = 9.3\)K, this frequency falls near 1350 GHz. Alternate materials such as niobium nitride \( T_c = 16\)K or high \( T_c \) superconductors based on YBCO \( (T_c > 90\text{K}) \), have much higher operational frequency limits, but acceptable tunnel junctions have yet to be formed from these compounds. Since SIS devices have been around for less time than Schottky diodes, there have been significantly less circuit variations proposed, however a large number of submillimeter-wave systems have now been constructed and there are hundreds of papers on the subject. Two nice, but older review articles can be found in Proc. IEEE \[^96,97\] and a general review of submillimeter wave low noise receivers with more than 200 references was published more recently by Carlstrom and Zmuidzinas \[^98\]. The reader may have to stray outside MIT for many of the recent results, which tend to appear in physics and astronomy journals (Applied Physics Letters and Astrophysical Journal) or in the Applied Superconductivity and NASA Space THz Technology conference digests.

Fig. 14. Left: 1200 GHz SIS tunnel junctions integrated on planar slot antenna circuit. Right: Current-voltage relationship with/without applied LO power and IF output signal (Courtesy A. Karpov and J. Zmuidzinas \[^91\]).
An alternative to THz SIS mixers which has gotten a great deal of attention in the last several years is the transition-edge or hot electron bolometer (HEB) mixer [99]. Unlike the InSb bolometers of [86], modern HEB mixers are based on extremely small microbridges of niobium, niobium nitride or niobium titanium nitride (and recently aluminum [100] and even YBCO [101]) that respond thermally to THz radiation (Fig. 16). These micron and submicron sized HEB's can operate at very high speeds through either fast phonon [102] or electron diffusion cooling [103]. For heterodyning, the bolometer has a voltage responsivity in the picosecond range, fast enough to track the IF (intermediate frequency) up to several GHz. Depending on the dimensions and the composition of the material forming the microbridge and the cooling mechanism – electron-phonon into a superconducting bath or electron-electron diffusion into a normal metal – intermediate frequency roll-offs of more than 15 GHz can be obtained [104]. Since the bolometer is inherently a resistive device, its RF response is limited largely by the signal coupling antenna. HEB noise performance is dependent on the difference between the operating (bath) and the material critical temperature as well as the sharpness of the transition between the superconducting and normal states of the bridge. Sensitivities are predicted to be only a few times higher than the quantum limit [104] and the required LO power is even lower than for SIS mixers, falling in the 1-100nW range [104]. Excellent mixer performance has recently been achieved near 2.5 THz ($T_c$=1050K DSB, only ~10x the quantum limit) [105,106,107] and operation above 5 THz has been reported [108]. The HEB has revolutionized broadband THz detectors and has already been deployed at
several observatories. HEB mixers are planned for operation on at least one space mission [15]. Some recent reviews of HEB mixing can be found in [104,109], but the field is advancing so quickly it is advisable to keep a close eye on the Space THz Technology and Applied SuperConductivity conference proceedings.

![Image of Niobium Hot Electron Bolometer](image)

**Fig. 16.** Left: 2.5 THz Niobium Hot Electron Bolometer twin-slot antenna mixer. Right: Current-voltage characteristics with and without LO power applied. Also shown is the dc (Vdc) and ac (AV) components of the IF output power. (Courtesy W.R. McGrath [106, 107]).

### 4.1.3 Direct Detectors

Direct detectors are rapidly encroaching into the realm of heterodyne systems for applications that do not require ultrahigh spectral resolution. Room temperature detectors have limited applications outside diagnostic measurements due to sensitivity constraints. Small area GaAs Schottky diodes [110] used as antenna coupled square law detectors, conventional bolometers based on direct thermal absorption and change of resistivity (e.g. bismuth [111]), composite bolometers which have the thermometer or readout integrated with the radiation absorber (bismuth [112] or tellurium [113]), microbolometers [114] which use an antenna to couple power to a small thermally absorbing region, Golay cells [115] based on thermal absorption in a gas filled chamber and a detected change in volume via a displaced mirror in an optical amplifier, an acoustic bolometer [116] which reads out the change in pressure of a heated air cell using a photo-acoustic detector, and a fast calorimeter [117] based on single mode heating of an absorber filled cavity, have all been used at THz frequencies. Calibration (mode matching) can be a serious problem for the antenna coupled (small area) detectors (diode and microbolometers), and response time is on the order of seconds for the calibrated, but poorer sensitivity (several microwatt level), acoustic bolometer and fast calorimeter. Cooled detectors take several forms, the most common commercial systems being helium cooled silicon, germanium or InSb composite bolometers [118] with response times on the microsecond scale. NEP is typically $10^{-13}$ W/√Hz for 4K operation and improves greatly at mK temperatures. Some traditional IR detectors also respond in the submillimeter including pyroelectric [119] which change their dielectric constant as a
function of temperature and direct photoconductors based on mechanically stressed gallium doped germanium (Ge:Ga) [120] or even HgCdTe [121]. Non-commercial cooled bolometers of many forms have been used since the 1970's for spectroscopy and astronomical observations. Especially important are the transition edge bolometers [122] based on the change of state of a superconductor. An excellent review up to 1994 can be found in Richards [123]. Recent progress includes detectors and arrays based on bismuth-coated suspended micromachined silicon, doped to obtain the desired resistive temperature dependence [124,125], absorber-backed neutron-transmutation-doped (NTD) germanium on silicon nitride "spider" bolometers [126], and superconductor-insulator-normal (SIN) metal tunnel junction composite bolometers [127]. NEP's for these devices are $10^{-17}$ to $10^{-18}$ W/Hz. The spider bolometer is expected to fly in space as part of the SPIRE (Spectral and Photometric Imaging REceiver) instrument on FIRST. SIS mixers also can, and have been, used as high sensitivity detectors [128]. More recently, detectors based on the HEB mixer geometries (antenna coupled) working as hot electron superconducting transition edge sensors [129] have been proposed. These devices are predicted to have NEP's near $10^{-20}$ W/Hz [130].

Before leaving the detector area it is worth mentioning one other device that is catching quite a lot of attention, the quantum dot single photon detector developed by S. Komiyama [131,132]. This detector (Fig. 17) uses a cold (50mK) single electron transistor (SET) and quantum dot in a high magnetic field. Incident THz photons are coupled into the quantum dot via small dipole antennas.

![Fig. 17. The Komiyama quantum-dot single-photon detector concept. Reprinted with permission from [131]. ©2000, Nature.](image_url)
Within the quantum dot an electron-hole pair created by the incident photon releases energy to the lattice, which causes a polarization between two closely coupled electron reservoirs. Electron tunneling occurs, causing a shift in the gate voltage of the SET. According to the authors, the detector has a sensitivity of 0.1 photons per second per $0.1 \text{mm}^2$ detector area due to the photomultiplication effect where $10^6-10^{12}$ electrons per photon are generated. Single photon detection for signals in the range of 1.4-1.7 THz has been recorded. The equivalent NEP is on the order of $10^{-22} \text{W/}\sqrt{\text{Hz}}$, more than 1,000 times more sensitive than the best bolometric devices, but the speed of the detector is presently around 1 ms. The authors have hopes of taking advantage of the potential of new RF-SET devices [133] that have intrinsic speeds near 10 GHz. Many other devices have been proposed as high sensitivity detectors based on bolometric effects or photon counting [134] including some high $T_c$ and even cooled semiconductor constructions [135]. Again, the author's advice is to keep a close eye on the aforementioned conferences to keep abreast of progress in this area.

4.2. Sources

Certainly the most difficult component to realize in the submillimeter-wave bands has been the THz source. There are several fundamental explanations for this. Traditional electronic solid-state sources based on semiconductors, i.e. oscillators and amplifiers, are limited by reactive parasitics or transit times that cause high frequency roll off, or they have simple resistive losses which tend to dominate the device impedance at these wavelengths. Tube sources suffer from simple physical scaling problems, metallic losses and the need for extremely high fields (both magnetic and electric) as well as high current densities. Optical style sources, solid-state lasers, must operate at energy levels so low (~meV) they are comparable to that of the lattice phonons (relaxation energy of the crystal) although cryogenic cooling can mitigate this problem. More successful techniques for generating THz power have come from frequency conversion, either up from millimeter wavelengths, or down from the optical or IR. Many approaches have been tried with the most successful milliwatt sources to date being direct laser-to-laser (far-IR to submillimeter) pumping or reactive multiplication through our old standby the GaAs Schottky diode. For generating narrowband microwatt or nanowatt power levels at THz frequencies a variety of techniques have been demonstrated including optical mixing in nonlinear crystals [136,137], photomixing (optical difference frequency mixing in a photoconductor coupled to an RF radiator) [138,139,140,141], picosecond laser pulsing [142,143], laser sideband generation [144,145], intersubband and quantum cascade lasing [146,147], direct semiconductor oscillation with resonant tunneling diodes (RTD) [148,149,150], direct lasing of gases [151], Josephson junction oscillations [152] and other techniques too numerous to mention. We can only take a brief scan of the more successful approaches and again the avid reader is relegated to the references for more details.

4.2.1 Upconverters

By far the most common technique for producing small amounts of power at frequencies above 500 GHz is through nonlinear reactive multiplication of lower frequency oscillators. The field was actually better off twenty years ago when solid-state (Gunn
and IMPATT diodes) and tube sources (carcinotrons, klystrons and backward wave oscillators) were more readily available at millimeter-wave frequencies. The demise of cold war funding, advances in III-V MMIC technology and the draw of enormous commercial markets in the communications bands below 100 GHz has left the field with few interested manufacturers of expensive custom-tailored high frequency components. Although tube sources are still trickling out of the last commercial supply house in Istok, Russia, and small stockpiles of the original Varian-then-Litton-now-Filtronic high frequency (>120 GHz) InP Gunn and Hughes-now-Boeing IMPATT devices can be found tucked away in the drawers of a few small millimeter wave component companies, the most reasonable approach in the source area today is to multiply up from microwave frequencies (20-40GHz). Although higher frequency (up to 200 GHz) Gunn, IMPATT and TUNNETT devices are in development within University groups, especially by Eisele [153] they are not available commercially. Fortunately, commercially based W-band (75-110GHz) InP MMIC power amplifier chips have appeared [154] and promise to extend baseband frequency coverage up to at least 200GHz within the very near future [155,156]. At 100 GHz, power levels from waveguide-combined amplifiers of 300-400 mW are readily available [157] and there is a large effort underway to develop spatial power combined sources with many times this output level [158,159,160]. The amps can be driven by commercially available microwave oscillators (VCO’s, DRO’s) and millimeter-wave upconverters.

The need for narrow band compact solid-state THz sources is being driven, at least partially, by space applications such as FIRST/Herschel [161] that cannot fly bulky power hungry lasers or short lived, very heavy, kV driven tube sources. In order to get from W-band to THz frequencies through solid-state upconversion, several octaves must be spanned. Despite no theoretical limitations [162,163] and the best efforts of many researchers over many decades, comb generators and higher order multipliers (>X4) continue to provide extremely poor conversion efficiencies compared to doublers and triplers [164]. The most efficient THz sources are therefore composed of series chains of these lower order multipliers. Like the room temperature downconverter technology, today’s multiplied sources most commonly use planar (as opposed to whisker contacted) GaAs Schottky barrier diodes mounted in single mode waveguide, although the literature is replete with optically coupled circuits. Small electrical size and assembly constraints have led to some unusual and extremely low loss device topologies [165] an example of which is shown in Fig. 18. An additional constraint for higher order multiplying (due to low overall efficiency) is the power handling capacity of the first few stages of the chain, which can be beefed up by adding multiple devices in series to distribute the heat and increase the breakdown voltage [166]. Multiplier chains driven by amplified sources at 100 GHz have reached 1200 GHz with 75 μW at room temperature and over 250 μW when operated cold (120K) [167]. Signals up to 2.7 THz have been obtained with this technique [168] and the reported efficiencies and output power are improving monthly, driven by the instrument needs (and accompanying funding!). Fig. 19. shows the current state of affairs. All of the solid-state device-based sources, including amplifiers, multipliers and oscillators (RTD’s, TUNNETT’s, Gunn’s) will of course benefit from spatial power combining techniques [169], a field in and of itself. The literature, especially the MTT and NASA Space THz Technology conferences, is filled with even more THz upconverter than downconverter designs and the interested reader can scan them until they begin seeing double!
Fig. 18. THz Schottky diode multiplier chips fabricated monolithically from GaAs. Left: 800 GHz balanced doubler with thin membrane frame mounted in waveguide block [165]. Right: 2400 GHz waveguide mounted substrateless doubler schematic (currently in process at JPL) [161].

Fig 19. Performance of some CW sources in the submillimeter wave range. Data: Schottky varactor [JPL recent results], RTD [149,150], Photomixer [53,140], Carcinotron [http://www.pi1.physik.uni-stuttgart.de/index.js.html], Laser lines [183].
4.2.2 Tubes, Lasers and Optical Downconverters

Although we cannot look very closely at other THz sources, a mention of the most likely contenders in this most important development area is certainly warranted. Despite their limited availability, THz tube sources [170] based on emission from bunched electrons spiraling about in strong magnetic fields (backward wave tubes or carcin­trons) offer the most power and frequency tuning range at submillimeter wavelengths. Bench top commercial units (as opposed to room-sized kW clinotrons and gyrotrons [171]) extend to 1200 GHz [172] with mW levels of available power, although multi­moding is a significant problem when coupling these sources to actual antennas (typical TE_{10} mode coupled power can be 1/1000th or less of the total output power from the tube). The Russian tubes (the only ones still available) operate on 2-6 kV and ~1 Tesla fields that can be achieved by samarium cobalt permanent magnets. They are sweepable over ranges exceeding tens of GHz at kHz rates and can be phase locked [173] for high stability applications. The tubes with magnet cost from $15,000-$75,000 each and have lifetimes of a few thousand hours at best. Despite the limitations, the tube concept is a slow one to die and there are currently several research groups trying to resurrect kly­strons or other tube configurations [174] for THz operation by bringing in modern monolithic fabrication techniques and high-density cathode development [175,176,177]. Whether these efforts will yield successful THz oscillators remains to be seen.

The next most commonly used sources at THz frequencies are IR-pumped gas lasers such as those produced commercially by DEOS [178], Edinburgh Instruments [179] and MPB technologies [180]. These are usually based on grating tuned CO_2 pump las­ers (20-100 W) injected into low-pressure flowing-gas cavities that lase to produce the THz signals. Power levels of 1 to 20 mW are common depending on the chosen line, with one of the strongest being that of methanol at 2522.78 GHz. In the 500 GHz to 3 THz regime not all frequencies are covered by strong emission lines and some of the nastier gases are best left in their shipping containers under today's environmental safety policies. None-the-less laser sources have been in use for many years for spectroscopy and as THz LO sources in receivers. A fully sealed autonomous CO_2 pumped methanol laser has been space qualified and will fly on the EOS-MLS instrument in 2003 [22]. Total plug power for this system is only 120W, with 20W CO_2 output power and more than 30mW of generated RF at 2523 GHz. Nice reviews can be found in [181,182] and [183]. The lasers are also used with harmonic generators to make side­band sources [145,184] which have much more flexible tuning. While on the subject of laser sources one cannot ignore the many laboratory efforts underway to develop direct semiconductor THz laser sources based on intersubband transitions and tailored quantum cascade laser devices, most notably by Capasso and Cho at Lucent [185,186]. Most of these techniques will require cooling of the devices [187] and there is still a long way to go in extending out to THz frequencies, but there has been significant progress over the last several years and complex tailorable layers grown on an ever expanding list of compound semiconductor materials hold hope for significant CW power sources in the not too distant future.

One cannot leave sources without mentioning today’s most commercially successful technique for generating THz energy - downconversion from the optical regime. Two principal methods have been exploited to produce both narrow- and broadband energy.
The first, photomixing [138,139,188], uses offset-frequency-locked CW lasers focused onto a small area of an appropriate photoconductor (one with a very short, <1 picosecond, carrier lifetime e.g. low-temperature-grown LTG GaAs) to generate carriers between closely spaced (<1 μm) electrodes (source and drain) printed on the semiconductor. The laser induced photocarriers short the gap producing a photocurrent, which is modulated at the laser difference frequency. This current is coupled to an RF circuit or antenna that couples out or radiates the THz energy. The resulting power is narrow band, phase lockable and readily tuned over the full THz band by slightly shifting the optical frequency of one of the two lasers. Both 780-820 nm Ti:Sapphire and 850 nm distributed Bragg reflector (DBR) semiconductor lasers have been used to match the band gap of the LTG GaAs. Simple optical fiber coupling to the photoconductor has also been employed [189] and work on new materials (Erbium Arsenide doped LTG GaAs) that would enable operation in the optical fiber band (1.3-1.55 microns) is ongoing [190]. Typical optical to THz conversion efficiencies are below 10^{-5} for a single device [189] and reported output power falls from ≈1μW at 1 THz to below 0.1 μW at 3 THz even with the newer distributed or traveling-wave photomixer designs [140]. Cooling, arraying and new materials hold great promise for these photomixers however and modest improvements will open up applications both in radiometry (as broadly tunable local oscillator sources) and communications [53]. A second, and perhaps the most widely employed optical technique for producing THz energy (although broadband) is based on using a short pulse (femtosecond) optical laser [191,192,193] (argon-laser-pumped Ti:Sapphire laser) to illuminate a gap between closely spaced electrodes on a photoconductor (e.g. silicon-on-sapphire or LTG GaAs) generating carriers which are then accelerated in an applied field (<100V). The resulting current surge, which is coupled to an RF antenna, has frequency components that reflect the pulse duration, i.e. THz rates. The same THz output spectrum can be obtained by applying short laser pulses to a crystal with a large second order susceptibility, χ^2 (field induced polarization) like zinc telluride [61]. Since the higher order susceptibility terms are indicative of nonlinear response, mixing occurs, producing a time varying polarization with a frequency response representative of the pulse length, i.e. THz oscillation. These techniques are the basis for the T-Ray systems [60,61] described previously. As with the photomixers, RF power may be radiated by antennas printed on the photoconductor or crystal, and typically have frequency content from 0.2-2 THz or higher depending on the laser pulse parameters. Average power levels over the entire spectrum are very low (nanowatts-to-microwatts) and pulse energies tend to be in the femto-to-nanojoule range [61]. Fiber coupled systems have been developed [60] and continual improvements are being made to these RF generators in order to increase the signal-to-noise of the T-Ray imagers. Finally, it is worth mentioning one last pulsed laser technique that shows promise for high levels of THz output power. The technique [194,195] uses a Q-switched Nd:YAG laser to illuminate a large LiNbO_3 sample causing optical parametric oscillation via the crystal χ^2 and polariton mode scattering. The parametric process creates a near IR photon close in wavelength to that of the Nd:YAG pump and a THz difference photon. A prism-coupler is used to extract the THz power and tuning can be accomplished by simply changing the incident angle of the Nd:YAG laser. The system is shown in Fig. 20 and pulsed power output (line width ≈15 GHz) is reported to be as high as 3mW at 1.8 THz!
5. Future Applications and Concluding Remarks

It is clear that THz technology is just beginning to come of age and many applications yet to be realized lie in wait. The most pressing component technology development remains in the area of THz sources, just as it did in 1963 [9] and in 1985 [10]. However advances on several fronts in the solid-state device and laser pumped photoconductor areas are pushing hard at this bottleneck and it should not be too many more years before we will have sufficient power to do heterodyne imaging, radar or communications with THz signals. The sensor side has progressed remarkably, and near quantum limited receivers up to 1 THz are becoming widespread. Beyond 1 THz, direct detectors are grabbing territory at an alarming rate and the promise of single photon counters in the submillimeter may usurp some quantum limited optical applications. Unless sources make some quick progress, direct detectors rather than heterodyne systems will likely dominate the THz regime in the near term, especially for space and low background signals. As sensors and sources become more available, more complex circuits and eventually complete instruments will follow. We have not even scratched the surface in this area and many more pages would be needed to do justice to the component developments that have taken place already at THz frequencies. On the instrument side we are just beginning to see emerging systems. A THz network analyzer is commercially available [196], near field antenna measurements have been performed at 640 GHz [197], a recently demonstrated millimeter-wave near field microscope technique [198] promises micron resolution THz imaging once sources have been developed, the T-Ray system is now being employed on a wide variety of samples from the electronics industry to medical diagnostics and will likely be the medium for introducing the public to
THz wavelengths for the first time. As laser systems advance so that the pulsed and higher power bench-top models are replaced by solid-state semiconductor devices there will be dramatic reductions in the instrument envelopes as well as tremendous cost savings. There is already an enormous program (by submillimeter standards) with nine research partners, recently started in Europe to evaluate and coordinate findings on the biological aspects of THz radiation and THz imaging [199]. Even in the spectroscopy area, commercial instruments are on the near horizon. High resolution FTS systems like FASSST may see use in military systems as chemical agent detectors and a substantial MURI (Multidisciplinary University Research Institute) program [200] promises to help open up biomedical applications for this technology in the US. In the space science community, submillimeter waves have already reached their golden era and the groundwork for a long string of astrophysics, Earth and planetary sensor systems has been laid. Ultimately, the author predicts that submillimeter systems will complement near IR and optical sensors for interferometric measurements in the search for extraterrestrial life, although the case has yet to be made. All of these exciting applications and countless undiscovered ones remain in wait while THz technology enters adulthood - *div venturus erat et ibi est multus labor faci* - it has been a long time coming and there is still much work to be done.

Acknowledgements
This article could not have been written without the many conversations and contributions the author had from his friends and colleagues in the submillimeter-wave field. He would like to thank individually those people who contributed specific text, drawings, references or comments which have been included in the manuscript: Dr. Don Arnone (Teraview), Professor Elliott Brown (UCLA), Dr. Michael Coulombe (Univ. of Lowell), Professor Frank DeLucia (Ohio State), Dr. Calvin Domier (UC Davis), Dr. Michael Gaidis (IBM), Dr. Erich Grossman (NIST), Dr. Sam Gulkis (JPL), Professor Tatsuo Itoh (UCLA), Dr. Alexandre Karpov (CalTech), Professor Susumu Komiyama (Univ. Tokyo), Dr. William Langer (JPL), Dr. David Leisawitz (NASA Goddard), Professor Neville Luhmann (UC Davis), Dr. Edgar Martinez (DARPA), Dr. John C. Mather (NASA Goddard), Dr. Shuji Matsura (Inst. Space Astronautical Sciences, Japan), Dr. William McGrath (JPL), Professor Daniel Mittleman (Rice Univ.), Dr. Eric Mueller (DEOS), Dr. Leon Newman (DEOS), Professor Tony Peebles (UCLA), Professor Tom Phillips (CalTech), Dr. Dwight Woolard (ARO), Dr. Harold Yorke (JPL) and Dr. Jonas Zmuidzinas (CalTech). The author would also like to thank the editor of this special issue of MTT, Dr. Leo Young, and the section editor Professor David B. Rutledge, for the opportunity to contribute and for support during the writing of the manuscript. Finally, the author would like to thank all the members of the JPL Submillimeter Wave Advanced Technology (SWAT) team for their many contributions to NASA applications and to THz technology – he is honored to have had the privilege of working with such a wonderfully talented group of individuals over the past 8 years. This task was supported by the California Institute of Technology Jet Propulsion Laboratory under contract with the National Aeronautics and Space Administration.
References


4 The Oxford English Dictionary dates the term terahertz back to at least 1970 where it was used to describe the frequency range of a HeNe laser.

5 In 1947, the International Telecommunications Union designated the highest official radio frequency bands (EHF-extremely high frequency) as bands 12-14, 300kMc-300MMc (1 MMc=1 THz).


7 Erich Grossman, Electromagnetic Technology Division (814), Radio (1), Room 124, NIST, Mail code 814.03, 325 Broadway, Boulder, CO 80303, personal communication.


13 Figures 3 and 4 courtesy William Langer, from the Herschel Space Observatory archive at JPL.


21 Fig. 5 based on: J.W. Waters, "A 'Focused' MLS for EOS," EOS Atmospheres Panel presentation, NASA GSFC, Dec. 9, 1991.


http://solarsystem.nasa.gov/roadmap/pdffiles/Rmap.pdf


See also: http://sci.esa.int/rosetta/


40 See for example the following classic texts:


41 The Ohio State University International Symposium on Microwave Spectroscopy, now in its 56th year.


51 Leon Newman, private communication.


MTT and IEEE Proceedings Special Issues and Papers include:
    MTT-43, no.4, “Special Issue on Terahertz Technology,” 1995;
    MTT-24, no. 6, “Special Issue on the Proceedings of the First Int. Conf. on Submillimeter Waves and their Applications,” 1976;
    MTT-22, no. 12, “Special Issue on the Proceedings of the First Int. Conf. on Submillimeter Waves and their Applications,” 1974;

72 Alexander, Stephen B., Optical Communication Receiver Design SPIE Optical Engineering Press ; London, UK : Institution of Electrical Engineers, cl997, Fig. 5.4, page 129 (the author thanks Hamid Hemmati of JPL for this reference).


115 A submillimeter wave Golay cell with an NEP of $10^{-10}\text{ W/}\sqrt{\text{Hz}}$ can be purchased from QMC Instr. Ltd., Cardiff Univ., Cardiff, UK, http://www.terahertz.co.uk


172 Tubes are distributed in the US by Insight Products, Brighton, MA or by ELVA-1, St. Petersburg, Russia. http://www.elva-1.spb.ru/


178 DeMaria Optical Systems, New Haven, CT. http://www.deoslaser.com

179 Edinburgh Instruments, Livingston, UK. http://www.edinst.com


197 SWAS, MIRO and MLS all used customized near field ranges to measure large antenna beam patterns at 470, 557 and 640 GHz respectively.


199 This program known as THz-Bridge (Tera-Hertz radiation in Biological Research, Investigation on Diagnostics and study of potential Genotoxic Effects) already has 9 university and industry partners investigating a wide range of THz applications and effects in the bio area. http://www.frascati.enea.it/THz-BRIDGE

200 "The Science and Technology of Chemical and Biological Sensing at Terahertz Frequencies Project," T. Crowe, PI, Awarded through US Army Research Office to University of Michigan, Ann Arbor, University of Virginia, Charlottesville, Stevens Inst. of Technology, Hoboken, N.I., University of California at Los Angeles and Univ. of Tennessee, Knoxville, 2001.
TWO-TERMINAL ACTIVE DEVICES FOR TERAHERTZ SOURCES

G. I. HADDAD, J. R. EAST, AND H. EISELE
Solid-State Electronics Laboratory, Department of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor, MI 48109-2122

The terahertz frequency range of the electromagnetic spectrum holds great promise for many applications including sensing, imaging, and communications. However, the availability of solid-state sources with reasonable power levels is well recognized as one of the major obstacles for systems applications in this frequency range. Here, the state of the art of active two-terminal devices is reviewed and highlighted with some exemplary experimental results. The potential, capabilities, and limitations of two-terminal devices regarding power generation at terahertz frequencies are discussed. Furthermore, a new device is described that has the potential of circumventing some of the limitations of existing devices.

Keywords: Ballistic devices, IMPATTS, TUNNETTS, transferred electron devices

1. Introduction

Two-terminal devices were the first solid-state devices to be employed for power generation and amplification at microwave frequencies. Although they have been replaced mostly by three-terminal devices at microwave and lower millimeter-wave frequencies, they still hold record performance in terms of power generation capability, particularly at higher millimeter- and submillimeter-wave frequencies. They also have the potential of reaching terahertz frequencies and generating significant power levels.

Several types of two-terminal devices are suitable for radio frequency (RF) power generation and they include Esaki tunnel diodes (TDs),\textsuperscript{1,2} resonant tunneling diodes (RTDs),\textsuperscript{3} transferred-electron or Gunn-effect devices (TEDs),\textsuperscript{4,5} and transit-time diodes. Transit-time diodes utilize various types of injection mechanisms including avalanche breakdown, interband tunneling, and barrier injection, and, depending on the injection mechanism employed, they are generally referred to as impact ionization transit-time (IMPATT),\textsuperscript{6} tunnel injection transit-time (TUNNETT),\textsuperscript{7} and barrier injection transit-time (BARITT)\textsuperscript{8,9} diodes.

All of these two-terminal active devices exhibit the property of a negative differential resistance (NDR), although the basic mechanisms for generating this NDR are quite different. Such devices can be used either as reflection-type amplifiers or oscillators. Reflection-type amplifiers can be realized either by using a circulator to separate the output from the input or by injection locking of the device operating as an oscillator. Mainly the use of these devices as oscillators is discussed here.

2. Two-terminal NDR devices as oscillators

An equivalent circuit for an oscillator utilizing an NDR device is shown in Fig. 1. The admittance per unit area $Y_D$ is given by
The device admittance is

$$Y_D = G_D + jB_D,$$

(1)

and the total admittance

$$Y_a = AY_D = AG_D + jAB_D,$$

(2)

where $A$ is the device area.

![Simplified equivalent circuit of an oscillator with a two-terminal device connected to a load.](image)

Fig. 1. Simplified equivalent circuit of an oscillator with a two-terminal device connected to a load.

The device impedance is

$$Z_a = \frac{1}{Y_a} = R_a + jX_a,$$

(3)

where

$$R_a = \frac{G_D}{A(G_D^2 + B_D^2)},$$

(4)

and

$$X_a = \frac{-B_D}{A(G_D^2 + B_D^2)}.$$

(5)

The oscillation condition is satisfied for $-R_a = R_s + R_L$ and $-X_a = X_L$. This results in

$$X_L = \omega_s L_s = -X_a = \frac{1}{\omega_s C_s},$$

(6)

where $\omega_s = \frac{1}{\sqrt{L_s C_s}}$ is the oscillation frequency.

For the devices under consideration and at high operating frequencies, $|B_D| >> |G_D|$ and Eqs. 4 and 5 reduce to

$$R_a = \frac{G_D}{AB_D^2},$$

(7)
and

\[ X_d = -\frac{1}{AB_D}. \tag{8} \]

The RF power generated by the device is given by

\[ P_{RF}(Gen.) = -\frac{1}{2} V_{RF}^2 AG_D, \tag{9} \]

where \( V_{RF} \) is the magnitude of the RF voltage.

Equation 7 and the matching condition for oscillation require

\[ -\frac{G_D}{AB_D^2} = R_s + R_L. \tag{10} \]

Therefore,

\[ A = -\frac{G_D}{B_D^2(R_s + R_L)} \tag{11} \]

and

\[ P_{RF}(Gen.) = \frac{1}{2} V_{RF}^2 \frac{G_D^2}{B_D^2(R_s + R_L)}. \tag{12} \]

The RF power delivered to the load \( R_L \) is reduced by the series resistance \( R_s \) and is given by

\[ P_{RF}(R_L) = P_{RF}(Gen.) \frac{R_L}{(R_s + R_L)} = \frac{1}{2} V_{RF}^2 \frac{G_D^2}{B_D^2 R_L \left(1 + \frac{R_s}{R_L}\right)^2}. \tag{13} \]

A knowledge of \( G_D \) and \( B_D \) for various operating parameters and designs of a particular device is necessary to estimate the power generation capabilities and the appropriate device areas from Eqs. 11 and 13, respectively.

Two-terminal NDR devices can be operated under pulsed conditions as well as in the continuous-wave (cw) mode, which results in two limits of how much RF power the device can generate. The first limit is electronic and the second is thermal. The electronic generation limit comes from matching to the load resistance \( R_L \) including the series resistance \( R_s \) of the device (and circuit). The thermal limitation affects how much direct-current (dc) input power the device can handle safely and is mainly determined by the thermal resistance of the device, which in turn depends on various parameters, including device area, heat sink type, and layer structure.
3. Fabrication technologies and oscillator circuits

Dc-to-RF conversion efficiencies of two-terminal NDR devices invariably decrease with operating frequencies and, at millimeter-wave frequencies and higher, the values are generally medium to low. As a consequence, almost all of the dc input power needs to be effectively dissipated as heat. Therefore, most fabrication technologies for two-terminal NDR devices attempt to minimize the heat-flow resistance from the active region to an external metal heat sink. Additionally, operation at millimeter-wave frequencies and above requires thin devices to reduce losses that result from the skin effect in the undepleted substrate and are a contributor to $R_s$. The integral heat sink (IHS) process is the most common fabrication technology for commercially available InP Gunn devices. It is also employed in the fabrication of some of the commercially available GaAs Gunn devices, but here a flip-chip mounting process is more commonly used. As other examples, BARITT diodes and some of the Si and GaAs IMPATT diodes were fabricated also using this type of fabrication process.

The advent of advanced growth techniques such as metalorganic chemical vapor deposition (MOCVD) or molecular-beam epitaxy (MBE) greatly helped simplify some of the fabrication tasks. Since these growth techniques allow lattice-matched epitaxial layers of different compositions to be grown on GaAs or InP substrates, fabrication technologies can be tailored to using highly selective chemical etchants. As an example, the flow chart of a recently developed fabrication process for devices with an integral heat sink is illustrated in Fig. 2. In the first step, the metalization for the $n$ ohmic contact (Ni/Ge/Au/Ti/Au) is evaporated or sputtered onto the surface. A thick gold layer is then electroplated onto this metalization to form the integral heat sink. The sample is mounted on a chemically inert carrier to provide additional mechanical support and protect the heat sink during the subsequent processing steps. The substrate is removed in a selective etchant of diluted $\text{H}_2\text{O}_2$, which does not attack the In$_{0.53}$Ga$_{0.47}$As etch-stop layer. Good ohmic contacts can be formed on InP or, with lower specific contact resistance, on In$_{0.53}$Ga$_{0.47}$As. Therefore, this In$_{0.53}$Ga$_{0.47}$As layer need not be removed, but may be etched away selectively in a standard solution of phosphoric or sulfuric acid, hydrogen peroxide, and water as indicated in Fig. 2. Such a solution does not attack InP. A photolithography step defines the openings on this InP surface (or In$_{0.53}$Ga$_{0.47}$As surface if left in place), where the metalization (Ni/Ge/Au/Ti/Au) for the other $n$ ohmic contacts on the second heavily $n^+$ doped layer is deposited. Excess metal outside the contacts is lifted off with the photoresist and, using another photolithography step, the contacts are selectively electroplated with several microns of gold to form a good bonding pad. The contact pad acts as a mask when the mesa of the diode is etched in a nonselective wet etchant. After the sample has been removed from the carrier, the contacts are annealed and the sample is diced into individual diodes. Diodes are then mounted in packages for appropriate RF circuits.

Much improved heat dissipation through diamond heat sinks enabled the most recent performance enhancements in InP Gunn devices. Similar processes for devices on diamond heat sinks are also the most commonly employed in the fabrication of state-of-the-art Si or GaAs IMPATT diodes and GaAs TUNNETT diodes for millimeter- and submillimeter-wave frequencies. The fabrication steps of the process for InP Gunn devices are based on the same aforementioned selective etchants and are similar to those
of Fig. 2, except that no thick electroplated metal layers are used on the heat sink side.\textsuperscript{16,17} Processes with selective etchants are also routinely used in the fabrication of Si or GaAs IMPATT diodes\textsuperscript{10,18,19} and GaAs TUNNETT diodes\textsuperscript{20,21} on either integral or diamond heat sinks.

![Diagram of Fig. 2 with steps in the fabrication of InP TEDs on integral heat sinks.](image)

The schematic of Fig. 3 illustrates the most common waveguide configuration employed in millimeter-wave oscillators with two-terminal NDR devices. Other circuits and their modifications are discussed elsewhere.\textsuperscript{2,10,17,22,23} A disk on top of the device package, \textit{i.e.}, the resonant cap, couples the RF power from the device into the waveguide and coarsely determines the oscillation frequency. The position of the back short is tuned for maximum output power, but in the fundamental mode, also changes the oscillation frequency.\textsuperscript{16} The configuration for devices operating in a second-harmonic mode is quite similar except that the post (pin) above the resonant cap plays a much more active role in the circuit. At the fundamental frequency, the post inductance forms a resonant circuit with the device capacitance and the fringe capacitance of the cap as well as the package parasitics, if present. The fundamental frequency is chosen to be below the cut-off frequency of the waveguide. Therefore, this signal cannot propagate in the waveguide, which creates a reactive termination for the device at the fundamental frequency. The resonant cap couples the RF power at the second-harmonic.
frequency into the waveguide, and the back short is tuned for maximum output power only, but, typically, has very little effect on the oscillation frequency. A well-known modification of this circuit makes the post mechanically tunable resulting in second-harmonic power extraction over a wide frequency range.24-26

![Schematic of a typical waveguide circuit for a millimeter-wave Gunn device oscillator.](image)

Fig. 3. Schematic of a typical waveguide circuit for a millimeter-wave Gunn device oscillator.

4. Basic properties of two-terminal solid-state NDR devices

Three basic mechanisms of NDR device operation, i.e., tunneling, transit time, and intervalley electron transfer, are considered here. They are responsible for the generation of NDR in the devices cited in the Introduction and determine the main device properties.

4.1 Tunneling devices

In this category, Esaki tunnel diodes (TDs) and resonant tunneling diodes (RTDs) are considered. TUNNETT diodes are not included here because transit-time effects do not play a dominant role in this group of tunneling devices. TDs and RTDs both exhibit a negative differential resistance in their current-voltage characteristics ranging from dc to very high frequencies. Esaki TDs were proposed in 19581 and were utilized in oscillators up to 100 GHz. However, the RF output power from these devices is mainly limited by the small RF voltage swing as well as the large junction capacitance and was much smaller than that from other two-terminal devices. Therefore, they are not in much use at this time. More recently, an RTD was proposed,3 which has a much better speed index, i.e., current-capacitance ratio I/C, and can be tailored to optimize the important parameters through heterostructure engineering. In this section, the focus is on RTDs, but it should be noted that similar characteristics are obtained from TDs. In TDs, tunneling takes place from the valence band to the conduction band of a heavily doped p'-n junction and thus is referred to as interband tunneling. However, in RTDs, tunneling takes place through the conduction bands of a double-barrier heterostructure as shown in Fig. 4. The basic mechanism for the negative-resistance property is also illustrated in the same figure. As is well known, tunneling electrons maintain their energy and must have available states to tunnel through. A typical diode structure and its current-voltage characteristic are shown in Figs. 5 and 6, respectively. The structure of
Fig. 5 utilizes AlAs barriers and GaAs quantum wells. Other material systems have also been employed and yielded devices with operation at much higher frequencies than those from the GaAs/AlAs material system.

In contrast to the other two-terminal devices, the negative resistance of an RTD persists over a very wide frequency range and extends down to dc. This causes issues with bias circuit oscillations and, therefore, the bias circuit parameters must be included in an analysis of power generation capability. The equivalent oscillator circuit for such an analysis is shown in Fig. 7.
Fig. 6. Theoretical and measured current-voltage characteristics for the RTD of Fig. 5.\textsuperscript{27}

An approximate analysis of the power generation capability of this device is based on a linearized current-voltage characteristic as shown in Fig. 8. The device is assumed to be biased in the middle of the NDR region, \textit{i.e.}, at \( V_{DC} = \frac{1}{2}(V_{p} + V_{v}) \) and \( I_{DC} = \frac{A}{2}(J_{p} + J_{v}) \).

The RF voltage swing is assumed to be limited to this NDR region, \textit{i.e.},

\[ V_{RF} = \frac{1}{2}(V_{v} - V_{p}) \]  \hspace{1cm} (14)

The negative conductance and the susceptibility of the RTD per unit area \( G_{D} \) and \( B_{D} \) are then given by

\[ G_{D} = \frac{J_{v} - J_{p}}{V_{v} - V_{p}} \]  \hspace{1cm} (15)

and

\[ B_{D} = \omega C_{D} = \omega \frac{\varepsilon_{s}}{W} \]  \hspace{1cm} (16)

Fig. 7. Simplified equivalent circuit of an oscillator with a resonant-tunneling diode connected to a bias circuit and a load.
respectively, where \( C_D \) is the capacitance per unit area, \( \varepsilon_s \), the dielectric constant, and \( W \), the depletion-layer width.

![Fig. 8. Linearized current-voltage characteristic of a resonant-tunneling diode.](image)

The diode area \( A \) and the RF power generated in the device are obtained from Eq. 9 and the matching condition for oscillation and are given by

\[
P_{\text{RF}}(\text{Gen.}) = \frac{1}{8} A \left( J_p - J_v \right) \left( V_v - V_p \right).
\]

The corresponding dc-to-RF conversion efficiency \( \eta \) is given by

\[
\eta = \frac{P_{\text{RF}}(\text{Gen.})}{P_{\text{DC}}} = \frac{1}{2} \left( \frac{J_p}{J_v} + 1 \right) \left( \frac{V_v}{V_p} + 1 \right) \left( \frac{V_v}{V_p} + 1 \right).
\]

As the operating frequency increases, \( C_D \) begins to affect the circuit-limited RF power of Eq. 13, which is now given by

\[
P_{\text{RF}}(\text{Circuit limited}) = P_{\text{RF}}(R_L) = \frac{1}{8} \left( \frac{J_p}{\omega_0 C_D} \right)^2 \left( 1 - \frac{J_v}{J_p} \right)^2 \frac{R_L}{(R_L + R_s)^2}.
\]

However, since the NDR in this device extends to dc, another limitation arises when bias circuit instabilities are to be avoided. This limitation is related to the bias circuit inductance \( L_s \) and can be expressed as

\[
\frac{L_s \left( A^2 G_D^2 \right)}{(AC_D)} < -R_s (AG_D) < 1
\]
where \( AR_c = \rho_s \) is assumed to be dominated by the specific contact resistance and, therefore, is approximately constant. In this case, the bias-circuit limited RF power is given by

\[
P_{RF}(\text{Bias circuit limited}) = P_{RF}(R_L) = -\frac{1}{2} A \left[ G_D + \rho_s \left( G_D^2 + B_D^2 \right) \right] V_{RF}^2.
\]  

(22)

For the experimental device of Fig. 6, reasonable estimates are \((V_s - V_p) = 0.5 \text{ V}, J_p = 40 \text{ kA/cm}^2, (J_p/J_s) = 3.5, \) and \( W = 70 \text{ nm}. \) The output power from such an RTD can then be estimated\(^{10}\) for both limits and the results are shown in Fig. 9. As can be seen from this figure, Eq. 22 presents a much more severe limitation on the power generation capability of RTDs for a finite value of \( L_s \) than Eq. 19.

Figure 10 summarizes the best experimental results obtained to date from RTDs and superlattice electronic devices (SLEDs) in several material systems,\(^{27,29-31,34}\) but a more comprehensive comparison can be found elsewhere.\(^{35}\) It can be seen from this data that the oscillation frequency of 712 GHz is the highest fundamental frequency achieved to date from a solid-state RF source. However, the measured RF power level was approximately 0.3 \( \mu \text{W}. \)

4.2 Transferred-electron devices

These devices utilize basic transport properties in bulk materials to generate the negative resistance. They are unipolar devices where no \( p-n \) junction is required as compared to most of the other two-terminal devices presented here. These devices exhibit low-noise performance and are well suited for local oscillator applications. They require materials with a particular band structure, which is found in several semiconductor materials, particularly III-V compounds. For a material to be suitable, it must possess the following qualities:

i. It must have at least two valleys in the conduction band.

ii. The minimum of the upper valley must be at least several \( kT \) above the minimum of the lowest or main valley in the conduction band at the lattice temperature \( T. \)

iii. The energy difference between the minimum of the upper valley and that of the lower valley must be less than the energy gap \( E_g \) in order to avoid avalanche breakdown.

iv. The transfer of electrons between the valleys must take place in a time that is much less than the period of the operating frequency.
Fig. 9. Top curves: predicted RF output power and diode area for matching into $R_o = 1 \Omega$. Bottom curves: predicted RF output power and diode area for obtaining stability with $L_s = 0.1 \text{nH}$. Linearized current-voltage characteristics for the diode of Fig. 6 are assumed.

Fig. 10. State-of-the-art RF power levels from resonant-tunneling diodes and superlattice electronic devices in the frequency range 30–1000 GHz.
v. The effective mass of electrons in the upper valley must be much higher than that in the main valley and thus their mobility in the upper valley is much lower than in the main lower valley.

Figure 11 shows a simplified conduction band structure of such a semiconductor material system and Fig. 12 shows the resulting velocity-electric field characteristic in bulk material. The region of negative differential mobility $\mu_d < 0$ in this characteristic is responsible for the negative differential resistance of the device and thus RF power generation. As can be seen from these two figures, most of the electrons reside in the lower valley and have a high mobility $\mu_1$ when the electric field $E$ is less than the so-called threshold field $E_{th}$. When $E > E_{th}$, electrons in the main valley gain enough energy $\mathcal{E}$ to transfer to the upper valley where the mobility is lower and thus the velocity $v$ decreases. This continues until most electrons have transferred to the upper valley for $E \gg E_{th}$. The velocity $v$ then starts increasing again but under a lower mobility $\mu_2$.

Fig. 11. Simplified energy-band diagram for a direct two-valley semiconductor showing electron transfer for (a) $E < E_{th}$, (b) $E > E_{th}$, and (c) $E \gg E_{th}$.

Fig. 12. Velocity-electric-field profile for the two-valley semiconductor of Fig. 11.

Current oscillations in GaAs and InP were first observed by Gunn$^{5,36}$ and were subsequently explained by the transferred-electron effect.$^{4,37}$ Several modes of operation of TEDs exist$^{2,38}$ and are caused by the negative differential mobility in the active region of the device. They strongly depend on the length and doping profile of the active region and result in notably different device properties. The operating frequency $f_{op}$ in a near transit-time mode is approximately given by
\[ f_{op} = \frac{v_T}{l}, \]  

(23)

where \( v_T \) is the effective transit velocity and \( l \) is the length of the device.\(^{10}\) However, the operating frequency varies from that given by Eq. 23 depending on the mode of operation.

The bulk negative differential mobility alone does not cause a negative differential resistance at low frequencies as seen previously in RTDs. However, it does result in a dynamic negative resistance at frequencies around \( f_{op} \) as shown elsewhere.\(^{10}\) Among the many semiconductor materials that exhibit the transferred-electron effect, only two materials, namely GaAs and InP, have received the most attention so far and resulted in the best device performance. Because the energy relaxation times in InP are smaller than in GaAs, InP TEDs have yielded excellent performance up to much higher frequencies than GaAs TEDs.\(^{16,17,22,39,40}\) This performance difference can be seen from Fig. 13, which shows the state of the art of GaAs and InP devices in the frequency range 30–400 GHz, but a more comprehensive comparison can be found elsewhere.\(^{22,35}\) Exemplary results of RF output power from InP Gunn devices are 34 mW\(^{40}\) and 1.1 mW\(^{17}\) at oscillation frequencies of 195 GHz and 315 GHz, respectively. The latter experimental result represents the most powerful fundamental solid-state RF source operated at room temperature.

4.3 Transit-time diodes

This group includes several very important devices whose operation depends on a particular current injection mechanism into and the transit time through the active region to create the proper phase relationship between the terminal RF voltage and current. This phase relationship creates a dynamic NDR, which results in RF power generation. Transit-time diodes have many common properties and their basic principle of operation can be explained by reference to Fig. 14. In such devices, carriers are injected into a depleted region of a total width \( W \) and they drift toward the charge collector region with the drift velocity \( v_D \), which depends on the electric field \( E \) in this region. Several mechanisms can be employed to generate and inject carriers. These include:

(i) Thermionic emission over a barrier: Such a barrier can be formed by a \( p-n \) or Schottky junction in forward bias or by a heterojunction of a layer with a wider bandgap than in the neutral and drift regions. This would result in a BARITT diode.\(^{8,9}\)

(ii) Tunneling through a barrier: Electron tunneling takes place in a heavily doped \( p^*-n^* \) junction under reverse bias. It can also take place through a heterojunction barrier and resonant tunneling through a double barrier. This would result in a TUNNETT diode\(^7\) or a quantum well injection transit-time (QWITT) device.\(^{41}\)

(iii) Avalanche multiplication through impact ionization: At high electric fields in a reverse-biased \( p-n \) junction, electrons and holes gain enough energy to create additional carriers through impact ionization from the valence to the conduction
band. This would result in carrier injection by avalanche breakdown and an IMPATT diode.\textsuperscript{6}

(iv) At very high frequencies where very narrow regions of carrier generation exist, both tunneling and impact ionization mechanisms are present and thus a mixed mode results. This would yield a MITATT (mixed tunneling-avalanche transit-time) diode.\textsuperscript{42,43}

![Graph showing RF power vs. frequency for GaAs and InP Gunn devices](image)

**Fig. 13.** Published state-of-the-art results from GaAs and InP Gunn devices under CW operation in the frequency range 30–400 GHz. Numbers next to the symbols denote dc-to-RF conversion efficiencies in percent.

The pulse of charge $Q$, which is injected into the drift region at location $W_c$, drifts under a high electric field at a drift velocity $v_Q$ and induces a current in the external circuit connected to the diode. The induced current density $J_{\text{ind}}$ in the external circuit is given by the Ramo-Shockley theorem,\textsuperscript{5}

$$J_{\text{ind}} = \frac{Q}{W} \left( v_Q - \frac{W}{W_c} \right) \frac{dW}{dt}.$$  \hspace{1cm} (24)

Under ideal conditions, the diode is always punched-through, \textit{i.e.}, $W$ stays constant, and the electric field $E$ is usually high enough so that the carrier velocity is saturated. Under these conditions, $v_Q = v_s$ and $(dW/dt) = 0$ and Eq. 24 reduces to
These properties and conditions are suitable for a simple and approximate large-signal analysis to be used, which determines the basic power generation capabilities of these devices. Under such large-signal conditions, the carriers are assumed to be injected as a sharp pulse into the drift region and to travel at a saturated velocity $v_s$. Voltage and current waveforms, as shown in Fig. 15, then result for all the aforementioned transit-time diodes.

$$J_{\text{ind}} = \frac{Q}{W} v_s.$$  \hspace{1cm} (25)

The voltage $V_T$ across the diode is given by
\[ V_T = V_{DC} + V_{RF} \sin \omega t, \] (26)

where \( V_{DC} \) and \( V_{RF} \) are the dc voltage and magnitude of the RF voltage, respectively. The current pulse is injected at phase angle \( \Theta_M \) with an effective width \( \Theta_W \). The induced current is represented by the peak current density \( J_{\text{max}} \) and the transit angle in the drift region \( \Theta_D = \omega W/v_x \). The properties of each of the different transit-time diodes under consideration are determined by \( \Theta_M, \Theta_W, \) and \( \Theta_D \). The RF power generated in such devices with area \( A \) is given by

\[ P_{RF} = \frac{A}{2\pi} \int J_{\text{incl}}(\omega t) V_{RF} \sin \omega t \, d(\omega t), \] (27)

which simplifies to

\[ P_{RF} = A V_{RF} J_{DC} \sin \frac{\Theta_W}{2} \cos(\Theta_M + \Theta_D) - \cos \Theta_M. \] (28)

The bias current density \( J_{DC} \) is given by

\[ J_{DC} = \frac{1}{2\pi} \int J_{\text{incl}}(\omega t) d(\omega t) = \frac{J_{\text{max}}}{2\pi} \Theta_D. \] (29)

Therefore, the dc-to-RF conversion efficiency \( \eta \) is

\[ \eta = \frac{P_{RF}}{P_{DC}} = \frac{V_{RF} \sin \frac{\Theta_W}{2} \cos(\Theta_M + \Theta_D) - \cos \Theta_M}{V_{DC} \frac{\Theta_W}{2} \Theta_D}. \] (30)

It is clear from the preceding equations that it is desirable to have \( \Theta_W \) as small as possible. Under the ideal sharp-pulse approximation, the assumption of \( \Theta_W = 0 \) is justified and the properties of the various transit-time diodes can be determined as follows:

\[ a. \quad \text{IMPATT Mode} \]

In this mode of operation, \( \Theta_M = \pi \) and Equation 30 reduces to

\[ \eta = \frac{V_{RF} \left( 1 - \cos \Theta_D \right)}{V_{DC} \Theta_D}. \] (31)

For \( \Theta_M = \pi, \) \( \eta = (2V_{RF}/\pi V_{DC}), \) and the maximum occurs at \( \Theta_D = 0.74 \pi \) where \( \eta = (2.27V_{RF}/\pi V_{DC}). \)
IMPATT diodes are very efficient relative to the other devices and generate very high power. This is because the ratio $V_{RF}/V_{DC}$ can approach 60% in materials such as GaAs and InP, and in turn efficiencies can approach 40% under ideal conditions. Also, the current densities in IMPATT diodes are very high and thus the power output is also high. IMPATT diodes, therefore, are the most powerful solid-state devices available at millimeter-wave frequencies. However, they are considered to be quite noisy because of the avalanche generation mechanism. The noise can be reduced by injection locking or by introducing a tunneling component to the injection mechanism as discussed next.

b. **TUNNETT and BARITT Mode**

In these modes of operation, $\Theta_M \equiv (\pi/2)$ and $\eta$ becomes

$$\eta = \frac{V_{RF} \sin \Theta_D}{V_{DC} \Theta_D} . \quad (32)$$

In this case $\eta$ is close to maximum at $\Theta_D = (3\pi/2)$ where $\eta = (2/3\pi)(V_{RF}/V_{DC})$. As can be seen from these expressions, the efficiency of TUNNETT and BARITT diodes is approximately one third of an IMPATT diode under ideal conditions. This is because there is an induced current during the positive half of the RF cycle extending from $\Theta = (\pi/2)$ to $\pi$ and thus the device absorbs power during this phase. However, $V_{RF}$ is higher and the capacitance lower. Both have a positive effect on power output. In addition, because of the carrier generation process, these devices show excellent noise performance, which is comparable to TEDs.$^{10,20,21}$

c. **MITATT Mode**

In this mode of operation, $\Theta_M$ varies from $(\pi/2)$ to $\pi$ depending on the ratio of tunneling to avalanche generation and therefore the efficiency and power output varies between the modes of a pure TUNNETT diode and a pure IMPATT diode.

d. **QWITT Mode**

In this mode of operation, $\Theta_M \equiv (3\pi/2)$ and $\eta$ becomes

$$\eta = \frac{V_{RF} \sin \Theta_D}{V_{DC} \Theta_D} . \quad (33)$$

For $\Theta_D = (\pi/2)$, $\eta = (2V_{RF}/\pi V_{DC})$, which is the same expression as for an IMPATT diode at $\Theta_D = \pi$. However, $V_{DC}$ and $V_{RF}$ are much smaller than those of an IMPATT diode, and, therefore, power generation is smaller. Also, as discussed in Section 4.1 for
RTDs, such a device is more difficult to stabilize because its NDR extends to dc, which limits the power generation capability further.

4.4 Device structures for transit-time diodes

The basic device structure for IMPATT, MITATT, and TUNNETT diodes is shown in Fig. 16 where $E_c$ is the critical field for breakdown. The generation region width $x_A$ can be controlled by the width of the $i$ layer between the $p^{++}$ and $n^*$ layers, and this results in the following different modes of operation:

(i) For $x_A > 100$ nm and $E_c x_A = 1$, avalanche breakdown occurs, which results in an IMPATT diode. For $x_A < 50$ nm and $E_c > 10^6$ V/cm, tunneling dominates, which results in a TUNNETT diode. For $50$ nm < $x_A < 100$ nm, both tunneling and impact ionization contribute significantly to the carrier generation and injection, which results in a MITATT diode.

(ii) As the frequency of operation increases, the depletion layer width shrinks and it becomes more difficult to control the width of $x_A$. At extremely high frequencies, the relationship of $E_c x_A = 1$ becomes very difficult to satisfy and thus would be difficult to operate the transit-time diode in the pure IMPATT mode. In this case, the TUNNETT mode dominates. It is therefore expected that TUNNETT diodes are more suitable than IMPATT diodes for frequencies approaching the THz region.

(iii) From the approximate waveforms of Fig. 15, it is relatively straightforward to estimate the RF output power and dc-to-RF conversion efficiency of these diodes after a so-called single-drift structure as shown in Fig. 16(a) is chosen. This is beyond the realm of this review and the reader is referred to several references.

IMPATT diodes in particular can also be implemented in so-called double-drift structures as shown in Fig. 16(b). Here, the generation region is in the middle and both electron and hole drift regions are present, but the basic operation is the same. Double-drift IMPATT diodes ideally generate approximately four times the RF output power of single-drift IMPATT diodes for the following reasons: The distance between the $p^{++}$ and $n^{++}$ contact regions of the diode is approximately double thereby halving the junction capacitance per unit area and doubling the breakdown voltage. At half the capacitance, the diode area can be made twice as large for the same impedance level. At the same bias current density $J_{DC}$ the bias current is then double and, for twice the bias voltage, the dc input power is four times as large. For the same dc-to-RF conversion efficiency, four times the RF power is generated. In reality however, the total thermal resistance of a double-drift structure is higher than that of a single-drift structure of the same area, and, therefore, a lower bias current density $J_{DC}$ must be chosen. As a consequence, approximately twice the power is actually generated in cw operation, which is still a substantial improvement over a single-drift structure.
The basic device structure and electric field profile for a BARITT device at the operating point are shown in Fig. 17. In a small region $W_s$ around the forward-bias injection point, the electric field $E$ is still below the value $E_s$, which is necessary for carriers to reach the saturation velocity $v_s$. The basic carrier drift and diffusion mechanisms that are involved in this region limit the ratio of $V_{RF}/V_{DC}$ to small values. Therefore, achievable dc-to-RF conversion efficiencies are much smaller than the ideal value from the maximum of Eq. 32. Again, RF output power and dc-to-RF conversion efficiency can be estimated after the structure is chosen and the reader is referred to several references on this subject.\textsuperscript{47-50}

![Fig. 16. Schematic layer sequence and electric field profile for (a) a single-drift transit-time diode and (b) a double-drift transit-time diode.](image)

![Fig. 17. Basic device structure and electric field profile for a BARITT diode.](image)

The state-of-the-art experimental results of these transit-time diodes in cw mode are shown in Fig. 18. Of course, significantly more power can be generated under pulsed conditions where thermal considerations are significantly relaxed. Si IMPATT diodes
yielded the highest RF power levels from any two-terminal NDR device in the frequency range 60–300 GHz.\textsuperscript{35}

Exemplary results of RF output power are 50 mW,\textsuperscript{51} 7.5 mW,\textsuperscript{52} and 1.2 mW\textsuperscript{52} at oscillation frequencies of 245 GHz, 285 GHz, and 301 GHz, respectively. In a cavity cooled to a temperature of 77 K, they were operated at the second highest oscillation frequencies of up to 430 GHz\textsuperscript{53} as a comparison with the results from RTDs in Fig. 10 shows. InP Gunn devices, however, generated the highest RF power levels from any fundamental solid-state RF source operated at room temperature above 300 GHz as a comparison with the results in Fig. 13 indicates.\textsuperscript{35}

Transit-time diodes so far have been realized mostly in Si and GaAs. Other semiconductor materials, such as, InP\textsuperscript{54} and more recent ones such as SiC\textsuperscript{55} and GaN, which have a larger $E_c$, may be capable of generating significantly higher RF power levels. As is well known, the critical field for breakdown $E_c$ and the saturated velocity $v_s$ are the basic material parameters that determine the RF power generation capability of transit-time diodes. In addition, the dc-to-RF conversion efficiency $\eta$ is also important and depends on the mode of operation and other basic material properties such as low field mobility and ohmic contacts. Therefore,
The figure of merit \((E_v v_s)^2\) is given in Table 1 for several semiconductor material systems. As can be seen from this table, some materials with a very high figure of merit have a potential that has not been tapped yet. Ultimately, of course, large \(E_v\) and \(v_s\) occur in vacuum and thus tunnel transit-time devices where electrons tunnel from cold cathodes and travel ballistically in the vacuum of the drift region may be very appropriate for THz power generation.

Table 1. Properties of Important Semiconductors for Power Generation.

<table>
<thead>
<tr>
<th></th>
<th>Si</th>
<th>GaAs</th>
<th>InP</th>
<th>6H-SiC (4H-SiC)</th>
<th>GaN</th>
<th>Diamond</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bandgap at 300 K (eV)</td>
<td>1.12</td>
<td>1.42</td>
<td>1.34</td>
<td>3.06 (3.26)</td>
<td>3.39</td>
<td>5.5</td>
</tr>
<tr>
<td>Electron mobility at 300 K (cm²/Vs)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>⊥ c-axis</td>
<td>1400</td>
<td>8500</td>
<td>4600</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>c-axis</td>
<td>400 (850)</td>
<td></td>
<td>80 (1020)</td>
<td></td>
</tr>
<tr>
<td>Hole mobility at 300 K (cm²/Vs)</td>
<td>450</td>
<td>400</td>
<td>140</td>
<td>90 (115)</td>
<td>150</td>
<td>1600</td>
</tr>
<tr>
<td>Breakdown field at (N_D \sim 10^{17} \text{ cm}^{-3}) (10⁶ V/cm)</td>
<td>0.61</td>
<td>0.65</td>
<td>0.75</td>
<td>2.5 (2.2)</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Thermal conductivity (W/cmK)</td>
<td>1.25</td>
<td>0.46</td>
<td>0.68</td>
<td>4.9</td>
<td>1.3</td>
<td>20</td>
</tr>
<tr>
<td>Sat. electron drift vel. at (E &gt; 5 \times 10^7 \text{ V/cm}) (10⁷ cm/s)</td>
<td>1</td>
<td>0.6</td>
<td>0.75</td>
<td>2</td>
<td>2.7</td>
<td>2.7</td>
</tr>
<tr>
<td>Dielectric Constant</td>
<td>11.8</td>
<td>12.8</td>
<td>12.6</td>
<td>9.7</td>
<td>9</td>
<td>5.5</td>
</tr>
<tr>
<td>Electronic (P_{RF}) Figure of Merit, relative to Si ((E_v v_s)^2)</td>
<td>1</td>
<td>0.4</td>
<td>0.9</td>
<td>70</td>
<td>80</td>
<td>2000</td>
</tr>
</tbody>
</table>

5. Noise performance of solid-state two-terminal NDR devices

Sensing applications generally require sources with excellent noise performance. Two-terminal NDR devices, such as RTDs, Gunn devices, TUNNETT diodes (solid-state or vacuum), and BARITT diodes, exhibit suitable properties, whereas IMPATT diodes are often considered as too noisy because impact ionization is their sole carrier generation mechanism. In general, the FM noise measure is the best figure of merit to compare the noise performance of free-running oscillators with different two-terminal NDR devices. At frequencies above D-band (110–170 GHz), however, these devices frequently do not operate in the fundamental mode, but in a higher-harmonic mode where the FM noise measure is much more difficult or even impossible to determine. As a consequence, Table 2 shows a comparison of the phase
noise as measured for free-running oscillators with different two-terminal NDR devices at select frequencies off the carrier between 100 kHz and 1 MHz. If the phase noise of an oscillator as measured in dBc/Hz indicates excellent noise properties for operation in the fundamental mode and increases by 20 log(n) or less for operation at the n\textsuperscript{th} harmonic and for the same off-carrier frequency, such a device is considered an excellent candidate for a low-noise RF source in a higher harmonic mode of operation.

Table 2. Phase noise of free-running oscillators using low-noise GaAs or InP millimeter-wave and submillimeter-wave two-terminal devices.

<table>
<thead>
<tr>
<th>Device</th>
<th>Phase Noise (dBc/Hz)</th>
<th>Off-Carrier Frequency (kHz)</th>
<th>Oscillation Frequency (GHz)</th>
<th>RF Output Power (mW)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>GaAs Gunn</td>
<td>&lt;- 80*</td>
<td>100</td>
<td>77</td>
<td>&gt;40*</td>
<td>13</td>
</tr>
<tr>
<td>GaAs Gunn</td>
<td>-70</td>
<td>100</td>
<td>80</td>
<td>55</td>
<td>58</td>
</tr>
<tr>
<td>GaAs Gunn</td>
<td>-100</td>
<td>1000</td>
<td>80</td>
<td>55</td>
<td>58</td>
</tr>
<tr>
<td>GaAs Gunn</td>
<td>-120</td>
<td>10000</td>
<td>80</td>
<td>55</td>
<td>58</td>
</tr>
<tr>
<td>InP Gunn</td>
<td>&lt;- 107</td>
<td>500</td>
<td>81</td>
<td>126</td>
<td>40</td>
</tr>
<tr>
<td>GaAs Gunn</td>
<td>-80</td>
<td>100</td>
<td>94</td>
<td>10</td>
<td>59</td>
</tr>
<tr>
<td>GaAs Gunn</td>
<td>-105</td>
<td>1000</td>
<td>94</td>
<td>10</td>
<td>59</td>
</tr>
<tr>
<td>InP Gunn</td>
<td>&lt;- 75</td>
<td>100</td>
<td>94</td>
<td>20</td>
<td>59</td>
</tr>
<tr>
<td>InP Gunn</td>
<td>-100</td>
<td>1000</td>
<td>94</td>
<td>20</td>
<td>59</td>
</tr>
<tr>
<td>InP Gunn</td>
<td>&lt;- 110</td>
<td>500</td>
<td>103</td>
<td>180</td>
<td>39</td>
</tr>
<tr>
<td>GaAs TUNNETT</td>
<td>&lt;- 94</td>
<td>500</td>
<td>108</td>
<td>40</td>
<td>20</td>
</tr>
<tr>
<td>GaAs TUNNETT</td>
<td>&lt;- 101</td>
<td>1000</td>
<td>108</td>
<td>40</td>
<td>20</td>
</tr>
<tr>
<td>InP Gunn</td>
<td>&lt;- 108</td>
<td>500</td>
<td>132</td>
<td>120</td>
<td>16</td>
</tr>
<tr>
<td>InP Gunn</td>
<td>&lt;- 103</td>
<td>500</td>
<td>151</td>
<td>58</td>
<td>16</td>
</tr>
<tr>
<td>InP Gunn</td>
<td>&lt;-94</td>
<td>500</td>
<td>199</td>
<td>19</td>
<td>60</td>
</tr>
<tr>
<td>GaAs TUNNETT</td>
<td>&lt;-94</td>
<td>500</td>
<td>209</td>
<td>9</td>
<td>21</td>
</tr>
</tbody>
</table>

* Reported as typical value, corresponding RF output power not mentioned.

6. Vacuum TUNNETT devices

As was discussed in Section 4.4, the RF output power and maximum operating frequency of conventional solid-state transit-time diodes are limited by the breakdown electric field $E_c$, the dielectric constant $\varepsilon_s$ and the saturated velocity $v_s$ that exist in all semiconductor materials. Also, because the dc-to-RF conversion efficiencies decrease with frequency, it becomes more and more difficult to remove the heat from such diodes.

IMPATT diodes have exhibited the greatest RF output power at millimeter-wave frequencies and generally have good efficiencies. However, at terahertz frequencies, the
Two-Terminal Active Devices for Terahertz Sources 417

avalanche region becomes very narrow and the total device length very short. Therefore, tunneling becomes the dominant injection mechanism, which reduces the efficiency. Since tunneling is a very fast process, the operating frequency of a TUNNETT diode can be extended to higher values than that of a pure IMPATT diode. As shown in Section 4.3, tunneling causes electrons to be injected into the drift region in phase with the RF voltage rather than at 90° out of phase as in avalanche-breakdown injection. If electrons are assumed to be injected at a saturated velocity $v_s$, the theoretical dc-to-RF conversion efficiency of a TUNNETT diode is approximately one third that of an IMPATT diode. TUNNETT diodes also exhibit much better noise performance than IMPATT diodes and thus are excellent candidates for local oscillator applications at terahertz frequencies. As in IMPATT diodes however, $v_s$, $E_c$, and $\varepsilon_s$ of the semiconductor material again limit the RF output power. A device structure has been proposed\textsuperscript{56} that also uses tunneling as the electron injection mechanism, but circumvents the aforementioned limitations by using ballistic transport in vacuum. This device structure is a more ideal TUNNETT diode, which is capable of generating significant amounts of RF power and can be tuned over a wide frequency range. This is due to the following reasons:

(i) In such a device, electrons are injected with low velocity and their velocity increases as they approach the anode resulting in a low induced current during the positive half of the RF cycle and high induced current in the negative half of the RF cycle. Therefore, the overall induced current waveform resembles more that of an IMPATT diode rather than that of a semiconductor TUNNETT diode, which yields higher dc-to-RF conversion efficiency.

(ii) The overall velocity is controlled by the applied bias voltage $V_{DC}$. This eliminates the limitations from a saturated velocity $v_s$. Therefore, the device is capable of generating much higher RF output power and it is also easier to tune the device electronically (by changing $V_{DC}$).

(iii) Since the drift region of this device is in vacuum, the critical field for breakdown is essentially very high (infinite in a perfect vacuum). This eliminates the power generation limit from the breakdown electric field $E_c$.

(iv) The dielectric constant $\varepsilon_s$ in vacuum is unity. This results in a reduced device capacitance per unit area and thus a higher impedance level per unit area for matching. A larger-area device can then be matched to an external circuit, which results in a higher power generation capability.

(v) Last, and by no means least, the heat in such a device is generated mainly at the anode, which is made of metal, and thus the heat can be dissipated more easily than inside a semiconductor device. Therefore, the thermal resistance of such a device is lower than that of a conventional semiconductor device. This is very important because the thermal resistance is usually a major limiting factor, particularly in cw operation.
The series resistance of such a device can be lower compared to a semiconductor device because there is no contact resistance involved at the anode side. This is extremely important, particularly at terahertz frequencies.

Because of the basic operation of the device, it is referred to as a ballistic tunneling transit-time device (BT$^3$D). The proposed device structure is shown schematically in Fig. 19(a) and is composed of a field-emitter array (FEA) coupled to a micromachined waveguide cavity as indicated in Fig. 19(b). Recent advances in etching technologies, micromaching, self-assembly, and carbon nanotubes enable the fabrication of such a device. One of the most important parameters for device performance is the current density that can be achieved from the FEA. For example, arrays were fabricated from well-controlled Si tips in the Solid-State Electronics Laboratory at the University of Michigan.$^{51,52}$ They are shown in the two micrographs of Fig. 20. When such arrays were coated with HfC to lower the work function,$^{53}$ current densities of 2,000–4,000 A/cm$^2$ per tip were achieved,$^{54}$ which are quite adequate for the proposed device. Another important aspect is the design, fabrication, and realization of the resonant cavity. This may be implemented by utilizing micromachining technology.

![Fig. 19. BT$^3$D (a) structure, (b) waveguide mount](image)

The proposed device has the potential of generating significant amounts of RF power from millimeter-wave frequencies up to terahertz frequencies with good dc-to-RF conversion efficiencies and low-noise performance as the following preliminary theoretical analysis of this device shows.

### 7. Basic properties of the BT$^3$D

The basic device structure is shown in Fig. 19(a) and consists of a field emitter cathode, a drift region, and a collecting anode, all sealed in vacuum. This small structure is mounted across a waveguide as shown in Fig. 19(b). The operation of the device can be explained with the help of Fig. 21. A voltage is applied through a bias circuit that provides RF isolation. The applied voltage combined with the device dimensions sets up an electric field across the drift region. If proper magnitude dc and RF fields are present at the cathode, field emission of electrons occurs at the cathode tips producing an injected current waveform peaked at a phase angle of $\pi/2$ as shown in Fig. 21. These electrons are then accelerated across the drift region and collected at the anode producing the induced or terminal current as discussed in Section 4.3. If the dimensions and the
amount of acceleration are chosen correctly, the terminal current is greater than the average current in the second half of the RF cycle. This leads to a negative conductance with RF power being available from the device. The combination of the drift region length and the bias voltage determine the operating frequency. If a transit time equal to the inverse of the operating frequency is assumed, the required bias voltage vs. drift region length can be found for different operating frequencies and typical results are shown in Fig. 22. The bias voltages \( V_{\text{DC}} \) and device dimensions for operation at THz frequencies are quite compatible with fabrication technologies using micromachining techniques.

![Fig. 20. Scanning electron micrographs of gated Si emitters fabricated by the self-aligned process. (a) Arrays of Si field emitters with high packing density of \( 1.1 \times 10^7 \) tips/cm\(^2\). (b) Sharp Si emitter with a gate-tip spacing of 140 nm.](image)

The transit-time properties of this ballistic device structure are investigated next. The device has the nearly triangular waveform of the induced current shown in Fig. 21. However, a more realistic analysis depends on the injection conditions and accounts for the time varying voltage across the drift region. The properties of the cathode are critical to the device performance. However, the basic form of the cathode current is an exponential dependence on the voltage or the cathode field, with the peak current occurring at the \( \pi/2 \) phase point. A computer simulation program has been developed to obtain the injected and induced currents for a range of RF conditions. The analysis for the BT\(^3\)D is more complex than that for a corresponding semiconductor transit-time diode. In a semiconductor, all the carriers, \( i.e., \) electrons or holes, within the drift region have the same velocity, and the induced current can be obtained by integrating over the charge within the device. For this BT\(^3\) device, the electron acceleration is constant and the velocity depends on the amount of time the electrons spend in the drift region. A double integration over time and injection angle is needed to obtain the induced current. The simulation program predicts the large-signal induced current waveform of the device. The ratio of the Fourier components of the fundamental RF current and the RF voltage gives the device admittance. This, along with the RF voltage, the embedding impedance, and the parasitic series resistance \( R_s \) gives the available RF power delivered to a matched load. This information can then be used to investigate a variety of device
structures, bias conditions, and operating frequency ranges. Some sample calculations are discussed next.

Fig. 21. Generic voltage and current waveforms in a ballistic transit-time device.

Fig. 22. Dimensions and dc bias voltages of a BT$^3$ device for various operating frequencies.

Figure 23 shows the small-signal $G_D$ vs. $B_D$ plot for a BT$^3$ device with a 750-GHz center frequency and bias current densities $J_{DC}$ of 100–300 A/cm$^2$. The solid lines in this figure are lines of constant current density and the dashed lines are lines of constant operating frequency. The small-signal admittance is independent of the RF voltage. This type of plot is useful for predicting matching conditions and estimating the general bandwidth of the device. Figure 23 shows the excellent broadband device performance, predicting an operating bandwidth of more than 500 GHz for a device with a center
frequency of 750 GHz. A second key parameter is the device $Q$ value, which is the ratio of the imaginary part to the real part of the device admittance at the operating point. This $Q$ value is nearly -1 for the highest current density. Such a low $Q$ value allows easier matching to external circuits and reduces the effect of parasitic resistances. The $Q$ value is small because of the longer drift region and smaller dielectric constant when compared with a semiconductor transit-time diode.

Fig. 23. Small-signal admittance $G_D + j B_D$ for a 750-GHz BT$^3$ device at 500 GHz, 750 GHz, and 1 THz.

Fig. 24. Small-signal admittance $G_D + j B_D$ for a 1-THz BT$^3$ device at 750 GHz, 1 THz, and 1.25 THz.
A similar small-signal analysis of a device designed for a center frequency of 1 THz is shown in Fig. 24. This figure shows some of the frequency scaling effects of this device. The combination of a higher frequency and shorter drift region increases the magnitude of the susceptance $|B_D|$ of the device. A larger current density is required for a comparable device operating point.

Although small-signal calculations provide matching and bandwidth information, a large-signal simulation is needed to estimate the RF output power available from a BT$^3$ device. The large-signal $G_D$ vs. $B_D$ plot for a device with a center frequency of 1 THz and at a bias current density $J_{DC}$ of 500 A/cm$^2$ is shown in Fig. 25. In this figure, the solid lines are lines of constant $V_{RF}$ and the dashed lines are again lines of constant operating frequency. Since the device admittance $G_D + j B_D$ is $J_{RF}/V_{RF}$ and $J_{RF}$ is approximately constant for a triangular induced current waveform, the conductance $G_D$ is proportional to $V_{RF}^{-1}$. This explains the variation shown in Fig. 25. The bias current densities $J_{DC}$ are still modest compared with the requirements of semiconductor-based transit-time diodes with modest RF voltage requirements. Even under large-signal conditions, the device offers very wideband low-$Q$ operation. The RF power available from the device also depends on the matching impedance and the parasitic loss.

![Fig. 25. Large-signal admittance $G_D + j B_D$ for a 1-THz BT$^3$ device at $J_{DC} = 500$ A/cm$^2$.](image)

Typical transit-time diodes operating at lower frequencies have load and parasitic resistance values of a few $\Omega$, and similar matching conditions are assumed in the analysis of the proposed BT$^3$ device. The large-signal information from Fig. 25 is used to determine the device area $A$ required for matching to the combination of the load resistance $R_L$ and the parasitic resistance $R_p$. The available RF power vs. RF voltage can be found using this area $A$ as described in Section 2. The RF power available depends on the matching impedance, with lower impedance values yielding more RF power. A parasitic series resistance $R_p$ greatly reduces the available power, both by absorbing a portion of the generated power and by forcing a smaller device area to meet the matching condition for oscillation. Typical current densities require device diameters on the order
of 20–50 μm. Figure 26 shows the available RF power vs. RF voltage at different bias current densities $J_{DC}$ for a device designed for operation at 1 THz. The area of the device $A$ was chosen to yield a small-signal device resistance $R_D$ of $-10 \ \Omega$ at $V_{DC} = 10 \ \text{V}$ and $V_{RF} = 1 \ \text{mV}$ and was kept constant for all values of $V_{RF}$. As $V_{RF}$ increases, $V_{DC}$ decreases slightly for a constant $J_{DC}$. $|R_D|$ also decreases and, therefore, $R_L < 9 \ \Omega$ is required to fulfill the matching condition for $R_s = 1 \ \Omega$. For the curves in Fig. 26, $R_L = 2 \ \Omega$ was assumed to be the smallest possible load resistance.

![Fig. 26. Available RF power from a BT$^3$ device vs. RF voltage $V_{RF}$ for $R_s = 1 \ \Omega$ and a range of current densities at 1 THz.](image)

Tables 3 and 4 provide a summary of additional simulation results for a 500-GHz device and a 1-THz device under different bias conditions. Similar to the simulations for Fig. 26, the device area $A$ was chosen to give $R_D = -10 \ \Omega$ for the device operating at the noted $V_{DC}$ and under the small-signal condition of $V_{RF} = 1 \ \text{mV}$. The RF output power $P_{RF}$ was then determined for the device under large-signal conditions and matched to $R_L = 2 \ \Omega$. In the large-signal simulations, $V_{DC}$ was reduced slightly from the value at $V_{RF} = 1 \ \text{mV}$ to maintain a constant $J_{DC}$. In all cases of Tables 3 and 4, again a series resistance $R_s$ of $1 \ \Omega$ was assumed. The two tables show the expected range of bias voltages $V_{DC}$ bias current densities $J_{DC}$ and generated RF power levels $P_{RF}$.

The simulation results of Fig. 26 as well as Tables 3 and 4 show that it is reasonable to expect RF output power levels in the mW range from BT$^3$ devices. These results also show how critically the device performance depends on the current-injection characteristics of the cathode. Therefore, a key component for terahertz vacuum device research will be a detailed investigation of cold-cathode technology to increase the available current densities. The advantages of higher current densities are clear from these simulation results.
Table 3. Estimated output power (in mW) vs. bias voltage and current density for a BT$^3$D device at 500 GHz with $R_D(V_{RF} = 1 \text{ mV}) = -10 \Omega$, $R_i = 1 \Omega$, and $R_L = 2 \Omega$.

<table>
<thead>
<tr>
<th>$J_{DC} (\text{A/cm}^2)$</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_{DC} (\text{V})$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>12</td>
<td>38</td>
<td>117$^\S$</td>
</tr>
<tr>
<td>30</td>
<td>19.4</td>
<td>66.8</td>
<td>181</td>
</tr>
</tbody>
</table>

$^\S R_L = 2.3 \Omega$

Table 4. Estimated output power (in mW) vs. bias voltage and current density for a BT$^3$D device at 1 THz with $R_D(V_{RF} = 1 \text{ mV}) = -10 \Omega$, $R_i = 1 \Omega$, and $R_L = 2 \Omega$.

<table>
<thead>
<tr>
<th>$J_{DC} (\text{A/cm}^2)$</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_{DC} (\text{V})$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.74</td>
<td>3.14</td>
<td>8.42</td>
</tr>
<tr>
<td>30</td>
<td>1.8</td>
<td>5.1</td>
<td>19</td>
</tr>
</tbody>
</table>

8. Summary and conclusions

The basic properties, power generation capabilities, and state-of-the-art experimental results of two-terminal solid-state devices were presented. Such devices are capable of generating significant amounts of power and remain useful at millimeter and THz frequencies. New materials such as GaN and SiC have the potential of increasing the power output significantly and, ultimately, vacuum based ballistic devices may be needed for generation of significant power levels at THz frequencies. The BT$^3$D presented here may be a good candidate for this purpose.

Acknowledgments

The authors would like to thank Stella Pang from the University of Michigan for her contributions to this work. This work was supported in part by NSF (Grant ECS 98-03781) and ARO under the MURI program on Biological and Chemical Sensing at Terahertz Frequencies (Contract DAAD19-01-1-0622).

References


60. H. Eisele, unpublished results.


64. S. W. Pang, personal communication.
MULTIPLIER AND HARMONIC GENERATOR TECHNOLOGIES FOR TERAHERTZ APPLICATIONS

R. M. WEIKLE, II and T. W. CROWE
Department of Electrical and Computer Engineering,
University of Virginia, Charlottesville, VA 22904-4743, USA

E. L. KOLLBERG
Department of Microelectronics,
Chalmers University of Technology, S-412 96 Göteborg, Sweden

Harmonic generation based on frequency multipliers has proven to be the most successful and widely used solid-state technology for generating power at submillimeter wavelengths. Over the last several years, the development of new device technologies, implementation of innovative circuits, and application of advanced integrated-circuit processing techniques to frequency multiplier design have resulted in unprecedented levels of performance throughout the submillimeter-wave frequency band. This paper reviews the technological innovations, device options, circuit architectures, and fabrication technologies that have made harmonic generation such a successful approach to source development in the submillimeter spectrum.

Keywords: Harmonic generation, frequency multipliers, Schottky diodes, varactors.

1. Introduction

Over the past decade, scientists and engineers have devoted an enormous effort to develop the technology required to make scientific measurements in the millimeter and submillimeter regions of the electromagnetic spectrum [1,2,3,4]. Still, the development and implementation of systems operating above 300 GHz continues to be challenging. This is due, in part, to the difficulties associated with building receivers and sources capable of operating with the needed performance at these frequencies. In recent years significant progress has been made on superconducting detectors, which now offer near quantum-limited performance at submillimeter wavelengths [5,6,7,8]. In addition, Schottky diode receivers operating at room temperature offer an excellent alternative for applications where the expense of cryogenic cooling is not warranted. Thus, the most serious obstacle remaining to the widespread use and study of the submillimeter-wave spectrum is the lack of reliable and tunable solid-state sources for use as local oscillators in heterodyne receivers and as signal sources for spectroscopy, radar, and communications systems.
A range of fundamental oscillators is available in the frequency bands below 300 GHz. These include Gunn oscillators, which work well up to about 150 GHz [9] and have been a workhorse technology for radio astronomy. More recently, systems based on lower frequency synthesizers followed by inexpensive frequency multipliers and a transistor amplifier have become the most versatile frequency source in the 100 GHz region. Amplifiers capable of generating hundreds of milliwatts of power with significant bandwidth at 100 GHz are now available [10], and further improvements to this technology are expected in the near future. Reaching even higher frequencies requires the use of additional frequency multipliers. To be useful for the broadest possible range of applications, these multipliers must achieve a range of performance criteria including high efficiency and power handling, as well as broad tuning bandwidth without the use of mechanical tuners. Additional practical requirements include (i) room temperature operation, (ii) compactness, (iii) high reliability, and (iv) low cost. Also, depending on the application, the trade-offs among the performance criteria can vary greatly. For example, radio astronomers now require LO sources for superconducting receivers that supply about 100 μW of power and can be electronically tuned across a full waveguide bandwidth. Atmospheric researchers, in contrast, generally require power only at fixed frequencies but with several milliwatts of power to drive room temperature Schottky mixers. Unfortunately, no single multiplier technology offers a comprehensive solution for all of these criteria across the full frequency band of interest. Thus, for each application the system engineer must balance a variety of trade-offs and select the technology that offers the best solution.

The goal of this chapter is to review the range of multipliers now available and thereby offer the reader a convenient summary of what is technologically possible and where significant challenges remain for future development. This review begins with a brief overview of multiplier technology. It continues with a review of recent results with frequency doublers and triplers. Higher order multipliers are then considered. Very broadband multipliers based on nonlinear transmission lines (NLTLs) are then discussed. The review ends with a brief discussion of sideband generators, which are a technology used to transform a narrow band frequency source into a tunable source by frequency mixing. The chapter then concludes with a brief summary and some discussion on future directions in multiplier technology.

2. Overview

Although researchers have been studying harmonic generators for decades, it is only in the last several years that the technology has become available to achieve acceptable performance throughout most of the submillimeter-wave frequency band. The most significant advances have come in the areas of fabrication technology and circuit simulation and design. Modern semiconductor processing technology allows the fabrication of fairly complex integrated diode circuits with both exceptional device characteristics and reduced parasitic circuit elements. Several groups are now fabricating III-V semiconductor circuits integrated on quartz substrates [11,12,13] or with only very
thin semiconductor layers within the critical circuit regions [14]. This reduces parasitic capacitances and significantly increases the cut-on frequency of unwanted higher order electromagnetic modes. Also, because all of the critical circuit components are aligned lithographically, the circuit is fabricated exactly as designed. There are no alignment errors caused by the manual soldering of the diode chip to the circuit, for example, and therefore the circuit designer can be much more aggressive in laying out an optimized circuit geometry.

The second major technological advance has been the recent commercial availability of fast computer based simulation tools that allow the designer to precisely model the performance of the entire circuit over broad frequency bands. Among the more prominent of these software tools are Agilent's Advanced Design System (ADS), Ansoft's High Frequency Structure Simulator (HFSS), and Applied Wave Research's Microwave Office (MWO). These simulators allow accurate determination of the optimum embedding impedance for the nonlinear diode as well as the passive circuit layout that achieves this impedance over the broadest bandwidth. Furthermore, these tools are now fast enough to allow the circuit designer to efficiently vary the diode and circuit parameters to achieve the best performance for the given input frequency, power level and desired bandwidth. These tools have fundamentally changed the manner in which multiplier circuits are designed, and have permitted the use of more intricate circuit architectures.

Although these technological advances offer the potential of new levels of performance, they do not in themselves guarantee success. All computer-based simulations are subject to fundamental limitations that designers must consider if they are to achieve realistic designs and optimized circuits. For example, elementary circuit design techniques often neglect the existence and effects of higher order modes. At terahertz frequencies, high-order modes can be excited readily, but are often neglected in simple circuit analyses based on low-frequency circuit techniques. It is up to the designer to be cognizant of this possibility, search for such occurrences, and find practical methods to eliminate them. Furthermore, the simple equivalent-circuit based models of nonlinear diodes are often incomplete and neglect high-frequency effects or other phenomena important in the terahertz regime. Although many groups have proposed methods to simulate the effects of velocity saturation [15], charge carrier inertia, plasma resonance and the skin effect, these simulations can be very computer intensive [16,17,18]. In addition, thermal effects, caused by the power dissipated in the diode can significantly affect both the series resistance and the thermal emission of carriers over barriers [19]. In most cases it is not practical to include all of these effects into the circuit simulation. Rather, the designer must develop clever rules to avoid these problems and then carefully confirm that the final design does not violate the original assumptions.

Optimal terahertz performance requires submicron anode features, micron sized airbridges and substrate removal processes. Although the semiconductor fabrication technology available today is truly remarkable, it is not difficult to design circuits that should yield wonderful performance in theory but are impractical to fabricate. For this reason it is critical that the circuit designers and device fabrication engineers work closely together. Through this collaboration impractical designs can be avoided and the
process complexity can be minimized. This not only improves the performance of the final circuits but also increases yield. Furthermore, it allows circuit fabricators to develop unique processes to achieve the specific circuit features required for optimum performance.

Fig. 1: The evolution of millimeter and submillimeter-wave balanced doublers. (a) The initial version demonstrated by Erickson in 1990 with two whisker contacts and a diode chip mounted on a coaxial pin [21]. (b) A similar design with a planar diode chip replacing the whisker contacts. This allows more diodes and therefore greater power handling as well as easier assembly and greater robustness [22]. (c) A version designed by Porterfield that replaces the coaxial pin with a planar circuit [23]. (d) A recent design by Erickson (fabricated at NASA/JPL). The diodes are now placed in the output waveguide and the semiconductor substrate is removed to improve Terahertz performance [20] (photograph courtesy of N. Erickson).
A good example of the evolution of high frequency multiplier circuit design is depicted in Fig. 1. In 1990 Erickson introduced a balanced doubler circuit that eliminated the need for a distributed filter to separate the input and output frequencies [21]. The key features of the balanced circuit are its symmetry and the orientation of the diodes. These features prevent the second harmonic from coupling to the input waveguide, thus eliminating the physical requirement of a distributed filter. In its initial implementation, the balanced doubler design used two whisker contacts to the same GaAs diode chip, as shown in 1(a). In a subsequent improvement, shown in 1(b), the whiskered diodes were replaced by a planar GaAs chip [22]. This greatly eased circuit assembly and also allowed several diodes to be integrated on each half of the chip, thereby improving power handling ability. Porterfield [23] then replaced the coaxial pin with a lithographically defined planar circuit. This greatly eased the assembly and allowed for better impedance matching and tuning. More recently, the entire circuit, except for the metal block containing the machined waveguides and channels, has been integrated onto a single substrate. This new design uses beam leads to eliminate the need for diode soldering and has allowed researchers to extend this basic circuit concept to terahertz frequencies [14,24]. This steady evolution of the balanced doubler circuit, which is made possible by both the improvements in fabrication technology and computer simulation tools, has led to major advances in performance. As detailed in the following sections, multipliers now generate more power, over greater bandwidths and at higher frequencies, and are far easier to assemble than was considered possible only a few years ago.

3. Frequency Doublers

Frequency doublers are in most regards the simplest multiplier architecture since no idler circuits are required and all unwanted harmonics are above the desired output frequency. Whisker contacted doublers have generated power throughout the 0.1 – 1.0 THz band for many years [25,21,26]. However, these doublers have had many drawbacks, mostly related to the use of whiskered diodes. These include poor power handling and bandwidth, difficult assembly and uncertain reliability. The most successful doublers now use planar diodes and a balanced circuit architecture, as described above. The use of planar and integrated diodes allows many Schottky anodes to share the available input power, thereby greatly increasing power handling. In addition, since power handling is no longer a major problem, the circuit designer can place more emphasis on improving doubling efficiency and bandwidth. Figure 2 shows a recent result with a high power doubler to 55 GHz. An efficiency of approximately 70% has been achieved and the 3dB bandwidth is roughly 10%. This doubler easily handles input power levels up to 1W and uses no tuners of any type. Although efficiency and power are decreased as the frequency increases, this basic circuit design can be scaled to virtually any frequency within the millimeter-wave band. At the other frequency extreme, integrated balanced doublers are now the best all-solid-state sources operating above 1 THz. Erickson has demonstrated a broadband doubler to 1.5 THz for the Herschel spacecraft [27]. This multiplier will be used as an LO source for a hot electron bolometric mixer and the required power level is
on the order of one microwatt. At an operating temperature of 60K initial test results have indicated power levels as high as 45 µW and greater than 1 µW from 1,450 – 1,600 GHz. This circuit requires no mechanical tuners and works in a varistor mode. As this technology is further developed, and as the pumping sources are improved, it is expected that this doubler circuit will generate the required power for the Herschel mixers over several hundred GHz of bandwidth and will be readily space qualified.

4. Frequency Triplers

As the order of frequency multiplication is increased, the circuit becomes much more complex. This is because all of the harmonic frequencies between the input and output frequencies must be terminated appropriately. In the case of a tripler, the input and output frequencies as well as the second harmonic idler frequency (2x input) must be optimized. In practice, this is difficult to achieve, especially when whiskered diodes are used.

![Balanced Doubler, Pin = 700 mW](image)

Fig. 2: The measured performance of a doubler to 55 GHz. A planar diode array is used to achieve high power handling and excellent efficiency across a wide bandwidth without any mechanical tuning [31].

Despite their inherent complexity, whiskered diode triplers have been rather successful and have been useful to above 1 THz [28,21,29]. More recently, these traditional circuit designs have been replaced by planar and integrated diode circuits that use symmetry to suppress even harmonics and thereby eliminate the need for idler tuning. These triplers fall into two general categories, (i) designs using arrays of diodes to create circuit
symmetry and (ii) designs using diodes that naturally yield symmetric capacitance-voltage characteristics, such as the heterostructure barrier varactor.

A. Triplers using Circuit Symmetry

A first example of a planar diode tripler using circuit symmetry has been presented by Erickson. He has used an anti-parallel diode pair with an integrated bias capacitor to achieve excellent broadband performance at about 300 GHz [30]. The efficiency is as high as 9% and useful power is obtained over nearly 100 GHz of bandwidth. This result is particularly significant because the tripler uses no mechanical tuners and relies on an integrated diode process developed at the Jet Propulsion Laboratory (JPL) for NASA applications. Such an integrated diode technology can potentially be scaled to significantly higher frequencies. In fact, JPL has recently fabricated a 1 THz tripler chip, depicted in Fig. 3 [24]. In this type of design the second harmonic power flows only in the diode loop and thus the idler circuit design need only consider that part of the circuit. Since modern design tools allow the circuit to be modeled very precisely and integrated circuit fabrication technology reproduces the circuit design very accurately, excellent performance can be achieved over broad frequency bands.

More recently, a commercial enterprise (Virginia Diodes, Inc.) has begun marketing planar Schottky diode frequency triplers for millimeter and submillimeter wavelengths. These are also broadband circuit designs that require no mechanical tuners. In addition they require no dc bias. Thus, the tripler block has only an input waveguide and an output

Fig. 3: Photograph of the Erickson/JPL 1 THz tripler, showing the beam leads at far left, output waveguide (next to the beam leads), diodes and radial stubs, and the waveguide probe in the reduced height input waveguide (right) (courtesy of N. Erickson).
waveguide. Since there are no other electrical connections to the outside world the component is very reliable and nearly impervious to electrostatic discharge. Figure 4 shows the uniformity of performance that is achieved with this integrated tripler design [31].

B. Triplers using Device Symmetry

A variety of varactor structures exhibiting symmetric capacitance-voltage characteristics have been proposed and demonstrated over the past decade, including the Schottky/2-DEG/Schottky diode [32], the back-to-back barrier-$n$ layer-$n+$ (bbBNN) diode [33], a symmetric Schottky diode pair with inhomogeneous doping [34], and the heterostructure barrier varactor (HBV) diode [35]. Of these devices, the heterostructure barrier varactor, or HBV, has proved to be the most useful and important. When pumped with an AC drive voltage, devices (such as the HBV) having a symmetric or even capacitance-voltage relation produce only odd-order harmonics and are thus natural frequency triplers. In principle, the absence of even-order harmonics simplifies circuit design by eliminating the need for a bias voltage or idler circuits. It has been further suggested that multipliers

![Broadband Frequency Triplers](image)

Fig. 4: The output power of a series of broadband frequency triplers. The input power is about 15 mW and the six units show excellent repeatability for the planar diode circuit design. This circuit design is readily scaled to other frequencies and higher power levels [31].
based on these devices should be capable of achieving significant bandwidths due to the absence of idlers, which tend to be narrowband resonant circuits [36].

C. Heterostructure Barrier Varactor Triplers

The heterostructure barrier varactor is a symmetric layered device consisting of alternating sections of high-bandgap semiconductor barrier layers separated by low-bandgap semiconductor modulation layers, as is depicted in Fig. 5(a). The high-bandgap material provides a barrier that prevents electron transport through the structure. When biased, carriers accumulate on one side of the barrier and are depleted from the opposite side, as illustrated in Fig. 5(b), causing a decrease in capacitance. Usually, HBV structures are grown by either molecular beam epitaxy (MBE) or metalorganic vapor-phase epitaxy (MOVPE). The ability to epitaxially grow multiple barriers in a single device is one of the primary advantages of the HBV.

HBV Material Systems

Among the most critical parameters that govern the performance of an HBV as a frequency multiplier are: (1) capacitance modulation ratio (2) conduction leakage current through the device, (3) series resistance, and (4) breakdown voltage. These device parameters are strongly dependent on the quality and type of the material from which the device is fabricated and a number of different material systems have been proposed and researched since the invention of the HBV. The first devices were fabricated from the GaAs/AlGaAs material system [37,38,39]. In this arrangement, AlGaAs is used as the barrier layer and GaAs acts as the modulation layer. The AlGaAs/GaAs material system has been studied for many years and one of its advantages is a well-developed processing technology. A crucial drawback of this material is the relatively low conduction band offset and resulting high conduction currents. Even as the mole fraction of aluminum increases, the barrier remains low due to indirect scattering of carriers between the conduction band T-valley and X-valley. The conduction currents through these devices increase rapidly with temperature and, consequently, result in very low efficiencies for large pump powers.

An attractive material system for HBV multipliers that overcomes many of the problems with the AlGaAs/GaAs system is that based on In0.53Ga0.47As/In0.52As grown on an InP substrate [40;41]. In0.53Ga0.47As offers a reasonable compromise between mobility and breakdown voltage while giving a large conduction band offset – particularly if an
AlAs layer is grown in the middle of the barrier. High multiplier efficiencies have been reported for HBV's fabricated from this material system and its future prospects are very promising [42].

A number of other material systems have been suggested for millimeter and submillimeter-wave HBV multipliers including InAs/AlSb on InAs and Si/SiO₂ on Si. Each of these systems has potential merits as well as drawbacks. AlSb is slightly mismatched to InAs, but thin layers (~140 Å) can be grown epitaxially and have shown very high conduction band offsets and good current blocking characteristics. Unfortunately, the offset in the valance band is negative and does not block hole current. In addition, the breakdown voltage of InAs is rather low and prototype diodes have shown no capacitance modulation, indicating interface traps between the InAs and AlSb layers.

Si/SiO₂/Si devices are based on bonding two thin silicon wafers together with an interfacial layer of SiO₂. The oxide layer is an effective block for conduction currents and the processing technology for this system is very highly developed. Unfortunately, the relatively low mobility of silicon will probably limit devices made from this material system to the microwave and lower millimeter-wave range [43].

**HBV Circuit Models**

Accurate and reliable device models are imperative for obtaining the best possible performance from any submillimeter-wave device. Over the years, a number of different models have been used to represent HBV devices, all with varying degrees of accuracy. For simple modeling of symmetric varactor devices, some researchers have used odd-order polynomials to represent the device's voltage-charge relation [44]. Dillner used the
fifth-order polynomial given in equation (1) to explore the effect of the capacitance-voltage curve and cutoff frequency (equation 2) on HBV tripler and quintupler conversion efficiency [45].

\[
V(q) = S_{\min} q + \left( S_{\max} - S_{\min} \right) \left( \beta \frac{q^3}{q_{\max}^2} + \gamma \frac{q^4}{q_{\max}^4} \right),
\]

(1)

\[
f_c = \frac{S_{\max} - S_{\min}}{2\pi R_s}.
\]

(2)

In the above expressions, \( S_{\max} \) is the maximum elastance, \( S_{\min} \) is the minimum elastance, and \( R_s \) is the series resistance. Figure 6 shows an equivalent circuit model for the intrinsic HBV (excluding parasitics) along with capacitance-voltage curves based on equation (1). The “sharp,” “flat,” and “cubic” curves are obtained by varying the parameters \( \beta \) and \( \gamma \) in equation (1). Simple \( C(V) \) models such as this can be useful in analyzing basic performance trade-offs and limitations of symmetric varactor multipliers.

![Fig. 6. (a) Simple lumped-element circuit model for the intrinsic HBV. The nonlinear capacitance (or elastance) appears in parallel with a conductance that represents conduction current through the device. (b) Typical capacitance-voltage curves for an HBV device, illustrating the “flat”, “cubic”, and “sharp” models.](image)
Solving Poisson's equation and applying thermodynamic principles leads to a more complete, physical model for the HBV capacitance [46,47,48]. Using this approach, the voltage-charge relation for an HBV can be expressed as,

\[
V(q) = N \left( \frac{bq}{\varepsilon_b A} + 2 \frac{sq}{\varepsilon_d A} + \text{sign}(q) \left( \frac{q^2}{2eN_d \varepsilon_d A^2} + \frac{4kT}{e} \left( 1 - \exp \left( - \frac{|q|}{2L_D A eN_d} \right) \right) \right) \right)
\]

(3)

where \( N \) is the number of stacked barriers, \( b \) is the barrier thickness, \( s \) is the modulation layer thickness, \( N_d \) is the doping of the modulation layers, \( e \) is the fundamental unit of charge, \( \varepsilon_b \) is the permittivity of the barrier layer, \( \varepsilon_d \) is the permittivity of the modulation layer, and \( A \) is the device area. \( L_D \) is the intrinsic Debye length for electrons, given by

\[
L_D = \sqrt{\frac{kT \varepsilon_d}{N_d e^2}}
\]

(4)

where \( k \) is Boltzmann's constant and \( T \) is temperature. Physics-based models such as this, when coupled with modern harmonic-balance simulators, provide a powerful tool for analyzing and designing high-frequency multipliers based on the HBV [48].

It is crucial, for accurate HBV multiplier simulations, to model not only nonlinear capacitance but to account for conduction currents through the device. Because current flow through an HBV depends critically on temperature, it is not sufficient to merely characterize the conduction current of the device under ambient conditions. A number of empirical temperature-dependent models have been suggested and used to model current flow through these devices with good results [49,50,51]. To properly represent the series resistance, particularly at terahertz frequencies, these models are usually augmented with an empirical current-dependence to account for velocity saturation effects [52].

**Structure and Fabrication of HBV Devices**

Although the HBV is very simple structurally, a variety of different geometries and associated fabrication techniques have been applied to these devices in an attempt to optimize their performance at submillimeter frequencies. The simple mesa structure, as shown in Fig. 5, is convenient for circuits employing whisker-contacted devices and
many of the initial multiplier circuits utilizing HBV's were of this type [37,53]. In this scheme, one contact is made with a sharpened whisker wire and the backside contact, which is formed on the highly-doped substrate, is soldered directly to a waveguide or circuit mount. More recently, attention has focused on adapting processing technologies to create planar HBV's with integrated contact fingers [54]. This approach has numerous advantages including,

- greater mechanical stability and robustness,
- precise control of the circuit features (and consequently, the parasitics) in the vicinity of the diode contact, and
- the potential to fabricate fully integrated circuits incorporating HBV's with filters, waveguide probes, and other passive circuit elements.

An example of a planar HBV and a fully integrated circuit fabricated incorporating an HBV with filters and probes is shown in Fig. 7.

Numerical simulations as well as measurements have shown that HBV's operating at elevated temperatures suffer reduced efficiencies due to the increase in thermionic emission over the barrier and higher series resistance. Thus, heat removal from the HBV becomes a critical factor, particularly for devices containing multiple barriers and operating at high pump powers [19]. The Chalmers group that invented the HBV has addressed this problem by developing a new fabrication process whereby the semiconducting substrate is removed and replaced with copper [55,56]. In this technology, a Ti/Pt/Au ohmic contact is evaporated onto the surface of the substrate and

Figure 7. (a) Scanning electron micrograph of a planar HBV device fabricated on InP. The airbridges and contact fingers are clearly shown. (b) A frequency tripler consisting of an HBV integrated onto a quartz substrate with filter and probe structures.
subsequently electroplated with copper to a thickness of approximately 50 μm. Afterwards, the semiconducting substrate is etched away to an InGaAs contact/etch stop layer. Then, HBV diode mesas are formed using standard photolithographic and etching techniques. An advantage of this technology is that the diodes are fabricated after the semiconductor substrate is removed. This post-processing approach reduces stress in the devices and eliminates alignment difficulties.

**HBV-Based Frequency Multipliers**

Since its invention in 1989, the HBV has been used in a wide variety of multiplier designs covering the upper millimeter and lower submillimeter-wave regions of the spectrum. Many of these designs have employed whisker-contacted or planar devices mounted in the common “crossed-waveguide” configuration [57]. This type of design consists of two waveguide channels (designed for single mode propagation at the fundamental and third harmonic, respectively), a cross channel that supports waveguide probes and a lowpass filter, and moveable backshorts for impedance tuning. The crossed-waveguide design tends to be favored in many cases because the electromagnetic environment seen by the HBV can be accurately predicted using commercial electromagnetic structure simulators, the embedding impedances can be conveniently adjusted, and the entire system can be easily cooled, if desired.

Other multiplier architectures that take advantage of planar HBV devices and newer fabrication technology are beginning to emerge. These include distributed multipliers based on nonlinear transmission lines and quasi-optical triplers that incorporate antenna-coupled HBV’s integrated into an array. Both of these circuit structures are made possible by the development of planar HBV diodes and may eventually address some of the bandwidth limitations that are inherent in non-TEM waveguide multiplier circuits. A compilation of HBV multipliers reported in the technical literature is given in Table 1.

**5. Higher Order Multipliers**

Although an ideal varactor diode can in principal be optimized for any harmonic order, in practice it is very difficult to achieve the required embedding impedance at so many frequencies. In fact the best frequency doublers and triplers use symmetry to suppress power coupling to the odd and even harmonics, respectively, so that the number of frequencies that must be considered is minimized. However, for the fourth harmonic and higher, symmetry alone cannot suppress all of the important unwanted harmonics. For this reason there has been very little development of frequency multipliers beyond the third harmonic. Until recently the only exception has been a few studies of comb generators using whisker contacted Schottky diodes as varistor multipliers [58,59].

Recently, Hesler [31] has successfully developed a novel circuit design for wideband frequency quintuplgers. This architecture uses symmetry to suppress even harmonics and
Table 1. Performance comparison of some representative HBV multipliers reported in the literature.

<table>
<thead>
<tr>
<th>Frequency (GHz)</th>
<th>Efficiency (%)</th>
<th>Output Power (mW)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 X 13.5</td>
<td>7.6</td>
<td>12</td>
<td>Krishnamurthi, 1995, [60]</td>
</tr>
<tr>
<td>3 X 31</td>
<td>10</td>
<td>91</td>
<td>Rahal, 1995, [42]</td>
</tr>
<tr>
<td>3 X 33</td>
<td>0.7</td>
<td>1250</td>
<td>Liu, 1993, [61]</td>
</tr>
<tr>
<td>3 X 43.5</td>
<td>7</td>
<td>10</td>
<td>Hollung, 2000, [62]</td>
</tr>
<tr>
<td>3 X 47</td>
<td>8</td>
<td>11.5</td>
<td>Hollung, 1999, [63]</td>
</tr>
<tr>
<td>5 X 34</td>
<td>0.78</td>
<td>0.029</td>
<td>Räisänen, 1995, [64]</td>
</tr>
<tr>
<td>3 X 62</td>
<td>2</td>
<td>0.80</td>
<td>Choudhury, 1993, [65]</td>
</tr>
<tr>
<td>3 X 70</td>
<td>2.8</td>
<td>1.4</td>
<td>Meola, 1998, [66]</td>
</tr>
<tr>
<td>3 X 72</td>
<td>5.4</td>
<td>5</td>
<td>Mélique, 1999, [67]</td>
</tr>
<tr>
<td>3 X 74</td>
<td>7.9</td>
<td>7.1</td>
<td>Dillner, 2000, [56]</td>
</tr>
<tr>
<td>3 X 74</td>
<td>5</td>
<td>2</td>
<td>Rydberg, 1990, [37]</td>
</tr>
<tr>
<td>3 X 82</td>
<td>4.8</td>
<td>4</td>
<td>Stake, 1999, [54]</td>
</tr>
<tr>
<td>3 X 82.5</td>
<td>12</td>
<td>9</td>
<td>Mélique, 1999, [67]</td>
</tr>
<tr>
<td>3 X 84</td>
<td>2.5</td>
<td>2</td>
<td>Jones, 1997, [68]</td>
</tr>
<tr>
<td>3 X 100</td>
<td>4</td>
<td>4</td>
<td>Duan, Qun, 2002, [96]</td>
</tr>
</tbody>
</table>

Fig. 8. The measured performance of an integrated frequency quintupler to 500 GHz developed as a local oscillator for superconductive receivers (courtesy of Virginia Diodes, Inc. [31]).
successfully tunes the third harmonic. Although seventh harmonic power can be measured at the output, it is typically down by 20dB from the fifth harmonic. Results for a quintupler to 500 GHz are shown in Fig. 8. This result shows the broadband nature of the circuit that is achieved without mechanical tuners or dc bias. Furthermore, this result is achieved with a GaAs-on-quartz integrated circuit so that no diode soldering is required and no bond wires are used. Therefore, the results are very repeatable and the component is extremely robust. Although the output power level is not particularly large, this level of performance is ideal for use as a local oscillator for superconductive receivers and for spectroscopy experiments.

6. Distributed Frequency Multipliers

Practical limitations on the bandwidth of lumped-element varactor multipliers and the inherent low efficiencies of broadband varistor devices make harmonic generation based on distributed reactances a natural and attractive approach to multiplier design. Researchers working in this field have focused primarily on nonlinear transmission lines (NLTL’s) using two fundamental designs: (1) “uniform” NLTL multipliers that employ a semiconducting substrate and Schottky contact to form the waveguiding structure, and (2) “discrete” NLTL’s consisting of a planar waveguide that is periodically loaded with varactor diodes. While uniform nonlinear transmission lines are relatively simple structures and are readily fabricated using standard Schottky diode processing techniques, ohmic losses associated with the Schottky contact and semiconductor substrate have limited doubling efficiencies to about 2% [69]. Better performance, in terms of efficiency and bandwidth, has been obtained with discrete NLTL’s due to the lower distributed shunt conductance and use of semi-insulating substrates.

The potential of discrete nonlinear transmission lines to produce second harmonic power with very high efficiency has been demonstrated by Wedding and Jäger who have achieved efficiencies in excess of 90% at rf frequencies [70]. In their approach, an artificial transmission line made of lumped elements is used to generate a second harmonic signal that propagates through the structure as a backward wave. Using this method, Wedding and Jäger were able to match the phase velocities of the fundamental and second harmonic signals, a condition that is necessary for high-efficiency operation. Unfortunately, their phase-matching scheme is quite narrowband, limiting the bandwidth of the circuit.

A unique advantage of discrete nonlinear transmission lines is their natural filtering property. The periodic structure of the discrete NLTL acts as a low-pass filter, preventing wave propagation above a Bragg frequency of (in the small signal approximation),

\[ f_b = \frac{1}{2\pi \sqrt{LC(v_b)}}, \]  

(5)
where $v_b$ is the bias voltage, $L$ is the inductance and $C(v_b)$ is the (nonlinear) capacitance associated with a single section of the transmission line. This filtering property of discrete NLTL's can be useful in reactively terminating unwanted high-order harmonics that are generated by the varactor diodes.

Typical discrete NLTL multipliers consist of a high-impedance coplanar waveguide with integrated reverse-biased Schottky diodes, as shown in Fig. 9. The diodes act as voltage-dependent capacitors that modulate the phase-velocity of a signal propagating along the transmission line. Because of this velocity modulation, an incident voltage transient applied to the NLTL experiences pulse compression and the traveling wavefront forms a "shock wave" with fall time on the order of a few picoseconds. The combination of dispersion and amplitude-dependent phase velocity results in the formation of solitons that are rich in harmonic content. For good quality Schottky varactors fabricated on GaAs, the spectrum of the resulting pulse train extends to approximately 300 GHz, limited by the cutoff frequency of the diodes and the Bragg frequency of the NLTL. For efficient second harmonic generation, the spacing of the varactors is chosen so that the third harmonic lies in the stopband of the periodic structure [71].

Fig. 9. Diagram of a distributed frequency multiplier based on a nonlinear transmission line.
A fundamental goal in the design of NLTL multipliers is minimization of transmission line and diode losses. The skin effect is a major loss factor for distributed multipliers and mitigating its contribution requires the use of relatively large conductor widths and lower-impedance transmission lines. This, however, necessitates the use of higher-capacitance varactors and makes impedance matching the circuit more difficult. Rodwell's group at the University of California at Santa Barbara (UCSB) has addressed these trade-offs by using diodes with a hyperabrupt doping profile to achieve a larger fractional capacitance swing than is possible with the standard abrupt junction diode. Using this approach, they have realized a Ka-band NLTL doubler with approximately 10 dB conversion loss and a 3 dB output bandwidth of 11 GHz [72]. In addition, the UCSB group has extended this technique to W-band NLTL doublers (with ~7 dB conversion loss and 11 GHz 3-dB bandwidth) and V-band NLTL triplers (with ~12 dB conversion loss and 22 GHz 3-dB bandwidth) [73].

More recently, a number of investigators have focused their attention on nonlinear transmission line multipliers using heterostructure barrier varactors (HBV's). In principle, the leakage current through an HBV can be blocked effectively through the use of multiple stacked barriers. This has the benefit of reducing ohmic losses as well as diminishing the effects of carrier velocity saturation. Numerical modeling of distributed HBV multipliers has demonstrated that relatively high efficiencies are attainable (conversion loss of ~ 5 dB at W-band) with 3-dB bandwidths that nearly cover a complete waveguide band (75-105 GHz) [74]. An experimental NLTL tripler employing 15 HBV devices mounted to a finline has given 10 dBm peak power at 130 GHz with approximately 10% bandwidth and 7% conversion efficiency [75].

7. Terahertz Sideband Generators

It is generally true that high-power submillimeter-wave sources tend to be narrowband. Far-infrared lasers, for example, generate up to 100 mW of power at discrete spectral lines with typically no more than 200 MHz of tunable bandwidth [76]. Although fundamental two-terminal solid-state sources, such as Gunn diode oscillators and transit time devices, continue to advance [77], such devices designed to produce submillimeter-wave radiation usually exhibit limited frequency tuning and relatively poor efficiencies. Similarly, submillimeter frequency multipliers that are optimized for high output power and/or high efficiency are often limited to narrow frequency bands due to the impedance transformers and filtering networks required to properly match the input and output, as well as idle unused harmonics [78,79]. Sideband generation provides an attractive method for obtaining tunable terahertz signals by mixing these narrowband, high-power submillimeter-wave sources with a broadly tunable low-frequency (microwave) oscillator.

Fundamentally, a sideband generator is a modulator designed to convert power from a fixed carrier frequency to an upper or lower sideband. This frequency conversion may be achieved by modulating either the amplitude or phase (or both) of a carrier wave with an applied pump signal. Terahertz sideband generation can be accomplished by a number of different methods, most notably through the use of electronic and mechanical tuning
elements. In some instances, researchers have employed optical choppers to modulate terahertz signals and produce offset sidebands of a few kilohertz. For broadband tuning (of a few GHz or more), GaAs Schottky tuners provide the most attractive option.

Sideband generators based on Schottky diode resistive mixers have been used for a number of years and have produced useful and tunable radiation at terahertz frequencies, but with relatively high conversion loss and limited output power. As an example, a whisker contacted 1T15 Schottky diode fabricated at the University of Virginia (UVA) [80] and mounted in a corner-cube antenna has given 10.5 μW of sideband power with a carrier-to-sideband conversion loss of 30 dB [81]. The power output, in principle, can be improved by using a power-combining array of sideband generators, as was demonstrated by Kurtz et al [82]. A different approach that has outstanding potential for sideband generation is based on parametric frequency conversion using a pumped nonlinear reactance.

Frequency upconverters based on varactor parametric devices have been a subject of interest since the early 1960's. Penfield and Rafuse described the fundamental theory of the abrupt-junction varactor upconverter in 1962 [83]. Initially, the interest in parametric upconverters focused on their application to multi-channel frequency division communication systems. Consequently, researchers gave considerable attention to the intermodulation properties and dynamic range of these devices in the microwave region [84,85,86]. Others have studied the conditions for optimum conversion efficiency, including the effects of embedding impedances and pump power level [87,88,89]. Advances in diode-based resistive mixers and the development of microwave and millimeter-wave field-effect transistors during the 1970's and 1980's largely eliminated the need for and use of parametric upconverters for communications. Varactor-based sideband generation, however, remains an attractive technique for submillimeter applications that lie significantly beyond the operating range of modern transistors.

The circuit and device design parameters affecting the performance of sideband generators are similar to those for resistive mixers and frequency multipliers. In particular, the conversion loss of a sideband generator depends on the impedance terminations at each sideband frequency as well as the waveform of the pump signal. Barber [90] and Saleh [91] have shown that the minimum conversion loss for a resistive mixer with all sidebands terminated in their optimum impedance occurs for a pump signal with low pulse-duty ratio. Typical sideband generators operating at submillimeter wavelengths utilize a pump signal at a relatively low frequency (in the microwave or millimeter-wave range of the spectrum). It is important to note that the use of a low-frequency pump results in an output spectrum consisting of a large number of closely spaced sidebands around the submillimeter-wave carrier. Consequently, it is often difficult or impractical to optimize the embedding impedance presented to each relevant sideband. A pragmatic approach is to terminate all sideband frequencies with identical matched loads. This results in a simple and broadband circuit with reasonable embedding impedances. Kelly has shown that a lossless mixer with all sidebands terminated in matched loads and conductance driven between a perfect open-circuit and a perfect short-circuit performs best at a 50 % duty cycle and yields a minimum conversion loss of 3.92
dB [92]. This conversion loss can be considered as a practical limit on the performance of submillimeter-wave sideband generators.

While many of the sideband generators designed for submillimeter applications have been based on resistive mixing, much recent work has focused on varactor-based sideband generation. In 2000, Kurtz demonstrated a varactor phase-modulator operating as a sideband generator at W-band in which a series resonant circuit modulates the reflection coefficient presented to an incident millimeter-wave signal. This rather crude, but simple, circuit consisted of a planar varactor diode suspended across a waveguide with bond wires and produced an overall carrier-to-sideband conversion gain of −9 dB with a 3 dB bandwidth of 10% [93]. Later, this technique was extended to 1.6 THz using a whisker-contacted varactor chip mounted in reduced-height waveguide. Submillimeter radiation was coupled to the circuit with an integrated diagonal feedhorn and a sliding backshort was used to optimize the embedding impedance presented to the diode. The circuit, which is shown in Fig. 10, resulted in a carrier-to-sideband conversion gain of −14 dB and a sideband output power of 55 µW [94]. These results represented a 17 dB improvement in conversion gain and a fivefold increase in power output over the previously published best performance which was obtained with a Schottky mixer diode mounted in a corner-cube [81].

Fig. 10. Photograph of a 1.6 THz varactor sideband generator employing a planar whisker contact and reduced-height waveguide. The submillimeter signal is input through a diagonal feedhorn (not shown in the figure) and the microwave pump is applied through a stepped-impedance lowpass filter [94].
Multiplier and Harmonic Generator Technologies for Terahertz Applications

Fig. 11. Photographs of (a) an integrated GaAs-on-quartz varactor Schottky diode optimized for sideband generation at 600 GHz and (b) a reflection coefficient modulator used as a sideband generator at 200 GHz [95].

Further improvements in varactor sideband generator performance can be expected with the application of planar diode processing technology and the realization of fully integrated sideband generator circuits. Xu and coworkers have investigated optimized circuit architectures and diode parameters for sideband generation using planar devices [95]. At 200 GHz, they have achieved an upper sideband conversion loss of 7–10 dB over a 20 GHz tuning range (see Fig. 11).

8. Summary

Semiconductor-based frequency multipliers and harmonic generators have played a critical role in the development of millimeter and submillimeter-wave systems. Although the extraordinary progress in millimeter-wave transistors and fundamental solid-state sources has allowed some replacement of harmonic generators in the millimeter-wave region, frequency multipliers remain a crucial technology for scientists and engineers working in the frequency range from 100 GHz to 10 THz. In recent years, the most important advances in submillimeter multiplier technology have resulted from the application of more sophisticated and robust device processing technologies that permit the replacement of lossy and dispersive substrates, the incorporation of multiple devices, and the direct integration of tuners, filters, and probe circuits. The application of these new processing technologies have allowed tighter control over parasitics in the vicinity of the diodes and enabled precise alignment and placement of diodes in the surrounding circuitry. Modern processing technology also permits more intricate circuit architectures and provides circuit designers the ability to be more creative and thereby realize multiplier topologies that were unachievable only a few years ago.
With the advances in semiconductor processing techniques comes the need for more precise simulation and design tools. Fortunately, the RF and microwave industry now has access to a number of commercially available circuit and electromagnetic simulation packages that have been shown to be useful for submillimeter-wave multiplier design. The utility of these tools in the design of terahertz multipliers has been demonstrated by a significant number of researchers and using them is now becoming standard practice.

Although terahertz multiplier design has reached a relatively high level of maturity in the past decade, it is becoming increasingly clear that no single terahertz source technology will be sufficiently flexible to serve all applications and requirements. Consequently, a number of researchers have begun to explore new circuit and system concepts that, potentially, can expand the functionality of a given source technology, such as frequency multipliers. One example that has been presented in this discussion is the sideband generator, a device that can extend the frequency tuning of a narrowband, high-power source through modulation. While it is certain that submillimeter-wave multipliers, enabled by advances in fabrication technology, will continue to display greater levels of sophistication it is likely that future terahertz sources will be composite systems that utilize frequency multipliers in conjunction with other components to produce greater flexibility and better performance for submillimeter-wave applications.

References


96. Y. Duan, X. Qun, private communication, University of Virginia, 2002.
Submicron InP-based HBTs for Ultra-high Frequency Amplifiers

M. Urteaga, S. Krishnan, D. Scott, Y. Wei, M. Dahlstrom, S. Lee, M.J.W. Rodwell

Department of Electrical and Computer Engineering, University of California
Santa Barbara, CA 93106, U.S.A.

Transistor bandwidths are approaching terahertz frequencies. Paramount to high speed transistor operation is submicron device scaling. High bandwidths are obtained with heterojunction bipolar transistors by thinning the base and collector layers, increasing emitter current density, decreasing emitter contact resistivity, and reducing the emitter and collector junction widths. In mesa HBTs, minimum dimensions required for the base contact impose a minimum width for the collector junction, frustrating device scaling. We have fabricated HBTs with narrow collector junctions using a substrate transfer process. HBTs with submicron collector junctions exhibit extremely high $f_{\text{max}}$ and high gains in mm-wave ICs. Transferred-substrate HBTs have obtained record 21 dB unilateral power gain at 100 GHz. Recently-fabricated devices have shown unbounded unilateral power gain from 40-110 GHz, and $f_{\text{max}}$ cannot be extrapolated from measurements. However, these devices exhibited high power gains at 220 GHz, the frequency limit of presently available microwave network analyzers. Demonstrated amplifier ICs in the technology include reactively tuned amplifiers at 175 GHz, lumped and distributed amplifiers with bandwidths to 85 GHz, and W-band power amplifiers.

Keywords: InP; HBT; small-signal amplifiers; G-band electronics.

1. Introduction

Device scaling – the reduction of device layer thicknesses and lithographic feature dimensions – is essential to extending the operating frequency of transistor-based integrated circuits. The benefits of aggressive device scaling are illustrated in silicon CMOS technology where progressive reduction in transistor gate length has been essential to the rapid increases in microprocessor speeds. III-V compound semiconductors offer inherent material advantages over silicon. These advantages include higher electron mobilities, higher electron saturation drift velocities, and stronger heterojunctions than Si/SiGe. Extending transistor technology towards THz frequencies will require combining these material advantages with deep submicron device scaling.

The gate lengths of III-V-based high electron mobility transistors (HEMTs) have been scaled to submicron dimensions. InP-based InGaAs/InAlAs HEMTs have exhibited impressive high frequency performance. Devices in this technology with 45 nm gate lengths have been reported with maximum current gain cut-off
frequencies ($f_T$) of over 400 GHz\textsuperscript{1}. Separately, transistors with 100 nm gate lengths and maximum frequencies of oscillation ($f_{\text{max}}$) of 600 GHz have been reported\textsuperscript{2}. HEMT-based multi-stage amplifiers with large power gains in the 140-220 GHz band have also been reported\textsuperscript{3, 4, 5, 6, 7}. State-of-the-art HEMT amplifier results include: a 3-stage amplifier with 30 dB gain at 140 GHz\textsuperscript{3}, a 3-stage amplifier with 12-15 dB gain from 160-190 GHz\textsuperscript{4}, and a 6-stage amplifier with 20 ± 6 dB gain from 150-215 GHz\textsuperscript{5}.

In contrast to HEMTs, aggressive scaling of III-V heterojunction bipolar transistors (HBTs) has not been prevalent. InP and GaAs-based HBTs are typically fabricated with emitter widths of 1-2 $\mu$m, and collector junction widths of 3-5 $\mu$m. By comparison, state-of-the-art Si bipolar and Si/SiGe HBTs are fabricated with < 0.2 $\mu$m emitter-base junction width. SiGe devices with 0.14 $\mu$m emitter-base junction widths have been reported with 92 GHz $f_T$ and 108 GHz $f_{\text{max}}$\textsuperscript{8}. Despite disadvantages in material properties, highly-scaled SiGe technologies will challenge III-V integrated circuits for market share in next generation >40 Gb/sec optical fiber communication systems.

The full benefits of scaling III-V HBTs are only realized if all transistor parasitics are simultaneously reduced. Devices with highly scaled emitter-base junctions have been fabricated for low power applications\textsuperscript{9}; however, reduced emitter dimensions have not necessarily correlated to improvements in device bandwidth. The parasitic capacitance of the base-collector junction lying under the base Ohmic contacts presents the most severe limit to HBT scaling. The geometry of the mesa HBT used throughout the III-V community is such that the minimum size for base Ohmic contacts places a lower limit on the size of the collector-base junction, preventing submicron scaling. Approaches to facilitate scaling of the collector-base junction include: removal of excess collector semiconductor using a lateral-etch undercut\textsuperscript{10, 11}, definition of extremely narrow base contacts using $> 10^{20}$/cm$^3$ base layer doping\textsuperscript{12}, and substrate transfer to allow lithographic pattern definition on both sides of the device epitaxial layers.

We have developed a transferred-substrate HBT technology in an InP-based material system. The process allows the emitter-base and collector-base junctions to be simultaneously scaled to submicron dimensions, resulting in dramatic increases in $f_{\text{max}}$. A record unilateral power gain of 21 dB at 100 GHz has been measured in the technology\textsuperscript{13}. Recently fabricated submicron devices have exhibited a small negative output conductance from 40-110 GHz, resulting in unbounded unilateral power gain in the 75-110 GHz band\textsuperscript{14}. As a result, $f_{\text{max}}$ cannot be extrapolated from these measurements. Other device results in the transferred-substrate technology include transistors with simultaneous 295 GHz $f_T$ and $f_{\text{max}}$\textsuperscript{15}, and double heterojunction transistors with 425 GHz extrapolated $f_{\text{max}}$ and 8 V common-emitter breakdown voltage\textsuperscript{16}.

In this paper, general scaling laws for HBTs are reviewed. The transferred-substrate process is subsequently described as a means of realizing the potential of a highly scaled III-V HBT for mm-wave applications. We then present an overview
of our measurement and calibration methods for on-wafer device measurements, as these factors are critical for accurate characterization of submicron devices. Measured transistor results are then presented, and difficulties in extending low frequency device models to high frequencies are described. Finally, ultra-high frequency HBT amplifier design is discussed, and results from the transferred-substrate technology are presented.

2. HBT Scaling

In general, transistor bandwidths are determined by carrier transit times and RC charging time constants. HBT transit times are reduced by decreasing the thicknesses of the base and collector epitaxial layers. Reduction of the HBT's epitaxial thicknesses will lead to an increase in base resistance and collector capacitance unless accompanied by lateral scaling of the base and collector junction widths.

The simplified cross-section of a mesa HBT shown in fig. 1 illustrates the difficulty in scaling the transistor's collector-base junction. The patterned etches and metal depositions that form the HBT junctions result in a device structure where the collector-base junction must lie beneath the full area of the base Ohmic contacts. To obtain low base contact resistance, the base Ohmic contact must be at least one contact transfer length \( L_{\text{contact}} \) wide at the sides of the emitter stripe. In an InGaAs base HBT with 400 Å base thickness and \( 5 \times 10^{19} / \text{cm}^3 \) doping, \( L_{\text{contact}} \approx 0.4 \mu \text{m} \). Processing tolerances for lithographic alignment may further limit the minimum collector-base junction dimensions.

In contrast to the mesa HBT, a simplified cross-section of an idealized HBT structure is shown in fig. 2. Here, the width of the collector-base junction has been effectively decoupled from the width of the base Ohmic contacts. Scaling of the collector-base junction width in this device is limited only by processing tolerances in aligning the emitter and collector stripes. Through substrate transfer, we are able to realize this idealized geometry by lithographically patterning both sides of the device epitaxy. The transferred-substrate process will be discussed further in the Section 3. Next, we will consider the factors that determine HBT bandwidth.

In literature, transistor bandwidth is commonly described by two figures-of-merit: the current gain cutoff frequency \( f_T \), and the power gain cutoff frequency \( f_{\text{max}} \). Independent of \( f_T \), transistors cannot provide power gain at frequencies above \( f_{\text{max}} \). Thus, \( f_{\text{max}} \) defines the maximum usable frequency of a transistor in narrowband reactively-tuned circuits. In more general analog and digital circuits, the transistor figures-of-merit may not accurately predict circuit performance. For instance, \( f_T \) is commonly used to evaluate a transistor's potential in digital logic applications. However, a detailed charge-control analysis of switching times reveals that device current density, collector-base junction capacitance and emitter resistance make much larger fractional contributions to logic gate delay than they contribute to the emitter-collector forward delay \( \tau_{\text{ec}} = 1/2\pi f_T \). In analog and digital circuits, \( f_T \) and \( f_{\text{max}} \) are used to provide a first-order estimate of device transit delays and of the magnitude of the dominant transistor parasitics.
Figure 1: Cross-section of a typical mesa HBT. The emitter-base junction has width $W_e$, length $L_e$ and area $A_e = L_e W_e$, while the collector-base junction has width $W_c$, length $L_c$ and area $A_c = L_c W_c$.

Figure 2: Cross-section of an idealized HBT with the collector-base junction lying only under the emitter. Such device structures can be formed using substrate transfer processes.
Figure 3: Hybrid-π small-signal HBT equivalent circuit. $C_{be,\text{diff}} = g_m(\tau_b + \tau_c)$.

We estimate below the cutoff frequencies from HBT parameters calculated from physical device properties and fit to a lumped-element device model. Experience has shown that a simple hybrid-π small-signal circuit model (fig. 3) is sufficient to describe all but the most highly scaled devices up to a frequency of 110 GHz. Concerns regarding the accuracy of model at higher frequencies and in describing highly scaled devices are discussed in Section 5. Those concerns notwithstanding, analysis using this first-order model proves excellent for determining those terms that limit transistor bandwidth.

The scaling analysis that follows has been presented in greater detail elsewhere. It is repeated here for completeness as the benefits of device scaling motivate our approach towards developing THz frequency electronics.

2.1. Factors determining $f_T$

Our approach to HBT scaling is determined from the parameters that limit device bandwidth. The current-gain cutoff frequency is given by

$$\frac{1}{2\pi f_T} = \tau_b + \tau_c + \frac{kT}{qI_c} (C_{je} + C_{cb}) + (R_{ex} + R_c)C_{cb},$$

where $R_{ex}$ and $R_c$ are the parasitic emitter and collector resistances, $\tau_b$ and $\tau_c$ are the base and collector transit times, $C_{je}$ and $C_{cb}$ are the emitter-base and base-collector junction capacitances, and $I_c$ is the collector current.

Examining each term separately, we begin with the base transit time $\tau_b$. If a linear grading of the base semiconductor bandgap energy with position is used to reduce $\tau_b$, then

$$\tau_b = \frac{T_b^2}{D_n} \left( \frac{kT}{\Delta E} \right) - \frac{T_b^2}{D_n} \left( \frac{kT}{\Delta E} \right)^2 \left( 1 - e^{-\Delta E/kT} \right)$$

$$+ \frac{T_b}{v_{\text{exit}}} \left( \frac{kT}{\Delta E} \right) \left( 1 - e^{-\Delta E/kT} \right),$$

where $T_b$, $D_n$, and $v_{\text{exit}}$ are the thermal velocity, diffusion length, and exit velocity, respectively.
where $\Delta E$ is the grading in the base bandgap energy, $T_b$ the base thickness, and $D_n$ is the base minority carrier diffusivity. For a typical InGaAs base at $5 \times 10^{19}/\text{cm}^3$ doping, 52 meV bandgap grading is sufficient to reduce $\tau_b$ by $\sim 2:1$. For a thick base layer or a large $v_{exit}$, $\tau_b \propto T_b^2$, with InGaAs base layers below $\sim 400 \ \text{Å}$ thickness, the exit velocity term in eqn. 2 adds a significant correction.

The collector transit time $\tau_c$ is the mean delay of the collector displacement current, and in first order analysis is given by $^20$, $^21$

$$\tau_c = \int_0^{T_c} \frac{(1 - x/T_c)}{v(x)} \, dx \equiv \frac{T_c}{2v_{eff}},$$

where $v(x)$ is the position-dependent electron velocity in the collector drift region and $v_{eff}$ is an effective electron velocity. $\tau_c$ is most strongly dependent upon the electron velocity in the proximity of the base, and becomes progressively less sensitive to the electron velocity as the electron passes through the collector $^21$. In HBTs with thin epitaxial layers, nonequilibrium electron transport is observed in the collector drift region $^22$. At low collector-base bias voltages, electrons may travel through a significant fraction of the collector drift region in the high velocity T-valley before acquiring sufficient kinetic energy (0.55 eV for InGaAs $^23$, 0.6eV for InP $^24$) to scatter to the lower velocity satellite L-valley. As a result, $v(x)$ is fortuitously highest near the base. This velocity overshoot effect significantly reduces collector transit times and in thin InGaAs or InP layers ($< 3000 \ \text{Å}$), $v_{eff} = 3-5 \times 10^7 \ \text{cm/s}$. By contrast, measured saturation velocities in thick InGaAs drift layers are in the range of $v_{sat} = 6-9 \times 10^6 \ \text{cm/s}$. $^23$

The $RC$ charging terms in eqn. 1 comprise a significant fraction of the total forward delay of submicron HBTs, and these terms must be considered in detail. Consider first the term $[kT/qI_c]C_{cb}$. The limits of collector current density are set by the onset of base push-out (the Kirk effect $^25$). At high collector current densities, electron space charge screening at the edge of the base-collector junction eventually leads to a collapse of the electric field. Holes may then diffuse into the collector effectively extending the base region and leading to an increase in base transit time and collector-base capacitance. It has been shown that GaAs HBTs may exhibit improved $f_T$ when biased close to the Kirk threshold $^22$, as the reduced electric field at the collector-base junction edge increases velocity overshoot and reduces the collector transit time. We will ignore these considerations in considering the contribution of $[kT/qI_c]C_{cb}$ to $f_T$.

From electrostatic considerations, the maximum collector current before base push-out is $^17$

$$I_{c,max} = A_e(V_{cb} + \phi)A_{v_{sat}}/T_c^2 \propto A_e/T_c^2,$$  

where $v_{sat}$ is an (assumed) uniform electron velocity within the collector, and the collector doping $N_d$ is chosen to obtain a fully-depleted collector at zero bias current and the applied $V_{cb}$. The collector capacitance is $C_{cb} = eA_c/T_c$. With the
HBT biased at $I_{c,max} \propto 1/T_e^2$, $(kT/qI_C)C_{cb} \propto T_e(A_c/A_e)$. This delay term is thus minimized by scaling (reducing $T_e$), but bias current densities must increase in proportion to the square of the desired fractional improvement in $f_T$.

The emitter charging time $(C_{je}[kT/qI_c]$ in eqn. 1) plays a significant role in determining $f_T$. If we were to assume that $C_{je}$ were simply a depletion capacitance, it would be reasonable to expect that this charging time could be minimized simply by making the emitter-base depletion region very thick, by use of very low emitter doping, combined with a thick bandgap grading region in the base-emitter heterojunction. The tradeoffs between the depletion capacitance and excessive charge storage in the depletion layer were considered in detail elsewhere and the results are repeated here. Using methods similar to those used to derive the collector transit time, we obtain

$$C_{je}/A_e = \epsilon/T_{eb} + \frac{\partial}{\partial V_{bc}} \left[ \int_0^{T_{eb}} (x/T_{eb}) qn(x)dx \right], \quad (5)$$

where $T_{eb}$ is depletion layer thickness and $n(x)$ is the electron density in the depletion region. The term $(kT/qI_c)C_{je}$ in eqn. 1 can be then written as

$$(kT/qI_c)C_{je} = \left( \frac{\epsilon A_e}{T_{eb}} \right) \left( \frac{kT}{qI_c} \right) + \frac{\Gamma T_{eb}T_b}{D_n} \int_0^1 \frac{n(\zeta T_{eb})}{n(T_{eb})} \zeta^2 d\zeta, \quad (6)$$

where $\Gamma = kT/\Delta E - (kT/\Delta E - D_n/v_{exit}T_b)e^{-\Delta E/kT}$ is a factor involving the base bandgap grading ($\Gamma \approx 1$ for an ungraded base) and $\zeta = x/T_{eb}$ is a normalized position variable.

The first term in eqn. 6 results from the depletion-layer capacitance, and is minimized using high bias current densities $J_e = I_e/A_e$; the second term reflects storage of mobile electron charge within the depletion layer, and is minimized by reducing $T_{eb}T_b$. This analysis clearly shows that the depletion region thickness cannot be indefinitely extended to reduce base-emitter junction capacitance, as charge storage in the region also contributes to the transistor's forward delay.

The delay term $R_{ex}C_{cb}$ is a major limit to HBT scaling for high $f_T$. Because of the relative sizes of the emitter and collector Ohmic contacts, in a well-designed submicron HBT, $R_c$ is 4:1 to 10:1 smaller than $R_{ex}$ and $R_cC_{cb}$ can be neglected in a first analysis. To calculate $R_{ex}$, we must consider the geometry of the emitter layer structure.

The emitter layer structure of a typical HBT (fig. 4) contains a heavily-doped and narrow-bandgap contact ("cap") layer, and a heavily-doped N++ wide-bandgap emitter layer. A portion of the emitter layer may be more lightly (N+) doped for reduced junction capacitance, and may be of several hundred Å thickness to avoid dopant diffusion from the N++ layer into the emitter-base junction. Accurately
calculating the emitter resistance requires the consideration of the resistivity and dimensions of each of the emitter layers. For submicron emitters, the junction width $W_{e,junct}$ is significantly smaller than the contact width $W_{e,contact}$ due to lateral undercutting of the emitter during etching of the emitter-base junction. For simplicity in scaling analysis, we will approximate

$$R_{ex} \approx \rho_e/A_e$$

(7)

where $\rho_e$ is a fitted parameter, approximately $50\Omega - \mu m^2$ for submicron InAlAs/InGaAs HBTs fabricated to date at UCSB.

The $R_{ex}C_{cb}$ charging time can now be examined. Since $C_{cb} = \epsilon A_c/T_c$,

$$R_{ex}C_{cb} = \left(\frac{\epsilon \rho_e}{T_c}\right) \left(\frac{A_c}{A_e}\right).$$

(8)

This term can constitute a significant delay. In HBTs we have fabricated with 275 GHz peak $f_T$, the substrate transfer process allows $A_c/A_e$ to be kept small at 2.3:1, yet $R_{ex}C_{cb}$ still constitutes 11% of the total $1/2\pi f_T$ forward delay. In mesa HBTs (fig. 1) $A_c/A_e$ is often larger than 2.3:1 and hence $R_{ex}C_{cb}$ will contribute a larger delay. Because $R_{ex}C_{cb} \propto 1/T_c$, thinning the collector to reduce $\tau_c$ also increases $R_{ex}C_{cb}$.

To increase HBT current gain cutoff frequencies, the base and collector layers must be thinned and the bias current density increased. Thinning the collector increases $R_{ex}C_{cb}$, imposing a limit to scaling. Limits to bias current density imposed by device reliability, and loss in breakdown voltage with reduced collector thickness, are two further potential limits to scaling. Finally, unless the device structure of fig. 1 is laterally scaled, vertical HBT scaling for increased $f_T$ will result in reduced power-gain cutoff frequencies $f_{max}$.
In an HBT with base resistance $R_{bb}$ and collector capacitance $C_{cb}$, the power-gain cutoff frequency is approximately $f_{\text{max}} \approx (f_r/8\pi \tau_{cb})^{1/2}$. The base-collector junction is a distributed network, and $\tau_{cb}$ represents an effective, weighted time constant. Because the base-collector junction parasitics are distributed, calculation of $\tau_{cb}$ is complex. To simplify analysis, we will first roughly approximate $f_{\text{max}} \approx (f_r/8\pi R_{bb}C_{cb})^{1/2}$, where $R_{bb}C_{cb}$ is the product of the base resistance and the full capacitance $C_{cb} = \epsilon A_c/T_c$ of the collector-base junction.

The base resistance $R_{bb}$ is composed of the sum of contact resistance $R_{\text{cont}}$, base-emitter gap resistance $R_{\text{gap}}$, and spreading resistance under the emitter $R_{\text{spread}}$. With base sheet resistance $\rho_s$, and specific (vertical) contact access resistance $\rho_v$, we have

$$R_{bb} = R_{\text{cont}} + R_{\text{gap}} + R_{\text{spread}}$$
$$R_{\text{cont}} = \sqrt{\rho_s \rho_v} / 2L_e$$
$$R_{\text{gap}} = \rho_s W_{eb} / 2L_e$$
$$R_{\text{spread}} = \rho_s W_e / 12L_e.$$ (9)

The base-collector time constant is then

$$R_{bb}C_{cb} = \left[ (\sqrt{\rho_s \rho_v} + \rho_s W_{eb}) \left( \frac{\epsilon}{2} \left( \frac{L_c}{L_e} \right) \right) \left( \frac{W_c}{T_c} \right) \right] + \left[ \left( \frac{\rho_s \epsilon}{12} \right) \left( \frac{L_c}{L_e} \right) \right] \left( \frac{W_c W_e}{T_c} \right).$$ (10)

Consider the influence of device scaling on the time constant $R_{bb}C_{cb}$. Decreasing the base thickness to reduce $\tau_{\text{b}}$ increases the base sheet resistivity $\rho_s$, increasing $R_{bb}C_{cb}$. Decreasing the collector thickness $T_c$ to reduce $\tau_c$ directly increases $R_{bb}C_{cb}$, as is shown explicitly in eqn. 10. Low $R_{bb}C_{cb}$, and consequently high $f_{\text{max}}$, is obtained by scaling the emitter and collector junction widths $W_e$ and $W_c$ to submicron dimensions. Reducing the emitter width $W_e$ alone reduces towards zero the component of $R_{bb}C_{cb}$ associated with the base spreading resistance (the second term in eqn. 10). In the normal triple-mesa HBT (fig. 1), the minimum collector junction width $W_c$ is set by the base Ohmic contacts which must be at least one contact transfer length ($L_{\text{contact}} = (\rho_c/\rho_s)^{1/2}$). As a result, the component of $R_{bb}C_{cb}$ associated with the base contact resistance (the first term in eqn. 10) has a minimum value, independent of lithographic limits, and consequently, $f_{\text{max}}$ does not increase rapidly with scaling. Given this minimum $R_{bb}C_{cb}$, attempts to obtain high $f_r$ by thinning the collector have resulted in decreased $f_{\text{max}}$, frustrating efforts to improve HBT bandwidths.
If the parasitic collector-base junction is eliminated, \( f_{\text{max}} \) will instead increase rapidly with scaling. The collector-base junction need only be present where current flows, e.g. under the emitter. We have fabricated such a device (fig. 2) using a substrate transfer process. If we neglect processing alignment tolerances, the emitter and collector junctions can be of equal width, hence \( W_c = W_e \). With submicron scaling of the emitter and collector junction widths, the first term in eqn. 10 dominates and scales as \( W_e \), and in our approximate analysis, \( f_{\text{max}} \) increases as the inverse square root of the process minimum feature size.

To more accurately predict \( f_{\text{max}} \), the distributed nature of the base-collector junction parasitics must be considered. Figure 5 shows a distributed model of a transferred-substrate HBT. Using a small grid spacing, we have entered the resulting network into a microwave circuit simulator (HP-EESOF 26) to calculate—without approximation—the HBT \( f_{\text{max}} \). Alternatively, analytic expressions for \( f_{\text{max}} \) can be developed from hand analysis of the distributed network of fig. 5. Among these is the model of Vaidyanathan and Pulfrey 27, which provides good physical insight into the problem. We now consider the Vaidyanathan/Pulfrey model applied to a transferred-substrate HBT 28.

Figure 6 shows an equivalent circuit representing the Vaidyanathan/Pulfrey model of a transferred-substrate HBT. We define three capacitances. \( C_{cb,e} = \varepsilon L_e W_e / T_c \) is the capacitance of the collector junction lying under the emitter. \( C_{cb,gap} = 2\varepsilon L_e W_{cb} / T_c \) is the capacitance of the collector junction lying under the gap between the emitter and the base contact. \( C_{cb,ext} = 2\varepsilon L_e W_{bc} / T_c \) is the capacitance of the collector lying under the base Ohmic contacts. Components of the base resistance are as defined in eqn. 9, with the exception of two additional resistances \( R_{\text{vert}} = \rho_{\text{c}} / 2L_w c_b \) and \( R_{\text{horiz}} = \rho_s W_{bc} / 2L_e \). \( R_{\text{vert}} \) represents the vertical access resistance through the base Ohmic contact over the path \( W_{bc} \), and \( R_{\text{horiz}} \) represents the lateral sheet resistance over that same path.

The \( R_{\text{vert}} / R_{\text{horiz}} \) network approximates the distributed network charging \( C_{cb,ext} \) in the mesh model of fig. 5. This approximation is valid under the condition that \( W_{bc} \leq L_{\text{contact}} \). If this condition does not hold, \( R_{\text{vert}} \) and \( R_{\text{horiz}} \) must be replaced by a finite element ladder network with a larger number of discrete elements, as in fig. 5. The model of fig. 6 further assumes that the total base width \( W_b \gg L_{\text{contact}} \). Typical geometries of the transferred-substrate HBTs we have fabricated meet both of the aforementioned assumptions.

From fig. 6, we note that charging resistance seen by \( C_{cb,e} \) and \( C_{cb,gap} \) contains the component \( R_x = R_{\text{cont}} \parallel R_{\text{vert}} + R_{\text{horiz}} = R_{\text{cont}} \). While the simplified lumped element model of fig. 6 approximates fig. 5 only if \( W_{bc} \leq L_{\text{contact}} \), the relationship \( R_x = R_{\text{cont}} \) is in general true even for \( W_{bc} > L_{\text{contact}} \), as are the expressions for \( f_{\text{max}} \) presented below.

In the limit of zero collector series resistance, Vaidyanathan and Pulfrey’s model, 27, 28 reduces to

\[
f_{\text{max}} = \sqrt{\frac{f'_r}{8\pi T_{cb}}},
\]

(11)
Figure 5: Distributed model of the HBT base-collector junction for accurate calculation of $r_{cb}$ in eqn. 11. With mesh spacing $\Delta x$, $\Delta G = L_e \Delta x / \rho_c$, $\Delta R = \rho_s \Delta x / L_e$, and $\Delta C = \epsilon L_c \Delta x / T_c$.

Figure 6: Lumped element circuit approximating the transistor mesh model of fig. 5. Terms $R_{cb,canc}$ and $C_{cb,canc}$ represent capacitance cancellation effect described by eqn. 15.
where
\[
\frac{1}{2\pi f_T} = \tau_b + \tau_c + \frac{kT}{qI_c} (C_{je} + C_{cb}),
\] (12)
and
\[
\tau_{cb} = C_{cb,e} (R_{cont} + R_{gap} + R_{spread}) \\
+ C_{cb,gap} (R_{cont} + R_{gap}/2) \\
+ (R_{cont}||R_{vert}) C_{cb,ext}
\] (13)

The model of fig. 6 accurately represents the distributed nature of the collector-base junction. We can approximate this network with the simple hybrid-\pi model of fig. 3 if we select \(C_{cbi}\) such that the correct transistor \(f_{max}\) is obtained. The components of fig. 3 are then given by \(C_{cbi} = \tau_{cb}/R_{bb}\) and \(C_{cbx} = C_{cb} - C_{cbi}\), where \(C_{cb} = C_{cb,e} + C_{cb,gap} + C_{cb,ext}\).

Figure 7 compares the \(f_{max}\) of mesa and transferred-substrate HBTs, computed using the finite-element model. For the transferred-substrate device, \(f_{max}\) increases rapidly with deep submicron scaling. Experimentally, we observe a more rapid variation of predicted \(f_{max}\) with collector width than is shown, and fig. 7 predicts a higher \(f_{max}\) than is experimentally observed for mesa HBTs. Series resistance in the base metallization and collector series resistance \(27\) (not modeled above, and not present in Schottky-collector transferred-substrate HBTs) are possible explanations for the discrepancy.

At high collector current densities, differential space-charge effects in the collector space-charge region result in \(C_{cb,e}\) smaller than \(\epsilon A_e/T_c\), and increase the HBT \(f_{max}\) \(29\), \(30\). In III-V materials at high fields, electron velocity \(v(\mathcal{E})\) decreases with increasing electric field. Modulating the collector voltage \(V_{cb}\) modulates the collector transit time \(\tau_c\) (eqn. 3), and partially modulates the space-charge in the collector drift region. This modulated space-charge partially screens the base from modulations in the collector applied field, and \(C_{cb,e}\) is reduced to \(30\)
\[
C_{cb,e} = \epsilon A_e/T_c - I_c \frac{d\tau_c}{dV_{cb}}
\] (14)
where \(\epsilon A_e / T_c\) is the normal dielectric capacitance of the junction. Experimental data confirming \(C_{cb}\) cancellation will be shown in Section 5.

The derivation of eqn. 14 is limited by the charge control assumption that changes in carrier concentration occur instantaneously. Clearly, this is not valid at frequencies approaching the inverse of the collector transit time. We now describe the dynamics of capacitance cancellation to first order in frequency, assuming a simplified velocity dependence \(1/v(\mathcal{E}) \simeq \kappa_0 + \kappa_1 \mathcal{E}\). We further assume that the electric field induced by mobile collector charge is small compared to both the DC
Submicron InP-based HBTs for Ultra-high Frequency Amplifiers

and AC applied field across the collector-base junction. Using a formalism similar to $^{20, 21, 30}$ it can then be shown that

$$\frac{dQ}{dV_{cb}} = \epsilon A_e/T_c - I_c \frac{d\tau_c}{dV_{cb}} \frac{1}{[1 + j\omega(2/3)\tau_c]}$$

(15)

where $d\tau_c/dV_{cb} = \kappa_1/2$. The term $-I_c d\tau_c/dV_{cb}$ is the capacitance cancellation of eqn 14, while the terms in $j\omega$ represent the dynamics of this effect. The differential equation of eqn. 15 can be represented by an equivalent circuit model consisting of the dielectric junction capacitance $C_{cb,e}$ in parallel with the negative capacitance cancellation term, $C_{cb,canc} = -I_c d\tau_c/dV_{cb}$, in series with a negative resistance, $R_{cb,canc} = (3/2)\tau_c/C_{cb,canc}$. These terms are included in the Vaidyanathan/Pulfrey transistor model of fig. 6. We note that $C_{cb,canc}$ and $R_{cb,canc}$ are charged through the total base resistance ($R_{b,cont} + R_{gap} + R_{spread}$). The elements $C_{cb,canc}$ and $R_{cb,canc}$ can therefore appear in the approximate hybrid-$\pi$ model appearing in shunt across $C_{cb}$. A revised hybrid-$\pi$ model including the capacitance cancellation terms is shown in fig. 8.

We caution that this derivation models only to the first order in frequency the dynamics of the space-charge redistribution in the collector region. However, the method, though approximate, is sufficient to predict that negative resistance effects may be observed in highly scaled devices, and may explain device results presented in Section 5.1.
Figure 8: Modified hybrid-π small-signal HBT equivalent circuit with additional negative capacitance $C_{cb,canc}$ and negative resistance $R_{cb,canc}$ elements to account for dynamics of capacitance cancellation.

3. Transferred-substrate HBTs

We now consider the transferred-substrate process as a means of realizing a highly scalable HBT. Substrate transfer provides access to both sides of the device epitaxial material, which allows for the simultaneous definition of narrow emitter and collector stripes. With the extrinsic collector-base capacitance greatly reduced, aggressive lithographic scaling without epitaxial scaling greatly increases $f_{\text{max}}$ at constant $f_{\text{T}}$. If high values of both $f_r$ and $f_{\text{max}}$ are sought, simultaneous lithographic and epitaxial scaling is required. Further improvements in device bandwidth will require operation at higher current densities and reduction of emitter parasitic resistance.

3.1. Growth and fabrication

The MBE epitaxial structure is grown on a Fe-doped semi-insulating InP substrate. Both single and double heterojunction transistors have been fabricated in the transferred-substrate technology. The single heterojunction transistors have an InAlAs/InGaAs emitter-base junction. The double heterojunction devices have an InP collector for increased breakdown, and may have an InP emitter for improved heat flow in the device. A chirped superlattice grade is used to smooth conduction band discontinuities at the heterojunctions. The InGaAs base is typically 300–400 Å thick, has $2kT$ bandgap grading, and is Be-doped at $5 \times 10^{19}$/cm$^3$. The transistor collector thickness is typically 2000–3000 Å and a N$^+$ pulse-doped layer placed 400 Å from the base delays the onset of base push-out at high collector current densities. High $f_{\text{max}}$ devices are typically fabricated with Schottky collector contacts which provide zero collector series resistance.

Figure 9 shows the process flow. Standard fabrication processes are used to
define the emitter-base junction, the base mesa, polyimide planarization, and the emitter contacts. The IC wiring environment consists of thin-film NiCr resistors, two levels of metal interconnects, and a PECVD Si₃N₄ insulator layer for MIM capacitors. The substrate transfer process commences with deposition the Benzocyclobutene (BCB) transmission-line dielectric (5 μm thickness). Thermal and electrical vias are etched in the BCB. The wafer is electroplated to metallize the vias and to form the ground plane. The wafer is then solder-bonded to a GaAs or AlN carrier substrate. The InP substrate is removed in a selective HCl etch, and collector contacts are patterned and deposited on the exposed collector epitaxy, completing the process. Fig. 10 shows a detailed device cross section.

Minimum device feature sizes are determine by the lithography system used in the process. The projection lithography system at UCSB can define emitter widths down to 0.5 μm. The relative sizes of the emitter and collector junctions are determined by lithographic alignment tolerances, and the collector stripe width must exceed the emitter stripe width by twice the lithographic alignment tolerance. Our projection lithography system aligns to 0.1–0.3 μm registration, depending on the time since maintenance. For deep submicron devices, a JEOL JBX electron-beam lithography system at UCSB is used. With this system, 0.2 μm emitter and 0.3 μm collector stripe widths have been realized. Collector alignment of better than 0.1 μm is achieved using local registration marks for each device.
Figure 10: Schematic cross-section of a transferred-substrate HBT

Figure 11: E-beam HBT: HBT structure with 0.3 \( \mu \text{m} \) emitter-base junction (a), and 0.7 \( \mu \text{m} \) Schottky collector stripe (b)
For the emitter-base junction, deep submicron scaling requires tight control of lateral undercutting during the base contact recess etch. The undercut both narrows the emitter and defines the liftoff edge in the self-aligned base contact deposition. For InAlAs emitters, a combination dry and wet etch is used. A CH$_4$ / H$_2$ / Ar dry etch removes the N$^+$ InGaAs emitter contact layer and etches into the InAlAs emitter. A HCl/HBr/Acetic selective wet etch then removes the InAlAs emitter, stopping on the InAlAs/InGaAs emitter-base grade. A timed nonselective Citric/H$_3$PO$_4$/H$_2$O$_2$ wet etch then removes the base-emitter grade, etching ~100 Å into the base. For recently fabricated InP emitter devices, an all wet etch process has been employed. A H$_2$O$_2$/H$_3$PO$_4$ etch removes the N$^+$ InGaAs emitter contact layer stopping on the InP emitter. A selective HCl/H$_3$PO$_4$ etch then removes the emitter semiconductor again stopping in the emitter-base grade. The same nonselective etch used for the InAlAs emitter is then used to etch into the base. For both emitter etch processes, a well controlled undercut of 0.1 μm is achieved.

The collector junction is defined by the stripe width of the deposited metal. Subsequent to collector deposition, a self-aligned wet etch of ~1000 Å depth removes the collector junction sidewalls (eliminating fringing fields) and reduces the collector junction width by ~2000 Å.

4. High Frequency Device Measurements

Prior to discussing transferred-substrate device results, we will consider issues related to high frequency device measurements. Submicron transistors have extremely small reverse transmission characteristics and low shunt output conductances. These features make device measurement and model extraction challenging even in the DC-50 GHz band covered by typical commercial vector network analyzers (VNAs). As state-of-the-art transistor bandwidths far exceed this frequency range, we would like to measure device scattering parameters (S-parameters) at as high a frequency as possible. Presently, VNA test set extensions are available covering frequencies up to 220 GHz, and 325 GHz systems will soon be available. Accurate device measurements at these frequencies require that careful attention be paid to measurement and calibration methodology.

4.1. Ultra-high frequency measurement systems

The 140-220 GHz VNA measurement system used at UCSB consists of an Agilent HP8510C network analyzer interfaced with Oleson Microwave Lab millimeter wave VNA extensions. Frequency synthesizers in the VNA test system generate 17.5-27.5 GHz RF and 11.6-18.4 GHz LO signals that are sent to the Oleson extenders through microwave coaxial cables. A harmonic multiplier chain in the extender upconverts the RF signal to the measurement frequency, and harmonic mixers downconvert stimulus and response signals obtained from a dual directional coupler at the extender's test port. The IF signals ( < 300 MHz) are then sent back to the VNA for processing. Full two-port transmission and reflection measurements can be ob-
tained with this system, and the dynamic range is greater than 50 dB. The test ports of the extenders are connected to on-wafer probes through short lengths of WR-5 rectangular waveguide. Due to the relatively high loss of the waveguide (12 dB/m), it is important to keep the connection length as short as possible to preserve the measurement system’s dynamic range. At UCSB, ~ 50 cm of waveguide with two right angle bends is used. This arrangement is found to provide adequate loss and sufficient range of motion for probe manipulation. GGB Industries ground-signal-ground wafer probes are mounted onto probe station micromanipulators. A waveguide-to-microcoax transition is internal to the probes, and the insertion loss of the probes is better than 3 dB across the band. Internal bias-tees in the wafer probes allow for biasing of active devices through the center probe conductor.

4.2. On-wafer calibration

To underscore the importance of measurement calibration for submicron device measurements, consider that a state-of-the-art InP HBT with a 300 GHz $f_T$ has a total forward delay of 0.53 psec, the same delay as ~ 100 μm length of transmission line in our on-wafer transmission line environment ($\epsilon_{ef} = 2.2$). One sees the importance of removing from transistor measurements the effects of all extraneous propagation delays and losses incurred in the measurement system up to the device-under-test. An accurate VNA calibration will place the measurement reference planes precisely at the input and output of the device-under-test. However, standard 12-term VNA error corrections do not account for leakage and coupling between on-wafer probes. Highly scaled transistors have extremely small reverse transmission characteristics (S12) and excessive probe-to-probe coupling can easily corrupt device measurements. Probe-to-probe leakage can be accounted for using more complicated 15- or 16-term VNA error corrections $^{33, 34}$. However, these corrections require precise characterization of calibration standards, and such characterizations are difficult to achieve for on-wafer elements, particularly at mm-wave frequencies. The use of 15- or 16-term error corrections can be avoided if the wafer probes are spaced far enough apart to provide sufficient isolation. Probe isolation that is at least 20 dB lower than S12 of the transistor is sufficient for accurate device characterization $^{35}$. Separation between wafer probes is achieved by adding lengths of 50 Ω on-wafer transmission line to the input and output of the device. At UCSB, a length of 230 μm transmission line at each port has been found to provide sufficient isolation.

Calibrating the network analyzer involves the measurement of various known calibration standards. A standard approach for device measurements is to utilize a separate calibration substrate that has on it an array of characterized calibration standards. These substrates are available commercially and cover various frequency ranges up to 110 GHz. The goal of these calibrations is to place the measurement reference planes at the wafer probe tips. There are several drawbacks to using this approach for precise device measurements.

As previously mentioned, the transistor is embedded on-wafer between lengths of
transmission line. If we use a probe tip calibration, the effect of the embedding structures must be eliminated from the transistor measurements. An ad hoc approach often used is to measure the capacitance of an open circuit embedding structure and then subtract this capacitance from the measured device S-parameters. This approach can lead to considerable error as the pad capacitance may be of the same order as the input capacitance of a submicron device. This approach also ignores the series resistance of the embedding structure, and the series inductance, which will have a considerable effect at mm-wave frequencies. A more precise determination of the electrical characteristics of the embedding structures may made by modeling the structures electromagnetically, or by measuring the test structures without devices and fitting the results to a lumped element model. In either case, this adds a level of complexity and the opportunity for further error in extracting device parameters.

The second drawback of probe-tip calibration approach is that calibration substrates generally have a different transmission line environment than the device under test. A standard VNA calibration assumes that only a single propagation mode exists at the calibration reference plane for both measurement and calibration. The discontinuity at the probe/wafer interface does not meet these conditions, and the field distribution at the discontinuity will depend on the transmission line environment that is being coupled into. As such, the probe-tip calibration on the calibration substrate need not apply to the substrate of the device-under-test. We expect discrepancies to increase at higher frequencies as the wavelength approaches the size of the probe tips.

The alternative to probe-tip calibration is to calibrate to the ends of the on-wafer transmission lines. This places the measurement reference planes at the input and output of the device under test, but requires the realization of custom on-wafer calibration standards. Depending on the calibration used, different types of standards are required. Certain VNA calibrations, such as the commonly used Short-Open-Load-Through (SOLT) calibration, require precise characterization of the electrical properties of the standards. Fringing fields and the distributed nature of the elements at mm-wave frequencies require that the elements be modeled electromagnetically or measured using an accurate on-wafer calibration. Again, this adds a level of complexity and the opportunity for further error in extracting device parameters.

The Through-Reflect-Line (TRL) calibration is well-suited for an on-wafer measurement environment. The calibration uses two transmission line standards one of which is designated "through", the other is designated "line" and differs from the through line by some electrical length $\Delta L$. The final "reflect" standard may be an open or short termination. An advantage of the TRL calibration is that the solution for the calibration error terms is overdetermined, and the reflection coefficient of the reflect standard and propagation constant of the line standard can also be calculated from calibration measurements. The only physical property that must be known is the characteristic impedance of the line standard. The characteristic
impedance can be determined analytically from transmission line models or electromagnetic simulations, or alternatively, it can be calculated from measurements of the line's capacitance and propagation constant \(^{38,39}\). It is important to note that line loss will lead to a characteristic impedance that has frequency dependent real and imaginary parts. The imaginary part can be large at low frequencies and should be accounted for in the measurement calibration.

An often cited disadvantage of the TRL calibration is that one line standard can only cover a 1:8 frequency span, with the ideal \(\Delta L\) being a quarter-wavelength at the center of the span. As such, multiple line standards are required to cover large frequency ranges, and low frequency line standards can take up a large amount of valuable wafer area. Multiple line standards may also be used to provide measurement redundancy in a band, and reduce the error due to probe placement repeatability \(^{40}\).

Quantitatively assessing the accuracy of a microwave calibration is difficult. To partially verify the calibration accuracy, we have re-measured calibration standards after calibration in the 75-110 GHz and 140-220 GHz bands. Measurement of a through standard after calibration gives an indication of probe-placement repeatability, as the calibration defines the measurement reference planes to be at the center of the through line. In the 75-110 GHz band, the measurement of a through line
shows better than 35 dB return loss, and $S_{21}$ has < 0.2 dB amplitude variation and < 0.3 degrees of phase variation. In the 140-220 GHz band, measurement of a through line shows better than 30 dB return loss, and $S_{21}$ has < 0.1 dB amplitude variation and < 1 degree of phase variation. As discussed previously, the TRL calibration does not assume a known reflection coefficient of the reflect standard. Measurement of the short or open reflection standards after calibration, therefore, provides a good indication of the quality of the calibration. In the 75-110 GHz band, measurement of the open standard shows < 0.25 dB amplitude variation and < 1.5 degrees of phase variation. In the 140-220 GHz band, the calibration appeared slightly poorer. Measurement of the open standard showed < 0.4 dB amplitude variation and < 3 degrees of phase variation. Figure 12 shows measurements on a Smith chart of the calibration standards in both frequency bands.

Independent of re-measuring the calibration standards, a quantitative estimate of calibration accuracy would require measurement of known on-wafer elements. Measurements, to be presented in Section 6, show excellent agreement between electromagnetic simulations of passive matching network elements and measured results. Additionally, device measurements presented in the next section show smooth variation across all frequency bands.

5. Device Results

Depending on the circuit application transferred-substrate devices can be aggressively laterally-scaled for ultra-high $f_{\text{max}}$, or both laterally and vertically scaled for simultaneously high values of $f_{\text{max}}$ and $f_T$. As we are concerned here with high frequency tuned circuit applications, we will consider only those devices scaled for high $f_{\text{max}}$. These devices are typically fabricated with 400 Å base thickness for low base resistance, and 3000 Å thick collectors for low collector-base capacitance. This results in moderate, not high, $f_T$.

Figure 13 shows microwave gains for a deep submicron single heterojunction transistor fabricated using electron-beam lithography, reported by Lee et. al. The emitter and collector junction dimensions are 0.4 μm x 6 μm and 0.7 μm x 10 μm, respectively. The measurements were made in the 10-50 GHz and 75-110 GHz frequency bands using the TRL calibration described in the previous section. At the time these measurements were made, the 140-220 GHz measurement set-up had not yet been obtained. With the device biased at $V_{ce} = 1.2$ V and $I_c = 6$ mA ($J_e = 2.5 \times 10^5$ A/cm²), the transistor exhibits an extrapolated $f_T$ of 204 GHz. Mason's invariant (unilateral) power gain is measured to be greater than 20 dB at 100 GHz. If we extrapolate at -20 dB/decade, an $f_{\text{max}}$ of > 1 THz is predicted. However, recent device measurements have indicated that the unilateral power gain of highly scaled devices does not show a well-behaved roll-off with frequency, and as such, a -20 dB/decade extrapolation should not be used. Prior to considering these recent device measurements, we discuss the use of the Mason’s gain to predict transistor $f_{\text{max}}$.

For a general two-port network Mason’s invariant (unilateral) power gain is given
Figure 13: Gains of a 0.4 µm × 6 µm emitter and 0.7 µm × 10 µm collector HBT fabricated using electron-beam lithography. Theoretical -20 dB/decade ($H_{21}$, $U$) gain slopes are indicated.
Figure 14: Variation of transistor gains with frequency, computed from a hybrid-π HBT model. Shown are the maximum available / maximum stable gains (MAG/MSG) in common-emitter, common-base, and common collector mode, and Mason's invariant, U, the unilateral gain

\[ U = \frac{|Y_{21} - Y_{12}|^2}{4(G_{11}G_{22} - G_{12}G_{21})} \]  

(16)

where \( Y_{21} \) and \( Y_{12} \) are network admittance parameters, and \( G_{11}, G_{22}, G_{12}, \) and \( G_{21} \) are the real parts of the admittance parameters. Mason's gain represents the power gain available from a network if it is unilateralized (reverse transmission = 0) using lossless reactive feedback. The gain is invariant with respect to embedding the device in a lossless reciprocal network, and consequently is independent of pad inductive or capacitive parasitics and independent of the transistor configuration (common-emitter vs. common-base). For HBTs well-modeled by a hybrid-π equivalent circuit fig. 3, Mason's gain conforms closely to a -20 dB/decade variation with frequency (fig. 14). In marked contrast, the maximum available / maximum stable gain is a function of the transistor configuration, and shows no fixed variation with frequency. \( f_{\text{max}} \) is unique; at \( f = f_{\text{max}} \) the MAG/MSG and U are both 0 dB. While measurements of U are generally useful for \( f_{\text{max}} \) extrapolation, monolithic amplifiers are not easily made unilateral, and the maximum stable / maximum available gain has more direct relevance in tuned mm-wave amplifier design.

Figure 15 shows microwave gains for a recently-fabricated submicron HBT. The
Figure 15: Gains of recently fabricated submicron HBT. In the indicated frequency bands, Mason’s unilateral power gain $U$ is unbounded as a result of a small negative output conductance $G_{22}$.

Emitter and collector junction dimensions are $0.3 \mu m \times 18 \mu m$ and $0.7 \mu m \times 18.4 \mu m$, respectively. The device was characterized in the 6-45, 75-110, and 140-220 GHz frequency bands. The device is biased at $V_{ce} = 1.1 V$ and $I_c = 5 mA$ ($J_e = 1.1 \times 10^5 A/cm^2$). The transistor measurements show a negative unilateral power gain across the 75-110 GHz band and over parts of the 140-220 GHz band. Above $\approx 45$ GHz, the unilateral power gain increases to infinity, and then becomes negative, a condition under which the addition of an appropriate small resistive attenuation results in infinite $U$.

From eqn. 16 we see that the denominator of the expression for Mason’s gain can be negative if a device has a negative real output conductance $G_{22}$ or a positive real feedback term $G_{12}$. The transistor of fig. 15 exhibits a very small negative output conductance that peaks at approximately $-1\, mS$, leading to the observed negative unilateral gain. An HBT modeled by the hybrid-$\pi$ transistor model cannot show a negative output conductance. We speculate that the effect arises from small secondary HBT transport effects in the collector region, either through the dynamics of base-collector capacitance cancellation as described in Section 2.2, or through weak IMPATT effects in the collector depletion region. These effects would not be seen in a typical III-V HBT because of the large positive output conductance arising from high-frequency feedback through $R_{eb}$ and $C_{ce}$. In contrast, submicron
transferred-substrate HBTs have an extremely small $R_{bb}C_{cb}$ time constant, and such effects can be observed.

The dynamics of capacitance cancellation may be the cause of the negative unilateral gain. A dramatic decrease in measured base-collector capacitance is observed with increased bias current as predicted from eqn. 14. The total collector-base capacitance $C_{cb}$ is determined from the measured variation with frequency of the imaginary part of the admittance parameter. At low frequencies, $\Im[Y_{12}] = j\omega C_{cb}$. The total $C_{cb}$ determined from $Y_{12}$ (fig. 17) shows a 2.4 fF decrease between 0.5 mA and 5 mA $I_c$. Unlike the standard hybrid-$\pi$ model of fig. 3, the modified model of fig. 8, with negative resistance and capacitance elements modeling the dynamics of capacitance cancellation, will show negative output conductance over certain frequencies. However, as of yet, we have been unable to fit a physically-based model of the dynamics of capacitance cancellation to measured S-parameters.

Another possible explanation for the negative output conductance is systematic errors in the S-parameter measurements. As discussed in Section 4, a great deal of work in our group has been put towards developing an accurate calibration methodology. However, measurements of $U$ are inherently difficult, as both products in the denominator of eqn. 16 approach zero for an HBT with small collector-base parasitics. Note we observed little variation in a transistor's S-parameters over a number of measurements with different calibrations. Further, the negative output conduc-
Figure 17: Collector-base capacitance extracted from $Y_{12}$, vs. emitter current

tance was observed for numerous devices of the same dimensions on the wafer. The transistor S-parameters also show relatively smooth variation across all of the measured frequency bands (fig. 16), showing no evidence of resonances or calibration artifacts.

A consequence of the observation of negative $U$ is that we cannot predict $f_{\text{max}}$ of these highly scaled devices from a -20 dB/decade extrapolation. Nevertheless, the maximum stable / maximum available gain of these devices is very high even at 200 GHz. Ultimately, a higher frequency network analyzer and/or the development of higher-frequency amplifiers and oscillators is necessary to determine the usable bandwidth of the devices.

The results described above were for single heterojunction devices with narrow bandgap InGaAs collectors. Recently, double heterojunction transistors have been fabricated with InP collectors for increased voltage breakdown. Figure 18 shows microwave gains and DC I-V characteristics for a double heterojunction transferred-substrate device with a 3000 Å thick collector. The emitter and collector junctions are 0.4 μm × 8 μm and 1.2 μm × 8.75 μm, respectively. The device has extrapolated cutoff frequencies of $f_r = 139$ GHz and $f_{\text{max}} = 425$ GHz. The common emitter breakdown voltage $BV_{CEO}$ is 8 V at $J_e = 5.0 \times 10^4$ A/cm$^2$. The negative output conductance observed in single-heterojunction devices has not been observed in double heterojunction devices, possibly because these devices have not yet been scaled to deep submicron dimensions with electron beam lithography.
Figure 18: RF gains and DC I-V characteristics of double-heterojunction transistor
Using the double heterojunction process large area power transistors have recently been fabricated for W-band power amplifiers. A multi-finger common-base device with a total emitter area of 128 $\mu$m has been measured with an extrapolated $f_{\text{max}}$ of 330 GHz. The transistor has a breakdown voltage of 7 V and a maximum collector current of $I_c = 100$ mA.

### 5.1. Device modeling

Figure 19 shows a small-signal hybrid-$\pi$ model derived from 6-45 GHz measurements of a submicron HBT. The transistor had a 0.4 $\mu$m x 6 $\mu$m emitter junction area and 0.7 $\mu$m x 6.4 $\mu$m collector junction area. The model is for a bias condition of $V_{ce} = 1.2$ V and $I_c = 3.6$ mA ($J_e = 1.5 \times 10^5$ A/cm$^2$). To develop the model, the transistor S-parameters are measured at various DC bias conditions, and the measured Y-parameters are analyzed to extract the bias-dependent parameters, such as the transconductance and emitter-base diffusion capacitance, and the bias-independent terms, such as the extrinsic emitter resistance.

In general, we observe that the hybrid-$\pi$ model of a submicron HBT shows good correlation with measured S-parameters, $h_{21}$, and $U$ in the DC-50 and 75-110 GHz bands. The model parameters are also consistent with measured bulk and sheet resistivities and junction capacitances. Base-width modulation in HBTs is negligible, hence $R_{ce}$ is very large. $C_{be,\text{poly}}$ is a metal-polyimide-metal overlap capacitance between the emitter and base contacts (sec. 10).

As previously discussed, the negative output conductance observed in recently fabricated submicron devices cannot be modeled with a standard hybrid-$\pi$ model. Additionally, we have found that device models developed in the 6-45 GHz band show poor agreement with measured device parameters in the 140-220 GHz band. Figure 20 shows measured and modeled $S_{11}$ and $S_{22}$ in the 6-45 GHz and 140-220 GHz bands for the transistor of fig. 19. $S_{12}$ and $S_{21}$ are omitted from the graph for clarity, but show similar discrepancies in the 140-220 GHz band. The poor agreement between model and measurements in the higher frequency band points to a weakness in extending the simple hybrid-$\pi$ model to these frequencies.

The results of fig. 20 are for a highly scaled device and the discrepancy may be due to collector-transport effects not included in the model. Model development for less highly scaled devices in the 140-220 GHz band have not been made at the time of this writing, and to the best of our knowledge no work has been done to characterize mesa HBTs in the 140-220 GHz band. It should be noted that the hybrid-$\pi$ model approximates to first order in frequency base and collector transit time effects. In contrast, the T-model (common-base) does not require such approximations. We note, however, that a T-model of the device, developed from 6-45 GHz measurements, also could not fit the measured S-parameters in the 140-220 GHz band.
Figure 19: Device equivalent circuit model at $V_{cc} = 1.2$ V and $I_c = 3.6$ mA.

Figure 20: Measured (solid line) and modeled (circles) $S_{11}$ and $S_{22}$ in 6–45 GHz and 140–220 GHz bands of device of fig. 19.
At the time of this writing, a physically justifiable small-signal device model has not been developed for the 140-220 GHz band. This further complicates estimations of transistor bandwidth, as we cannot predict the power gain roll-off versus frequency. Higher frequency amplifier and oscillator designs will also be limited by the lack of a predictive model.

6. HBT Amplifiers

6.1. Amplifier Design

Wideband analog amplifiers and high frequency tuned amplifiers have been fabricated in the transferred-substrate HBT technology. Prior to presenting amplifier results, we consider some of the design issues faced specifically in ultra-high frequency tuned amplifiers.

Because of process complexity and yield issues, first generation tuned amplifier designs have emphasized simple design strategies and low transistor counts. Circuit design is performed using Agilent’s Advanced Design System software. Small-signal transistor models are developed in-house using the procedures described in Section 5.1. Large-signal model development is more complex, and physically based models have been developed for power amplifier designs. As described in the previous section, poor correlation has been found between the small-signal hybrid-π model and measurements in the 140-220 GHz band. First generation designs in this frequency band were shifted from their intended design frequencies due to model discrepancies. Improved designs are now being developed based on measured transistor S-parameters.

Tuned-amplifier designs have typically utilized transmission line matching networks, with lumped passive elements for stabilization and biasing. At mm-wave design frequencies, electromagnetic simulation of passive elements is essential. A planar method-of-moments electromagnetic simulator is used to model critical passive elements and transmission line discontinuities. The transferred substrate technology provides a low-dielectric (\( \varepsilon_{ef} = 2.7 \)) microstrip wiring environment with a thin substrate thickness (5 μm). Standard microstrip CAD models have been found adequate to describe straight sections of transmission line to the frequency limits of our measurement system. The design approach of utilizing electromagnetic simulation of unique passive elements with standard microstrip models has shown excellent agreement with measurements. Figure 21 shows modeled and measured S-parameters for the matching network of the single-stage amplifier described in the next section. A matching network test structure without an active device was realized on-wafer, and the model and measurements show good agreement across the 140-220 GHz band.

The 5 μm BCB microstrip dielectric used in the transferred substrate process was selected to provide a low inductance, low cross-talk wiring environment for densely
packed mixed-signal IC applications. The thin dielectric also improves thermal heat-sinking, and provides low inductance access to the backside ground plane. In tuned amplifier design, these advantages are offset by the high resistive losses incurred in the transmission line matching networks. For the amplifier using the matching network of fig. 21, simulation of the circuit with lossless matching networks resulted a 2.0 dB increase in the gain. Given that resistive losses increase inversely with substrate thickness for a line of constant characteristic impedance, increasing the dielectric thickness may be beneficial for circuits utilizing transmission line matching networks.

6.2. Amplifier results

Figure 22 shows a chip photo and measured S-parameters of a single-stage G-band tuned amplifier. The transistor used in the design had an emitter junction area of 0.4 μm × 6 μm, and a collector stripe of 0.7 μm × 6.4 μm. The amplifier employed a simple common-emitter topology. Shunt-stub tuning at the input and output of the device was used to conjugately match the transistor at the intended design frequency. A shunt resistor at the output was used to ensure low frequency stability, and a quarter-wave line to a radial stub capacitor bypassed the resistor at the design frequency.

The amplifier exhibited a peak gain of 6.3 dB at 175 GHz, with a gain of better than 3 dB from 140 to 190 GHz. Both the input and output return losses were better than 10 dB at 175 GHz. The gain-per-stage of the amplifier is amongst the
Figure 22: Single-stage tuned G-band amplifier with 6.3 dB gain at 175 GHz.
Figure 23: W-band balanced medium-power amplifier (chip photo) and schematic for cascode power amplifier. The balanced amplifier has 7 dB gain and produces 10.7 dBm saturated output power at 78 GHz. The cascode amplifier had 8.5 dB gain at 75 GHz with a 1 dB gain compression output power of 9.4 dBm.
Figure 24: Distributed amplifier in the transferred-substrate process. The amplifier exhibits 11.5 dB gain and approximately 80 GHz bandwidth.
Figure 25: $f_T$-doubler resistive feedback amplifier with 8.2 dB low-frequency gain and a DC-80 GHz 3-dB-bandwidth
highest reported from any transistor technology in this frequency band. Multi-stage amplifier designs based on this first generation design are currently being fabricated. Simulation of a three-stage amplifier design predicts a peak gain of 20 dB at 175 GHz.

W-band HBT amplifiers are being developed in the transferred-substrate technology for phased-array antenna applications. First-generation designs utilizing the single heterojunction process were reported in (fig. 23). A cascode amplifier exhibited 8.5 dB gain at 75 GHz with a 1 dB gain compression output power of 9.4 dBm. A balanced amplifier employing two cascode stages had a gain of 7.9 dB at 78 GHz and a 1 dB compression output power of 10.7 dBm. Second generation power amplifier designs using the higher breakdown InP double heterojunction technology have recently been fabricated. A common-base power amplifier exhibited 8 dB gain and a saturated output power of greater than 16 dBm.

In addition to tuned amplifiers, broadband analog amplifiers for optical fiber receivers have also been fabricated. Results from this effort include: 80 GHz distributed amplifiers (fig. 24), 50 GHz differential amplifiers, and Darlington and $f_T$ - doubler resistive feedback amplifiers (fig. 25). Greater than 400 GHz gain-bandwidth product has been obtain from a single Darlington stage.

7. Conclusions

Extending transistor bandwidths towards terahertz frequencies requires aggressive device scaling. High bandwidths are obtained with heterojunction bipolar transistors by thinning the base and collector layers, increasing emitter current density, decreasing emitter contact resistivity, and reducing the emitter and collector junction widths. HBT amplifiers have been demonstrated in the 140-220 GHz band, and transistors show high levels of available gain at the frequency limits of state-of-the-art measurement systems. The physical characteristics of submicron HBTs make accurate measurement and modeling difficult. Next generation amplifier designs will require further scaling of minimum device feature sizes, and accurate modeling of active and passive circuit components.

Acknowledgments

Work at UCSB was supported by the ONR under grants N0014-99-1-0041, N00014-01-1-0066 and N00014-01-1-0024, by the AFOSR under grant F4962096-1-0019, and by the ARO under the Quasi-Optical MURI PC249806.

References


26. HP-EESOF Series IV microwave circuit simulation program. Hewlett-Packard Company, 3000 Hanover Street, Palo Alto, CA 94304, USA.


44. Advanced Design System. Agilent Technologies, 3000 Hanover Street, Palo Alto, CA 94304, USA

45. M. Urteaga, D. Scott, T. Mathew, S. Krishnan, Y. Wei, M.J.W. Rodwell, "185 GHz Monolithic Amplifier in InGaAs/InAlAs Transferred-Substrate HBT Technology", 2001 MTT-S International Microwave Symposium Technical Digest, May, Phoenix, AZ.


This page is intentionally left blank
THz Generation by Photomixing in Ultrafast Photoconductors

E. R. Brown
Professor of Electrical Engineering
University of California, Los Angeles
Los Angeles, CA 90095-1594, USA

This article reviews the fundamental concepts, technology, and recent developments on the topic of THz photomixers made from low-temperature-grown GaAs, ErAs:GaAs, and ErAs:InGaAs. It is designed to be comprehensive and useful both to students and to engineers and scientists new to the THz field.

Keywords: THz generation, ultrafast photoconductivity, interdigitated electrodes, planar antennas, tunable solid-state and semiconductor lasers.

I. Introduction

A. The challenge of solid-state THz sources

The THz region of the electromagnetic spectrum has long been the realm of basic sciences, such as molecular chemistry and astrophysics, but has not been broadly utilized for commercial or military applications because of longstanding problems in the generation of coherent radiation, and the difficulty in propagating this radiation through the lower atmosphere. To date the most powerful THz sources have been fundamental oscillators, such as molecular gas lasers, free-electron lasers, and certain types of vacuum tubes (e.g., backward-wave oscillators). As shown in the Fig. 1, the BWOs easily provide enough output power (> 1 mW) for laboratory sciences (e.g., molecular chemistry), but they are bulky and expensive and generally not considered to be suitable for portable systems.

As in other spectral regions, the most promising sources for portable systems are solid-state devices, of which many approaches have been proposed and developed over the past few decades. At the low end of this region (~0.1 to 0.2 THz), active monolithic microwave integrated circuits (MMICs) offer the best overall performance, and improvements are steadily occurring through advancements in GaAs- and InP-based heterostructure transistor technology. Of particular interest are solid-state power amplifiers (SSPAs) that can increase the power from a separate solid-state oscillator by 10 dB or more. The SSPA not only increases the power to levels not achievable from spectrally-pure oscillators, but it electrically isolates the oscillator from the ultimate load. At THz frequencies it is often difficult to control the load impedance, so isolation is often necessary to prevent oscillations that would otherwise occur.

On the long-wavelength end of the THz region up to about 300 GHz, fundamental-frequency oscillators have been developed including InP Gunn diodes and silicon IMPATT diodes. The highest reported output power from IMPATT diodes is
plotted in Fig. 1. The highest known fundamental oscillator frequency at room temperature of 712 GHz was achieved by the resonant-tunneling diode, although its output power at this frequency is below 1 μW. The curve clearly shows a trend of rapidly diminishing output power with frequency. There are several reasons for this, the primary one being that fundamental oscillators are limited to operate at or below a maximum oscillation frequency \( f_{\text{max}} \) that is determined by a variety of internal electrical time constants, carrier transit times, and parasitic (e.g., circuit loss) effects. These effects cause the gain (or magnitude of negative resistance) to start dropping at frequencies well below \( f_{\text{max}} \), and the power falls in kind.

At the upper end of the THz region (~3 to 10 THz) photonic techniques are quite promising, particularly lasers based on the quantum-cascade of electrons in multiple-quantum-well structures. As in other semiconductor-based lasers, the great challenge is to realize cw, room-temperature operation. Such lasers are often limited in performance by nonradiative relaxation mechanisms when the photon energy becomes comparable to the energy quanta of lattice vibrations (i.e., phonons), as occurs at the meV levels of THz photons. Reducing the device temperature alleviates this problem, but only up to the point were defect or carrier-carrier scattering (which are weakly dependent on temperature) takes over, and at the practical cost of cryogenic operation.

To avoid the difficulties faced by fundamental solid-state oscillators, researchers have long sought to produce coherent power in the THz region by harmonic multiplication of fundamental sources from the long-wavelength end, and nonlinear-optical (e.g., three-wave) mixing from the short-wave end. Of the two techniques, harmonic multipliers have been more successful and are now the most common solid-state coherent source up to frequencies around 1 THz. They provide the highest output power, although tunability is usually limited. The state-of-the-art in solid-state THz
Three-wave mixing entails the injection of two frequency-offset laser beams into an optical waveguide device filled with an asymmetric quantum-well structure or a nonlinear-optical material such as LiNbO$_3$. Because of the substantial $\chi^{(2)}$ susceptibility in the material, the two beams will produce a third beam at the difference frequency provided that all three have the same phase velocity. Unfortunately, three-wave mixing has never produced useful THz power levels in spite of the use of powerful drive lasers. The problems stem from the difficulty in achieving precise phase matching and in the fundamental limitation on the conversion efficiency imposed by the Manley-Rowe relation, which states that at most one photon can be produced at $|\omega_2 - \omega_1|$ for one photon each at $\omega_1$ and $\omega_2$. In other words, the best that three-wave mixing can ever do in terms of optical-to-electrical (O-E) conversion efficiency, $e$, is $\sim |\omega_2 - \omega_1| / 2\omega_1$, which is $\sim 0.2\%$ in going from 1 $\mu$m to 300 $\mu$m.

B. Photomixing in ultrafast photoconductors

The subject of this chapter is an alternative optical mixing technique similar to three-wave mixing but optoelectronic in nature. It is usually called photoconductive
mixing, or photomixing, because of its common application in photoconductors. As shown in the schematic diagram of Fig. 2, photomixing entails driving a high-speed photoconductive device with two frequency-offset lasers. The internal photoelectric effect, described further in Sec. III, generates an electrical photocurrent at the difference frequency. The difference-frequency current, in turn, generates THz power by connecting it to a suitable load, such as a planar antenna.

Three technological breakthroughs have occurred during the past decade that make photomixers a promising THz source alternative: (1) growth and fabrication of semiconducting material having photocarrier lifetime less than 1 ps, (2) modern microfabrication techniques that allow sub-micron electrode features to be patterned on the photoconductor surface, leading to sub-picosecond electrical time constants, and (3) integration of photomixer elements with compact planar antennas, leading to efficient coupling of the THz radiation to free space. A good example is the interdigitated-electrode photomixer coupled to a planar antenna, such as the dipole shown in Fig. 2.

An ancillary breakthrough that strongly supports the photomixer approach has occurred in the field of solid-state and semiconductor lasers. Solid-state materials, such as Ti:Al$_2$O$_3$, have been developed with unprecedented values of gain-bandwidth product so can provide high levels of power tunable over 10s of nm. Various optical techniques such as distributed Bragg reflectors, distributed feedback structures, and external cavities have all been integrated with semiconductor laser diodes to produce sources with useful output power (>1 mW) and high spectral purity. And in the popular fiber-optic telecomm band around 1550 nm, the erbium-doped fiber amplifier (EDFA) can boost the power of spectrally-pure laser-diode sources up to levels of 1 W or more.

A key advantage of photomixing over the other THz sources mentioned above is frequency tunability. Fig. 3 shows how little wavelength offset $\Delta \lambda$ is required to

---

**Fig. 3.** Difference frequency as a function of wavelength offset for drive lasers at two popular wavelengths – 780 and 1550 nm.
produce THz difference frequencies. The formula for this offset is simply $\Delta v = c|\Delta \lambda|/(\lambda_1 \lambda_2)$, which is approximately given by $c|\Delta \lambda|/(\lambda)^2$ for $\lambda_1 \approx \lambda_2$. One curve assumes $\lambda = 780$ nm near the center of the Ti:Al$_2$O$_3$ output power spectrum. The other curve assumes $\lambda = 1550$ nm near the center of the EDFA gain spectrum. Note that at 780 nm and 1550 nm, a 1 THz difference occurs for a $\Delta \lambda$ of 2.05 nm and 8.05 nm, respectively. Both offsets are readily achieved with modern tunable laser-diode technology. 

As shown in Fig. 1, the output power of photomixers coupled to planar antennas has been limited to a few $\mu$W below 1 THz, and progressively lower values at frequencies above 1 THz. One goal of this chapter is to present the analysis and design trade-offs that have been developed for photomixers on GaAs photoconductors - the best-understood photomixers to date. A second goal is to summarize recent research aimed at improving the output power and O-E efficiency by judicious changes of the design shown in Fig. 2(b). This includes reductions in the RC-rolloff, increases in the photoconductive gain, reductions in the thermal resistance, or some combination of the above. Two new design approaches will be addressed: (1) resonant-optical-cavity (ROC) photomixers with AlAs heat spreaders, and (2) distributed photomixers with optical coupling through dielectric waveguide.$^{10,11}$

A third goal is to address recent efforts to extend THz photomixing to In$_{0.53}$Ga$_{0.47}$As photoconductive material on InP substrates. This has the advantage of being compatible with spectrally-pure 1.55-$\mu$m drive lasers, such as the external-cavity diode laser (ECDL), and EDFA amplifiers. A related benefit is fiber-optic coupling of the drive lasers to the photomixer elements. Not only does this bring to bear a wealth of fiber-optic components and measurement capability, but it enables two-dim photomixer arrays. The array architecture could be similar to that used in microwave phased-array antennas - planar resonant antennas separated from their nearest neighbors by approximately $\lambda/2$, where $\lambda$ is the free-space THz wavelength.

II. Photomixer Technology

A. Materials and fabrication

The most successful photomixer materials demonstrated to date have been low-temperature-grown GaAs (LTG-GaAs) and ErAs-GaAs (ErAs:GaAs). The LTG-GaAs, which was first analyzed and applied to photomixers at Lincoln Laboratory in 1992,$^{12}$ is grown by molecular-beam epitaxy at temperatures around 200°C such that good crystallinity is maintained while incorporating approximately 1% more arsenic than gallium. With such deviation from stoichiometry, the photocarrier lifetime in as-grown material is very short (around 1 ps), but the electrical properties, such as the mobility, are rather poor. Annealing of the material at temperatures around 500°C causes much of the excess As to coagulate into nm-scale precipitates. Fortunately, the As precipitates maintain the short lifetime, but enhance the carrier mobilities to useful levels (>100
cm²/V-s), and increases the resistivity and electrical breakdown field significantly. This occurs in part because of the re-arrangement of As atoms on Ga sites. Such anti-site defects are associated with rather well defined defect states near the center of the GaAs band gap, so they support very rapid electron-hole recombination via the Shockley-Read-Hall mechanism, remain neutral at ambient temperatures, and do not impact ionize under high internal fields. Because of these remarkable effects, many researchers consider LTG GaAs a paradigm of defect engineering – an interesting branch of semiconductor materials science. A problem with LTG GaAs is the difficulty in controlling the growth temperature, the associated excess-As incorporation, and the photocarrier lifetime. MBE technology is simply not well suited to growth around 200°C.

To avoid the growth problems of LTG GaAs, researchers at the University of California, Santa Barbara, began around 1998 to investigate ErAs:GaAs – a composite material in which thin layers of ErAs are embedded in GaAs during growth. ErAs is a semi-metallic material, similar in many ways to As, that has a rocksalt crystal structure with a conventional lattice constant only ~1.6% different than GaAs. Typically, photomixer material is produced by depositing approximately 1 monolayer of ErAs during MBE growth of GaAs around 535°C. For such thin layers the ErAs self-assembles into nm-scale islands, and the GaAs growth proceeds with high crystallinity between the ErAs layers. Because of the semi-metallic nature of the ErAs, the nanoparticles tend to act as efficient nonradiative recombination centers, similar to the As precipitates in LTG GaAs. Hence, the material displays significantly higher mobility than LTG-GaAs and a photocarrier lifetime that depends on the time required for excess electron and holes to diffuse to ErAs nanoparticles.

After growth of the epitaxial layer, electrodes are deposited on the top surface by microfabrication. The most useful design has been the interdigitated electrodes consisting of narrow (~0.2 µm) metallic fingers varying in length between 4 and 20 µm.
The electron micrograph of a typical interdigitated structure is shown in Fig. 4. The gap spacing between neighboring fingers is a constant for a given structure that typically lies between 0.4 and 2.0 μm, but is always greater than the finger width to keep the metal fill-factor low and, thus, enable high optical coupling efficiency in the vertical direction. The fingers are patterned in polymethylmethacrylate (PMMA) resist by electron-beam lithography, the PMMA is developed chemically, and electrode metals are deposited by evaporation, and the metals on top of the undeveloped PMMA are lifted off by dissolution of the undeveloped PMMA.

The metal layers in the electrodes typically consist of ~50 nm of Ti and ~150 nm of Au – a recipe adopted mostly for fabrication reasons. The Ti has superior adhesion to GaAs than Au so that lift-off occurs with much greater yield. The Au is used because of its superior electrical and thermal conductivity. The Ti/Au metallization is deposited without any ion implant or thermal annealing step to drive the metal or dopants into the semiconductor. As explained in Sec. II.B below, this is based on the widespread experience that high-quality ohmic contacts are automatically obtained without such processing. Once the interdigitated-electrode fabrication is complete, a planar antenna is fabricated. The antenna type is an important part of the photomixer and will discussed further below. Note that the use of top-side electrodes and planar antennas generally require the growth and fabrication of photomixer structures on semi-insulating (SI) GaAs substrates. Any intentional doping in the substrate would likely short out the electrical current between the top-side electrodes and cause significant THz absorption in the planar antenna.

**B. DC electrical characteristics**

The interdigitated-electrode structure is one of the fundamental building blocks of microwave integrated circuits (MICs), having been introduced in the 1960s as a very compact planar capacitor. Its electrical characteristics are accurately predicted by electrostatics once the nature of the metal-semiconductor contact is defined. As in many other metal-semiconductor combinations, the analysis of the contact behavior is best done by first examining dc current-voltage (I-V) characteristics. The room-temperature I-V curves of most, if not all, LTG-GaAs and ErAs:GaAs photomixers tested to date are practically symmetric with respect to voltage and nearly linear out to a rather high voltage that depends on the gap width. This suggests that the contacts are non-rectifying, or ohmic, and that the measured resistance is determined primarily by transport in the bulk material. Note, however, that we can not deduce from a single I-V curve the contact resistance relative to the bulk resistance. These relative resistances are very important in photomixer operation. Note also that the non-ohmic behavior at high bias is rather gradual compared to the avalanching or Auger effects that occur in normal GaAs material at high bias. This is indicative of a soft breakdown effect whereby the current increases nonlinearly beyond a certain bias voltage, perhaps due to impact ionization, but avalanching does not occur.
The ohmic behavior of contact metals on LTG-GaAs or ErAs:GaAs may seem surprising until one recalls the nature of non-ohmic, or rectifying, contacts on GaAs. In most cases when a common metal is deposited on GaAs the contact is rectifying because of the Schottky barrier that occurs at the metal-semiconductor interface. The Schottky barrier results from the fact that the GaAs surface is usually defective, creating a high density of energy levels (i.e., density of states) deep in the band gap. In order that the Fermi level be flat between the bulk and surface, as required by thermodynamic equilibrium, the electron potential energy must bend, creating the Schottky barrier. In contrast to normal GaAs, LTG-GaAs and ErAs:GaAs both appear by experimental means to have an abundance of deep levels in the bulk, comparable in density to the surface states. Therefore, the Fermi level is thought to be approximately flat between the
bulk and surface, so that no Schottky barrier appears to electron transport from the bulk in to the metal. Note, however, that a barrier of roughly one-half the gap energy appears to electron transport from the metal to the GaAs. The transport in this direction can also be non-rectifying because electrons can proceed from the metal through mid-gap defect states in the GaAs by hopping or similar mechanisms. Again, it is not known what the resulting contact resistance is compared to the bulk resistance, which may involve both defect-related hopping and conduction-band drift processes.

The ramifications of bulk-Fermi-level pinning and ohmic contacts on photomixer performance are significant and often overlooked by researchers in the field. First, Fermi-level pinning makes the background free-carrier concentration very low and the associated dark resistance very high. As in all photoconductive effects, low background carrier concentration is important to maximizing the effect of photogenerated carriers. Second, very low background concentration means that stationary positively-charged impurities and defects will be balanced by an equal concentration of stationary negative charges, so that to first order the semiconductor is everywhere space-charge neutral. This means that the screening length in the material is very long and high bias voltage between the interdigitated electrodes can produce high electric fields deep in the bulk. Third, non-rectifying contacts mean that the device can possibly display photoconductive gain – a somewhat mysterious effect buried in the literature on photoconductivity. This requires that the contact resistance be much lower than the bulk resistance, meaning that the majority of bias voltage drops across the bulk.

C. Electrostatics and capacitance

As stated above the low free-carrier concentration in LTG GaAs and ErAs:GaAs imply that the screening length is long and that high electric fields can penetrate from the electrodes well into the bulk material. In interdigitated-electrode structures, the electric lines of force are non-analytic but simple to compute through numerical solution to Laplace’s equation, \( V \cdot E = 0 \), where \( E \) is the bulk electric field vector. Representative solutions are shown in Fig. 5 for a GaAs photomixer consisting of 0.25-\( \mu \)m-wide electrodes and a 0.8-\( \mu \)m-wide gap on a 1.5-\( \mu \)m-thick LTG-GaAs epitaxial layer. As expected intuitively, the largest electric field occurs at the surface and is approximated by \( E_{\text{max}} = V_B/W_G \) where \( W_G \) is the gap width. Fig. 5 shows the lines of force between two adjacent electrodes under a bias voltage that creates an \( E_{\text{max}} \) (which occurs at the bottom corners of each electrode) is just under \( 1 \times 10^6 \) V/cm.14
Given the lack of free-carrier concentration and assuming a uniform dielectric constant equal to the permittivity $\varepsilon_r$ of the background semiconductor, one can also calculate the electric potential distribution in the material and from that, estimate the capacitance. The most popular means of doing this is by conformal mapping techniques that result in the following useful expression for the capacitance:

$$C = \frac{K(k) \varepsilon_0 (1 + \varepsilon_r)}{K(k') W_E + W_G} A$$  \hspace{1cm} (1)

where $K$ is the complete elliptic function of the first type,

$$k = \tan^2 \left[ \frac{\pi W_E}{4(W_E + W_G)} \right]$$ \hspace{1cm} (2)

and $k' = (1-k^2)^{1/2}$, and $W_G$ is the gap width. In Eqn (1), $A$ is the active area assuming a square geometry, so that $A \approx L_E \left[ N_E W_E + (N_E - 1) W_G \right]$ where $L_E(N_E)$ is the electrode length (number). The elliptic function can be approximated for small and large $k$ by

$$\frac{K(k)}{K(k')} \approx \frac{\pi}{\ln[2(1 + \sqrt{k'})/(1 - \sqrt{k'})]}, \quad 0 < k < 1/(2)^{1/2}$$ \hspace{1cm} (3)

and

$$\frac{K(k)}{K(k')} \approx \frac{\ln[2(1 + \sqrt{k})/(1 - \sqrt{k})]}{\pi}, \quad 1/(2)^{1/2} < k < 1$$ \hspace{1cm} (4)

An important assumption behind (1) is that the total dielectric thickness is much greater than the electrode width $W_E$ or the gap width $W_G$. This is satisfied in most photomixer structures studied to date because their fabrication on SI substrates means that electrostatic calculations should not, to first order, differentiate between epitaxial and substrate material. Since GaAs SI substrates are typically 400-to-500 $\mu$m thick and $W_G$ is usually less than 2 $\mu$m, the accuracy of (1) should be very good.

An important practical consideration in all interdigitated photomixers is that $W_G$ is generally made greater than $W_E$ to obtain high external quantum efficiency (see Sec. II.D). Hence according to (2), $k << 1$ and the approximation of (3) becomes most useful. It is then clear that $K(k)/K(k') << 1$ and $C < \varepsilon_0 \varepsilon_r A/W_G \equiv C_p$, where $C_p$ is the equivalent parallel-place capacitance. This justifies the conventional wisdom that interdigitated-electrode structures provide low specific capacitance along with their planarity and relative ease of fabrication.
D. Laser drive and coupling

Since photomixing entails the injection of two frequency-offset lasers into the active region of an interdigitated-electrode structure, two key issues are the laser type and wavelength, and the optical coupling to the photomixer active region. The laser quality can be a critical issue in photomixing in applications where spectral purity and wide tunability are of utmost importance. The first laser-type to achieve useful THz results with LTG GaAs photomixers were Ti:Al$_2$O$_3$ solid-state lasers operating around 780 nm. Besides being the first solid-state laser with spectral tuning of at least 50 nm (limited only by the mirror set), Ti:Al$_2$O$_3$ lasers can readily produce power levels in excess of 100 mW and can generate an output at a single frequency in a single spatial mode if operated in the ring-cavity configuration. The next laser to produce useful THz results was the semiconductor (GaAs/AlGaAs) diode operated in a distributed-Bragg-reflector (DBR) cavity. A third possibility for driving GaAs photomixers is the master oscillator power amplifier (MOPA) laser operating around 780 nm. While lacking the tunability of the Ti:Al$_2$O$_3$ laser, the MOPA produces comparable power in a more compact package that includes the pump laser.

The primary goal of the photonic coupling process is to get as many of the photons as possible absorbed in the semiconductor without reflection from the electrode metal or semiconductor-air interface. Since the electrodes have the form of a two-dimensional grating, the first consideration in optical coupling is polarization. As in all gratings, the extinction is greatest when the polarization of the incident photons is parallel to the lines of the grating. Hence the photons from both lasers used to drive a photomixer are generally oriented to be perpendicular to the electrodes on incidence. In this case, the photonic coupling factor $T_p$ from free space into the active region can be estimated as

$$T_p \approx \frac{W_G T_s}{W_E + W_G}$$

where $T_s$ is the intensity transmission coefficient through the semiconductor-air surface. According to Fresnel's equation, $T_s = 4n/(n+1)^2$ where $n$ is the wavelength-dependent LTG-GaAs refractive index. For example, around $\lambda = 0.8 \mu$m, $n = 3.77$ and $T_s \approx 0.66$. So assuming the parameters $W_E = 0.2 \mu$m and $W_G = 1.0 \mu$m of the typical interdigitated electrode structure, we find $T_s \approx 0.66$ and $T_p \approx 0.55$.

E. Photomixing in interdigitated-electrode structures

The majority of THz photomixers demonstrated to date have been fabricated from closely-spaced (< 2 \mu m) interdigitated electrodes, as shown in Fig. 2(a). The gap between neighboring electrodes is made comparable to or larger than the electrode width so that a majority of light incident from the top side enters the photoconductive material. Strong photomixing can occur if the incident light consists of two frequency-offset lasers,
such as GaAs/AlGaAs laser diodes. Then the quadratic nature of the cross-gap absorption with respect to the electric field, discussed further in Sec. III.A., mixes the fields and creates an ac photocurrent between the electrodes oscillating at the difference frequency between the lasers.

The ac photocurrent is usually transformed into THz power by coupling the electrodes to a low-loss antenna, such as the resonant dipole shown in Fig. 2(b). According to the principles discussed in Sec. III, the THz output power is limited by the external quantum efficiency of the photomixer, the photocarrier recombination time, circuit-related (e.g., RC) rolloff of the photomixer circuit, and by deleterious heating effects that occur with high optical drive intensity. Note that the photomixer behaves in some ways like a field-effect transistor (FET), but with a photonic gate instead of the usual metal gate. So unlike an FET, the bandwidth of the photomixer is not limited by the gating effect because optical coupling requires no metal electrode and, therefore, adds no capacitance.

As shown in Fig. 1, the output power of photomixers coupled to planar antennas has been limited to just under 1 μW below 1 THz, and progressively lower values at frequencies above 1 THz. These power levels have been obtained with total optical pump power of ~50 mW and interdigitated structures having electrode widths of ~0.2 μm, gap widths of ~1.0 μm, and total active area of ≈50 μm². These dimensions are a result of several engineering trade-offs to be discussed at length in this chapter. If the gaps are made much smaller to achieve high photoconductive gain, the capacitance gets large and limits the power and O-E efficiency at sub-THz frequencies through RC rolloff, where R is antenna radiation resistance. If the gaps are made much larger to reduce the capacitance, the inverse transit time and associated photoconductive gain g decay, limiting the output power and O-E efficiency at moderate drive power levels. If the area is made much larger to allow for more optical pump power, the capacitance increases and causes RC roll off well below 1 THz. If the area is made much smaller, burnout invariably occurs because of the high junction temperature that results from the combination of optical drive and electrical bias.

Besides its small area and low specific capacitance, another key benefit of the interdigitated-electrode photomixer is that it naturally creates a balanced current feed for simple planar antennas such as dipoles, slots and a class of self-complementary spirals. No balun or similar circuit is required. The spirals are particularly useful because of their inherently wide instantaneous bandwidth (~1 decade or more). If judiciously aligned with the operating band of the photomixer, the spiral antenna will combine with the photomixer to create a THz free-space source with exceptionally high tuning bandwidth – a THz sweep oscillator. Such a source has never existed in the THz region and, as in lower frequency bands, could be a strong enabler for scientific and technological purposes.
THz Antennas

The first antennas to be developed for experimental purposes were planar spirals. Perhaps the most successful planar spiral to date is the self-complementary logarithmic spiral shown in Fig. 6. It consists of two metallic arms connected to opposite ends of the interdigitated electrodes. The edge of each arm emanates out on a locus given by $r_0 e^{\phi}$, and each edge is rotated from its nearest neighbor by 90°. As true for all self-complementary planar antennas, the log spiral displays special electromagnetic characteristics, namely a driving-point impedance that is real and independent of frequency over a wide operational bandwidth (determined by the inner and outer radii). For the two-turn spiral shown in Fig. 6, this bandwidth can easily by one decade. The value of the driving-point resistance is given by

$$R_A = \frac{60\pi}{(\varepsilon_{\text{eff}})^{1/2}}$$  \hspace{1cm} (6)$$

where $\varepsilon_{\text{eff}}$ is the effective dielectric constant given by $(1 + \varepsilon_t)/2$. For GaAs, $\varepsilon_t \approx 13.0$ so that $R_A = 72 \Omega$. Such a spiral has produced measurable power from a photomixer up to 3.8 THz. Another remarkable feature of self-complementary spirals is their insignificant load susceptance $B_L$ over the bandwidth.
To get more power from a given photomixer, but over a limited bandwidth, it is possible to couple it to a resonant antenna, such as a planar dipole or slot. The increase in power stems from the following two facts. First, if such an antenna is operated slightly away from its intrinsic resonant frequency, it may provide a high enough $B_L$ to cancel the capacitive reactance of the photomixer. Second, the antenna load resistance may be high enough that even when operated at a frequency away from resonance, the radiated power will be significantly higher than that from a broadband antenna.

A good example of a useful planar antenna is the full-wave dipole. It’s driving-point impedance on a semi-infinite GaAs half-space is shown in Fig. 7. Note that its resonant resistance is 250 $\Omega$ - over three times greater than that of the self-complementary spiral.

G. Coupling to free space

Independent of the antenna type, a significant problem with any antenna-coupled photomixer on a GaAs substrate is coupling out the radiation to free space. Eqn (6) assumes that the antenna is located on a half-space of dielectric constant $n_s = (\epsilon_r)^{1/2}$. An elegant and useful solution is to do this coupling through a back-side hemispheric lens made out of a high-$n$ material, as shown schematically in Fig. 8. This technique was pioneered by D. Rutledge’s group at Caltech. A good choice for the lens material is a low-cost, high-resistivity semiconductor, such as float-zone-grown silicon. If the photomixer antenna is located at the center of curvature of the lens and the lens is thick enough so that the spherical surface is in the far-field of the antenna, then all the radiation from the antenna reaches the spherical surface at normal incidence and passes in to free space with a reflection coefficient of $(n_L-1)^2/(n_L+1)^2$. This is a great improvement over the slab substrate, but yields a highly diverging beam from planar
antennas having large beam solid angle. In such cases a common practice is to locate the antenna behind the center of curvature at a point where the radiation will be refracted in the forward direction. If the hemispherical surface and set-back are fabricated in the same dielectric material, the resulting optic element is called a hyperhemispherical lens.

If \( n_L \) is perfectly matched to \( n_S \) and the antenna is located behind the center of curvature by a distance \( r/n \), then the lens focusing is aplanatic. This is a property from geometric optics which states that all rays from a given point in the object plane (in this case, the planar antenna) are refracted to a parallel bundle after the lens, as shown in Fig. 8. Clearly, most planar antennas can not be accurately analyzed by geometric optics. A treatment of the focusing problem by modal (Gaussian-beam) analysis shows that the optimum set-back is just short of \( r/n \), as might be expected.

Other free-space coupling techniques have been considered, such as end-fire antennas. But as of the time of this composition, none has yielded performance competitive to that of the hyperhemispherical lens.

**H. Measurement of THz power**

Measurement of power in the THz region is a surprisingly difficult task, particularly at microwatt levels and below. As in the lower millimeter-wave region, thermal detectors are often the most accurate detector, although not the most sensitive. This is because of their relative simplicity and ability to efficiently absorb the radiation. The practical requirement on any thermal detector is the sensitivity be good enough to detect the very low THz output that occurs when the photomixer is first packaged, the drive lasers are just being coupled, and the THz output optics are not yet well aligned.
Nearly every experiment starts in this state, and the power that reaches the detector is roughly 1 nW. If the nW level can be measured with confidence, then all the alignments can be made quickly by maximizing the THz signal.

The best candidate for cryogenic operation is the composite bolometer operating at 4.2 K (liquid helium). This device consists of a thin absorbing film in an integrating cavity on which is mounted a small Si (or perhaps some other semiconductor) thermistor having a high thermal coefficient of resistance, \((1/R)\, dR/dT\). The key advantage of cryogenic over room-temperature operation is reduction in the thermal noise while maintaining high responsivity. Around 1 THz the 4.2-K Si composite bolometer can provide an optical NEP of roughly \(1 \times 10^{-12} \text{ W-Hz}^{-1/2}\). So in the typical bandwidth of 10 Hz, the minimum detectable power is \(~10\, \text{pW}\). The disadvantage of the composite bolometer compared to the Golay cell is uniformity of response. Because of standing waves that can occur in an integrating cavity, the spectral variation in responsivity can be a factor of ten or more — a fact not appreciated by many people in the field.

III. Photomixing Theory

A. Photonic mixing

From fundamental semiconductor physics we know that the generation of electron-hole pairs by cross-gap photon absorption can be described semi-classically with the carrier wave functions expressed as quantum-mechanical (e.g. Bloch) wave functions and the electromagnetic energy expressed through the classical Poynting vector or, equivalently, the intensity. What results (from Fermi's golden rule) is an expression for the photocarrier generation rate \(g(t)\)

\[
g(\vec{r}, t) = \frac{\alpha l(\vec{r}, t)}{h\nu} = \frac{\alpha E^2(t)}{Z_0 h\nu}
\]

where \(\alpha\) is the absorption coefficient, \(E\) is the time-dependent optical electric field measured in the medium at the point \(\vec{r}\) where absorption occurs, \(Z_0\) is the intrinsic impedance of the medium, and \(h\nu\) is the optical photon energy. Note that the (MKS) units for \(\alpha\) are \(\text{m}^3\cdot\text{s}^{-1}\). The expression \(\alpha\) is sometimes called photoelectric law, first deduced by Einstein.

In the visible part of the electromagnetic spectrum or anywhere nearby, \(E(t)\) oscillates with period or order 10 fs or less. This is so much less than the typical transport times in semiconductors that the oscillating \(g\) can not be detected by electrical means. But suppose that the electric field consists, by linear superposition, of two independent optical fields \(E_1\) and \(E_2\). And for practical purposes, suppose that these fields are sinusoidal in time (e.g., the coherent state from a laser). We can then write
\[ g(\vec{r},t) = \frac{\alpha[\mathcal{E}_1(t) + \mathcal{E}_2(t)]^2}{Z_0 h v} = \frac{\alpha[\mathcal{E}_1 \cos \omega_1 t + \mathcal{E}_2 \cos(\omega_2 t + \phi)]^2}{Z_0 h v} \]  

(8)

\[ g_e(\vec{r},t) = \frac{\alpha}{Z_0 h v} \left[ E_1^2 [1 + \cos 2\omega_1 t] / 2 + E_2^2 [1 + \cos 2(\omega_2 t + \phi)] / 2 + \right. \]

\[ E_1 E_2 \cos(\Delta \omega t - \phi) + E_1 E_2 \cos((\omega_1 + \omega_2) t + \phi) \]

(9)

where \( \phi \) is the phase difference between the two fields and \( \Delta \omega = \omega_1 - \omega_2 \). If \( \Delta \omega \) is small enough to be much less than the electrical bandwidth of the external circuit, then only two terms are significant in the \textit{electrically-active} generation rate \( g_e \)

\[ g_e(\vec{r},t) = \frac{\alpha}{Z_0 h v} \left[ E_1^2 / 2 + E_2^2 / 2 + E_1 E_2 \cos(\Delta \omega t - \phi) \right] \]

(10)

where \( I_1 \) and \( I_2 \) are the time-averaged intensities of each field at the point of absorption. The third term of ( ) is the beat-note or difference-frequency term, which can comprise a significant fraction of \( g_e \) if \( I_1 \) and \( I_2 \) are comparable. Given a total optical intensity \( I_T = I_1 + I_2 \), it is simple to prove that the division between \( I_1 \) and \( I_2 \) that maximizes the contribution to \( g_e \) is simply \( I_1 = I_2 \). Clearly \( \alpha \) is an important parameter. For example in LTG GaAs at \( \lambda = 0.85 \) \( \mu m \), it is found \( \alpha = 10,000 \) \( \text{cm}^{-1} \).\[122\]

\section*{B. Optoelectronic transfer function}

A second aspect of photomixing and photoconductive detectors, in general, is the connection between photogeneration and electrical current – an issue that is often glossed over in textbooks, so will be addressed in some detail here. We assume that the photogeneration occurs in the presence of a static electric field that quickly separates the electron-hole pair and accelerates each in opposite directions. The electric field is created by two separate electrodes fabricated on or within the absorbing medium and having a potential difference \( V_B \) between them. As for interdigitated structures, if the electrodes are not parallel plates, the electric field is not uniform and the path followed by each carrier toward the appropriate electrode depends on where the photocarrier was generated. And because each photocarrier is generated with zero or near-zero velocity, there will be a dynamic dependence of the velocity that is universal in semiconductors and makes photocarrier collection a complicated issue.

Since the acceleration of free carriers is given by \( eE/m^* \), there is a field-dependent acceleration time during which the velocity increases to a maximum value \( v_{e,\text{max}} \) or \( v_{h,\text{max}} \), that depends critically on the carrier type, the specific semiconductor, and
the temperature. For example in GaAs at room temperature, it is known that $v_{e,\text{max}} \approx 2 \times 10^7 \text{ cm/s}$ and $v_{h,\text{max}} \approx 7 \times 10^6 \text{ cm/s}$.\textsuperscript{23} After reaching the maximum, both carrier types begin to scatter strongly with phonons and approaching a thermal steady-state in which they give up more energy to the lattice than they gain by acceleration between scattering events. So the carrier velocity saturates at a value that depends on the carrier type, the semiconductor, the temperature, and the electric field, and can be written analytically as

$$v = \frac{\mu E}{1 + \mu E/v_{\text{sat}}}.$$ \textsuperscript{24} For GaAs at room temperature and for $E > 1 \times 10^5 \text{ V/cm}$, one has $v_{e,\text{sat}} \approx 0.7 \times 10^7 \text{ cm/s}$ and $v_{h,\text{sat}} \approx 0.6 \times 10^7 \text{ cm/s}$.\textsuperscript{25}

To simplify the analysis of photomixers, we represent the complicated spatial and dynamic behavior of photocarrier collection in the following way. First, we assume that the path followed by each photocarrier toward the electrodes is along the electric line of force on which it is first generated. Hence each photocarrier can be associated with a transit length $L$ that is the path integral along the line of force and that depends on the point $\vec{r}$ in the structure where the photocarrier is generated. Second, we assume that each photocarrier type can be represented by an average velocity, $v_e$ or $v_h$, chosen in the range $v_{e,\text{sat}} < v_e < v_{e,\text{max}}$ and $v_{h,\text{sat}} < v_h < v_{h,\text{max}}$. With these two assumptions, each photocarrier can be associated with a transit time, $T_e$ or $T_h$, given by $L(\vec{r})/v_e$ and $L(\vec{r})/v_h$, respectively. In materials as complicated as LTG GaAs and ErAs:GaAs, the validity of these assumptions is very difficult to assess so is best tested by agreement between modeling and experiment.

Given these assumptions, analysis of the external electrical current in photomixers proceeds in the following way. We note from quantum mechanics that the process of photon absorption behind $g(r,t)$ is generally very short — on the order of a few femtoseconds. So we can approximate it as being instantaneous, and then the external current from a single photon absorption represents the optoelectronic impulse response, $h(t)$. According to the Ramo-Shockley theorem, the drift of each photocarrier towards the electrodes must induce a current in the external circuit to conserve overall energy.\textsuperscript{26} If the velocity for both carriers is saturated, the external current immediately after photoabsorption at time $t=0$ takes on the form $i_e(t) = (e/T_e)\theta(t)$ for the electron and $i_h(t) = (e/T_h)\theta(t)$ for the hole, where $\theta$ is the unit step function. Once generated, the photoelectrons and photoholes are assumed to recombine exponentially with lifetime $\tau_e$ and $\tau_h$, respectively.\textsuperscript{27} Thus, the time-dependent optoelectronic impulse response function takes on the form,

$$h(\vec{r},t) = [(e/T_e)\exp(-t/\tau_e) + (e/T_h)\exp(-t/\tau_h)]\theta(t) \quad (11)$$

From general principles of linear system theory, the differential output electrical current can now be found by convolving the input generation rate $g_e(t)$ with $h(t)$ so that
$d i(\vec{r}, t) = \int_0^\infty g_e(\vec{r}, t') \delta(\vec{r}, t' - t) dt'$

$$= \int_0^\infty g_e(\vec{r}, t') \left( \frac{e}{T_e} \exp[-(t' - t) / \tau_e] + \frac{e}{T_h} \exp[-(t' - t) / \tau_h] \right) dt'$$

(12)

We can simplify the analysis somewhat by re-writing $g_e(\vec{r}, t)$ in phasor form $g_e(\vec{r}, t) = g_0 + \text{Re}\{ g e^{j \omega t} \}$ so that $g_0 = (\alpha/hv)(I_1 + I_2)$ and $g$, a complex constant, is given by $(\alpha/hv)2(I_1I_2)^{1/2}e^{-j\phi}$. Straightforward integration of (12) then leads to the photomixing short-circuit current

$$d i(\vec{r}, t) = eg_0 \left( \frac{\tau_e}{T_e} + \frac{\tau_h}{T_h} \right) + e \frac{\tau_e}{T_e} \text{Re} \left\{ \frac{\tilde{g} e^{j \omega t}}{1 - j \Delta \omega \cdot \tau_e} \right\} + e \frac{\tau_h}{T_h} \text{Re} \left\{ \frac{\tilde{g} e^{j \omega t}}{1 - j \Delta \omega \cdot \tau_h} \right\}$$

(13)

Completion of the Re{} operator in (13) leads to the physically measurable result

$$d i(\vec{r}, t) = eg_0 \left( \frac{\tau_e}{T_e} + \frac{\tau_h}{T_h} \right) + e \tilde{g} \left\{ \frac{\tau_e \cos(\Delta \omega t - \phi) - \Delta \omega \tau_e \sin(\Delta \omega t - \phi)}{1 + (\Delta \omega \tau_e)^2} + \frac{\tau_h \cos(\Delta \omega t - \phi) - \Delta \omega \tau_h \sin(\Delta \omega t - \phi)}{1 + (\Delta \omega \tau_h)^2} \right\}$$

(14)

Inspection of the limiting behavior of Eqn (14) as $\Delta \omega \to 0$ indicates that both the dc and ac terms are multiplied by the factors $\tau_e/T_e$ and $\tau_h/T_h$. In principle, these factors can be significantly greater than unity, and thus are traditionally called the photoconductive gains $G_e$ and $G_h$. Physically if $G_e$ or $G_h > 1$, then more than one electron or hole can be delivered to the external circuit for each photon absorbed by the detector. This is not a violation of the laws of thermodynamics. Any difference between the power delivered to the load resistor and the incident optical power is provided by the required bias supply. This is an important distinction between ohmic MSMs and PiN photodiodes (or Schottky MSM detectors), which do not require external bias. On the other hand, PiN photodiodes and Schottky MSM detectors can not provide a photoconductive gain greater than unity under any condition.
C. Integrated photocurrent

It is important to realize that the expression in ( ) describes only the current that arises from photogeneration at point \( \vec{r} \) in the active region. To obtain the total electrode current \( i_t \), we must integrate \( i(\vec{r}, t) \) over the active volume \( V \)

\[
i_t = \int_{V_{nr}} d\vec{r} \int_{V_{nr}} \frac{Ge}{1 - j\Delta\omega \cdot \tau_e} + G_h \frac{Ge}{1 - j\Delta\omega \cdot \tau_h} dV
\]

(15)

For the case of interdigitated electrodes, this integral is complicated by the fact that \( g_0 \), \( \tilde{G} \), \( G_e \) and \( G_h \) all depend on the location in the active region where \( dV \) occurs. Two approximations can be made to make the volume integral tractable and thus useful for design purposes. The first is that the photons are absorbed according to the Lambert-Beer law so that the intensity at a depth \( z \) and any lateral position in the material has the form \( I(z) = I_0 \cdot T_P \cdot e^{-\alpha z} \), where \( I_0 \) is the incident intensity. This would be precise in a uniform absorbing region, but the electrodes will certainly create shadowing effects that cause dependence of \( I(z) \) on lateral position. In general, we expect the accuracy of this to improve as the ratio of \( W_E/W_G \) decreases.

The second approximation is that photocarriers, once generated, drift to the electrodes along the same line of force as the one on which they were generated. We can then quantify the electron and hole transit times \( L_e/\nu_e \) and \( L_h/\nu_h \) according to the path length for each line of force. Since according to the boundary conditions on Maxwell's equations the electric field at the electrodes must be perpendicular to the to the surface, we recognize that the lines of force in Fig. (5) are approximately half-elliptical. From geometry we can estimate the perimeter of each half-ellipse according to the expression \( L = \pi \frac{(W/2)^2 + z^2}{2}^{1/2} \) where \( W = W_E + W_G \) and \( z \) is the depth of the ellipse midway between the electrodes. Combining this with the Lambert-Beer law leads to a simplification of Eqn (15) to produce the expression

\[
i_t = \frac{e\alpha T_p}{hv} \cdot \frac{v_e \tau_e + v_h \tau_h}{(I_{1,0} + I_{2,0})} \left\{ \frac{v_e \tau_e}{1 - j\Delta\omega \cdot \tau_e} + \frac{v_h \tau_h}{1 - j\Delta\omega \cdot \tau_h} \right\} \]

(16)

where

\[
J = \int_{V_{tot}} \frac{\sqrt{2} \exp(-\alpha z)}{\pi[(W/2)^2 + z^2]^{1/2}} dV = A \int_0^\infty \frac{\sqrt{2} \exp(-\alpha z)}{\pi[(W/2)^2 + z^2]^{1/2}} dz = AK
\]

(17)
where $A$ is the active area and $K$ is a dimensionless depth integral.

It is customary to normalize this integral with respect to the total average current that would occur if all the incoming photons were absorbed and produced exactly one charge carrier in the external circuit. In this case one would have $i_0 = eP_{\text{in}}/hv = e(I_{\text{1,0}} + I_{\text{2,0}})A/hv = e(P_{\text{1,0}} + P_{\text{2,0}})/hv$, where $P_{\text{in}}$ is the total average incident optical power. This leads to the expression

$$i_0 = \frac{e(P_{\text{1,0}} + P_{\text{2,0}})}{hv}\{\eta + \Re\left[\frac{2(P_{\text{1,0}} \cdot P_{\text{2,0}})^{1/2}e^{-j\Delta \omega t}e^{j\omega t}}{P_{\text{1,0}} + P_{\text{2,0}}}\left(\frac{\eta_e}{1 - j\Delta \omega \cdot \tau_e} + \frac{\eta_h}{1 - j\Delta \omega \cdot \tau_h}\right)\]\}$$

(18)

where $\eta$ is the external quantum efficiency,

$$\eta = \alpha T_p K[v_e \tau_e + v_h \tau_h] = \eta_e + \eta_h$$

(19)

and $\eta_e$ and $\eta_h$ are the electron and hole contributions. Dividing the dc term of Eqn (18) by $P_{\text{in}}$ leads to the optical responsivity,

$$S = \frac{e\eta}{hv}$$

(20)

Eqns (18)-(20) are very useful in photomixer design and optimization.

Shown in Fig. 9 is the plot of $\eta$ as a function of $W_G$ in the range of 0.1 to 2.0 micron. The plot is parametrized by $\alpha$ in the range of 2,000 to 20,000 cm$^{-1}$. All other parameters are fixed as follows: $W_E = 0.2$ micron, $v_e = 1 \times 10^7$ cm/s, $v_h = 1 \times 10^7$ cm/s, $\tau_e = \tau_h = 1$ ps, $T_S = 0.66$. These parameters apply to a typical LTG-GaAs photomixer. Not surprisingly, each curve shows a maximum that moves to higher $\eta$ and lower $W_G$ as $\alpha$ increases. For example, at $\alpha = 10,000$ cm$^{-1}$, the expected value at a drive $\lambda$ of 0.78 um, we find $\eta_{\text{max}} = 0.052$ at $W_G = 0.40$ um. The presence of a maximum makes sense physically since for $W_G \ll W_E$, most of the incident light is blocked by the electrodes, and for $W_G \gg W_E$ the photoconductive gain for both carrier types drops rapidly, independent of where they are generated. As $\alpha$ increases, more and more photocarriers will be generated near the surface, so that narrower gaps become beneficial from the standpoint of photoconductive gain.

From the THz generation standpoint, the most important aspect of Eqn (18) is the time-varying term which can be written in phasor form as $i(t) = \Re\{i e^{j\omega t}\}$, where
Fig. 9. Plots of the quantum efficiency of an interdigitated-electrode structure as a function of gap dimension assuming ohmic contacts and constant electron and hole velocities equal to $0.6 \times 10^7 \text{ cm/s}$.

\[
\tilde{i}(\Delta \omega) = \frac{2e}{h\nu} \left( (P_{1,0} \cdot P_{2,0})^{1/2} e^{j\phi} \frac{\eta_e}{1 - j\Delta \omega \cdot \tau_e} + \frac{\eta_h}{1 - j\Delta \omega \cdot \tau_h} \right)
\]  

(21)

From fundamental circuit theory, the magnitude of Eqn (21) can be considered as the maximum short-circuit current that the photomixer can deliver to any load,

\[
i_s = \sqrt{\tilde{i} \cdot \tilde{i}^*} = \frac{2e}{h\nu} \left( (P_{1,0} \cdot P_{2,0})^{1/2} \frac{\eta_e^2}{1 + (\Delta \omega \cdot \tau_e)^2} + \frac{\eta_h^2}{1 + (\Delta \omega \cdot \tau_h)^2} \right)
\]

\[
+ \frac{2\eta_e \eta_h [1 + (\Delta \omega)^2 \tau_e \tau_h]}{[1 + (\Delta \omega)^2 \tau_e \tau_h]^2 + [\Delta \omega(\tau_e - \tau_h)]^2}^{1/2}
\]

(22)

The last term in this expression is a cross contribution from electrons and holes. It results from the fact that the drift currents from a given electron-hole pair are fully correlated.
through the simultaneity of generation. Note that when \( \tau_e = \tau_h = \tau \) and \( \eta_e = \eta_h = \eta/2 \), this term is equal to the sum of the previous two, and we get

\[
i_s = \frac{2e}{h\nu} \left( P_{1,0} \cdot P_{2,0} \right)^{1/2} \left[ \frac{\eta^2}{1 + (\Delta \omega \cdot \tau_e)^2} \right]^{1/2}
\]  

(23)

And if we maximize this with respect to the total optical power \( P_0 \), we find \( P_1,0 = P_{2,0} = P_{in}/2 \), and

\[
i_s = \frac{\eta P_{in}}{h\nu} \left[ \frac{1}{1 + (\Delta \omega \cdot \tau_e)^2} \right]^{1/2} \equiv SP_{in} \left[ \frac{1}{1 + (\Delta \omega \cdot \tau_e)^2} \right]^{1/2}
\]  

(24)

From Eqn (24) we see that the magnitude of ac photomixer current at \( \Delta \omega = 0 \) is exactly equal to the dc photocurrent. So the dc responsivity is a good indicator of the THz photomixer performance. For example, given the parameters used in Fig. 9, \( \eta_e = \eta_h \), and \( \eta_{\text{max}} = 0.052 \), \( h\nu = 1.59 \text{ eV} \) that \( S = 0.033 \text{ A/W} \). We also see that the 3-dB bandwidth \( B \) is defined simply by \( \Delta \omega \tau_e = 1 \) or \( B \equiv \Delta f = (2\pi \tau_e)^{-1} \). So the product of the external quantum efficiency and \( B \) becomes

\[
\eta \cdot B = \frac{\eta}{2\pi \tau_e} = \frac{\alpha T_p K_{V_e}}{\pi}
\]  

(25)

which is independent of photocarrier lifetime. This expression thus serves as a figure-of-merit for photomixers that includes optical, electrostatic, and semiconductor-transport effects. For example, in the GaAs photomixer of Fig. (9) and at the drive wavelength of 0.78 micron, we find \( \eta B = 9.1 \text{ GHz} \).

D. Thermal effects

Like most photonic-based sources the photomixer is limited in ultimate performance by thermal effects that must be considered in any design exercise. The two primary sources of heat are the optical power absorption and the Joule heating from photocurrent flowing in a bias field. THz photomixers generally operate under conditions that simplify the analysis of these sources. First, for reasons that will become clear shortly, THz photomixers are generally designed with sub-picosecond \( \tau_e \) and \( \tau_h \). This means that the vast majority of photocarriers recombine nonradiatively and that the heat from photon absorption is approximately \( P_{in} \). Second, in order to get the most efficient collection of photocarriers generated deep in the material, the photomixers are generally operated near the maximum possible bias voltage \( V_B \) between electrodes,
which can be approximated by \( E_B W_G \), where \( E_B \) is the breakdown field in the material. Hence, the Joule heating term can be approximated by \( I_P V_B \) and the total heat is

\[
P_Q \approx P_{in} + S P_{in} V_B = P_{in} \left( 1 + \frac{e \eta}{h \nu} E_B W_G \right)
\]  

(26)

Given this heat generation, the temperature of the photomixer can be estimated as follows. To maintain the low capacitance required for THz operation, \( A \) is typically made about 100 square microns or less. Thus the largest lateral dimension is typically 10 microns or less, which is generally much less than the thickness of the substrate (\( \approx 200 \) micron) on which the photomixer is fabricated and always less than the width (~ 1 mm or greater) of the photomixer die. In the ultrafast semiconductors (e.g., LTG GaAs) on which photomixers are typically made, \( \alpha \sim 10,000 \) cm\(^{-1} \), so that the vertical depth of heat generation lies within roughly 1 micron. So if the top side is air and the bottom side is semiconductor, we can model the photomixer as a physically small disk of heat lying on a thermally-conducting half-space. In this case we can write

\[
T_J = T_0 + \frac{P_Q}{k_\sqrt{2}} \left( \frac{r_{eq}}{k} \right)
\]  

(27)

where \( T_J \) is the junction temperature (i.e., at the air-semiconductor interface), \( T_0 \) is the ambient temperature, and \( k \) is the bulk thermal conductivity. In this expression, \( R_{TH} = \left[ \frac{(2)^{1/2} \kappa r_{eq}}{r_{eq}} \right]^{-1} \) is intermediate between the value for a circular heat source at the surface, \( R_{TH} = \left[ \frac{\kappa r_{eq}}{r_{eq}} \right]^{-1} \), and the value for a hemispherical source, \( R_{TH} = \left[ 2 \pi \kappa r_{eq} \right]^{-1} \). Clearly the photomixer is a 3D heat source but the vertical dimension, \( \sim 1/\alpha \), is generally much less than the lateral dimensions. And since photomixers generally have a square active geometry (for simplicity of fabrication), they can be analyzed through this expression with little error using an equivalent radius, \( r_{eq} = \left( \frac{A}{\pi} \right)^{1/2} \).

Like all semiconductor devices, photomixers have a maximum junction temperature \( T_{J,\text{max}} \) above which either catastrophic failure or rapid degradation occur. This allows us to define a maximum optical drive power by combining Eqns (26) and (27) to write

\[
P_{in,\text{max}} \approx \frac{(T_{J,\text{max}} - T_0) \kappa \sqrt{\pi A}}{(1 + \eta e E_B W_G / h \nu) \left( 1 + \eta e E_B W_G / h \nu \right)}
\]  

(28)

This expression has the interesting property that \( P_{in,\text{max}} \) is weakly dependent (square root) on the active area and practically independent of \( \eta \) in the range \( \eta < e E_B W_G / h \nu \). This is called the photothermal-limited regime. For \( \eta > e E_B W_G / h \nu \) \( P_{in,\text{max}} \) begins to fall and approaches a dependence of \( 1/\eta \) called the electrothermal-limited regime. This explains
Fig. 10. Maximum total optical drive power as a function of external $\eta$, showing the effect of heating caused by the dc photocurrent of high-$\eta$ devices.

why the majority of photomixers have been designed with a “back-off” in quantum efficiency well below unity.

These concepts can be illustrated on the canonical LTG-GaAs photomixer having $W_G$ fixed at 1.0 $\mu$m but a variable $\eta$ that is usually controlled in practice by varying the lifetime. According to experiments done on LTG-GaAs photomixers as a function of optical drive, the maximum $\Delta T_{J,\text{max}}$ has been found to be around 118°C, above which the devices fail rapidly.\(^{28}\) One theory for this failure is mechanical failure of the LTG-GaAs film caused by the in-plane stress induced by the CTE mismatch with the GaAs substrate and the large $\Delta T$. Fig. 10 shows the curves according to Eqn (28) for $\eta$ in the range $10^{-4}$ to 1.0 and $A$ parametrized between 5 and 400 sq. microns. For each curve the drop-off in $P_{\text{in,\text{max}}}$ begins to occur for $\eta > 0.01$ – a typical value displayed by the most useful THz photomixers. This is a surprise at first to people accustomed to using photoconductive detectors in optical receivers where $\eta$ is designed to be as high as possible and generally exceeds 0.5. The problem with photomixers having $\eta$ much higher than 0.01 is pre-mature burn-out from dc photocurrent-induced Joule heating.
E. THz circuit effects

The short-circuit current of Eqn (21) represents the conduction current that flows between the electrodes in the presence of photogeneration and a dc bias field. From Maxwell’s generalization of Ampere’s law we also know that a displacement current $\frac{dE}{dt}$ must flow between the electrodes. From a circuit standpoint, this displacement current can be represented by the capacitance $C_P$ calculated for the electrodes in Sec. II.C. This capacitance is almost always significant at THz frequencies and leads to the equivalent-circuit shown in Fig. (11). In this circuit $G_P$ is the differential resistance of the photomixer at the bias point, given approximately by $Y_{p\text{in}} = G_{p\text{in}}/V_b$. As in the lower RF bands, most THz load circuits can be represented as a complex admittance $Y_L = G_L + jB_L$, at least over limited bandwidths. But only the real part of $Y_L$ is capable of sinking the power corresponding to the difference-frequency generation term.

Given this circuit we can calculate the current phasor in the load by simple current divider action between $G_P$, $C_P$, and $Y_L$:

$$\tilde{I}_L \approx \frac{\tilde{I}(\Delta \omega)}{1 + j\Delta \omega \cdot R_L C} = \frac{2e}{h \nu} (P_{1,0} \cdot P_{2,0})^{1/2} e^{-j\phi} \times$$

$$\left[ \frac{\eta_{se}}{1 - j\Delta \omega \cdot \tau_{se}} + \frac{\eta_{sh}}{1 - j\Delta \omega \cdot \tau_{sh}} \right] \left[ \frac{G_L}{(G_L + G_p) + j(\Delta \omega \cdot C_D + B_L)} \right]$$

(29)

and the power in the load is
\[ P_L = \frac{|\tilde{i}_L|^2}{2G_L} \approx 2\left(\frac{e}{\hbar \nu}\right)^2 P_{1,0} \cdot P_{2,0} \times \]
\[ \left| \frac{\eta_e}{1 - j\Delta \omega \cdot \tau_e} \right|^2 \left( \frac{G_L}{(G_L + G_P)^2 + (\Delta \omega \cdot C_D + B_L)^2} \right) \]

This expression shows that the circuit introduces additional frequency dependence to the load power through the reactive part (i.e., \(B_L\)) of the load circuit. For the optimum case of \(P_{1,0} = P_{2,0} = P_{in}/2\), and for the special case of \(\tau_e = \tau_h \equiv \tau\) and \(\eta_e = \eta_h \equiv \eta/2\), we find

\[ P_L = \frac{1}{2} |\tilde{i}_L|^2 R_L \]
\[ = \frac{1}{2} S^2 P_{in} \cdot \frac{1}{1 + (\Delta \omega \cdot \tau)^2} \cdot \frac{G_L}{(G_L + G_P)^2 + (\Delta \omega \cdot C_D + B_L)^2} \]

This expression is very useful for estimating the output of THz photomixers. It can be analyzed in two general cases: (1) broadband loads for which \(B_L \ll 0\), and (2) resonant loads for which \(\Delta \omega C_D \approx B_L\) at the resonant frequency. It can also be used for performance optimization, as discussed further in Sec. V.A.

IV. Experimental Results/Comparison with Theory

A. DC characteristics and photocurrent

As in many semiconductor devices, the first and often best indicator of the quality of a THz photomixer is its dc current-voltage (I-V) curve. Shown in Fig. 12 are the I-V curves for typical LTG-GaAs and ErAs:GaAs devices taken at UCLA. Both devices consist of eight 0.25-μm electrodes separated by 1.0-μm gaps. The active area is 9x9 μm. Both devices display a very high differential resistance near zero bias, the LTG-GaAs exceeding 1 GΩ, and the ErAs:GaAs reaching just under 10 MΩ. At around 40 V in the LTG-GaAs sample there is a threshold where the current begins to increase rapidly and the differential resistance drops. This behavior has been modeled as a soft breakdown based on impact ionization of shallow traps. The resulting space charge can form a potential hill that blocks the terminal current – a mechanism commonly known as space-charge-limited current. The maximum bias field that can be sustained in the LTG-GaAs device is approximately 100 V, above which catastrophic failure occurs. This corresponds to a surface electric field of 10^6 V/m, which may be the highest breakdown field ever obtained in a GaAs material.
In contrast, the ErAs:GaAs sample displays a much higher absolute current, a more gradual increase with bias voltage, and no clear threshold effect. One model for this behavior is based on the fact that ErAs:GaAs is actually a superlattice consisting of ~20 nm lengths of normal GaAs separated by monolayer-thick layers of embedded ErAs nanoparticles. The purity of the normal GaAs should be very high, with high photocarrier mobility and no traps to ionize. And because there are no traps, the background concentration of free carriers is probably much less than in the LTG-GaAs. Under bias the only bottleneck in the transport is the ErAs nanoparticles, whose scattering cross section should decrease with increasing carrier kinetic energy $U$ with a power law that depends on whether the nanoparticles are charged or not (e.g., for ionized
impurities, $\sigma \propto U^2$). This explains qualitatively the gradual increase in current with bias voltage as a simple field-dependent mobility in the superlattice structure.

The other important characterization of photomixers is the dc photocurrent vs bias voltage. Both LTG-GaAs and ErAs:GaAs photomixers display a linear dependence of photocurrent on bias voltage up to a rather high bias. This is not predicted directly by the model of Sec. III, but is consistent conceptually if we consider the strong dependence of internal electric field on depth in the structure, as displayed by Fig. 5. This tends to reduce the average transit time and, therefore, increase the photoconductive gain with increasing bias. In a device having ~1 micron gaps, the largest increase tends to occur for photocarriers generated ~1 μm below the surface where the electric field is ~ 10-times lower than that at the surface. This makes the acceleration time to velocity saturation longer than at the surface, until the fields become so high that even at 1-μm depth the photocarrier acceleration time is a small fraction of the overall transit time.

B. Broadband THz output

The simplest THz loads to design and fabricate are the broadband self-complementary antennas, such as the log-spiral shown in Fig. 6. Submicron resolution is not required on any features except the interdigitated electrodes. The antenna pattern is guaranteed to be rather symmetrical about the optical axis of the photomixer, and dc biasing is rather trivial. One need only wrap the spiral with enough turns so that the lowest frequency of interest radiates away before reaching the outer extent of the spiral. Then, to dc bias the device one need only bond wire to the outer extent. No chokes or other RF passives components are required.
Fig. 14. THz output power from resonant (twin slot) antenna

Shown in Fig. 13 is the highest reported THz output power from a photomixer coupled to a broadband antenna - an LTG-GaAs device coupled to a two-turn log spiral antenna. The interdigitated structure has eight 0.2-micron-wide electrodes, seven 0.9-micron gaps. The LTG-GaAs had a measured, small-bias lifetime of 0.25 ps. The structure was fabricated at the driving point of a two-turn logarithmic spiral antenna. The measured power output is just above 1 μW up to about 1 THz and then falls rapidly at higher frequencies. Superimposed on this plot is the theoretical maximum broadband power predicted by Eqn. (31) for the following parameters: $C = 2.1 \text{ fF}$, $G_L = 0.014 \text{ S} >> G_P$, $B_L = 0$, $P_{in} = 78 \text{ mW}$, $v_e = v_h = 0.6 \times 10^7 \text{ cm/s}$, and $S = 6.1 \text{ mA/W}$. The agreement between experiment and theory in terms of frequency roll-off is remarkably good, both curves approaching 12 dB/octave at frequencies for which $\omega \tau > 1$ and $\omega C/G_L > 1$. The discrepancy in absolute power is about a factor of two – also very good considering the number of factors that can reduce the power coupled from the photomixer to the THz bolometer in the experiments.

C. Resonant THz output

Resonant load antennas are not as simple to design or implement as broadband antennas because of difficulties in simulation and design. By definition, resonant antennas have a significant conductance and susceptance that depend strongly on frequency and on the geometry of the antenna. And resonant antennas do not generally provide a symmetric beam or a simple means of dc biasing of the device.

The first reported resonant antennas coupled to photomixers were full-wave dipoles and slots$^{29}$ In contrast to log spirals, both displayed a resonant THz output power having a peak consistent with the resonant frequency for the given antenna type.
However, with a slot or dipole it is always questionable whether all the power is being collected as both types have distinctly asymmetric antenna patterns in the E and H planes. To improve the radiation pattern and provide even better cancellation of the electrode capacitance, the group at Lincoln Lab also fabricated a photomixer in a twin dipole antenna after successful demonstration of this approach with Schottky- and SIS diodes. The structure consisted of two half-wave (i.e., $0.5 \lambda_e$) dipoles separated by a little under $0.5 \lambda_e$ where $\lambda_e = \lambda_0/\eta_{\text{eff}}, \eta_{\text{eff}} = [(l+\varepsilon)/2]^{1/2}$, and $\varepsilon_e$ is the real part of the THz GaAs dielectric constant. The interdigitated structure was located at the mid-way point between the dipoles.$^{30}$

Shown in Fig. 14 is the output power from the twin slot antenna. The interdigitated structure had four 0.2-micron-wide electrodes and three 1.8-micron gaps. The LTG-GaAs had a measured, small-bias lifetime of 0.25 ps. The measured power output is now 3.0 $\mu$W at 850 GHz, 2.0 $\mu$W at 1050 GHz, 0.8 $\mu$W at 1600 GHz, and 0.2 $\mu$W at 2700 GHz. Superimposed on this plot is the theoretical maximum resonant power predicted by Eqn. (31) for the following parameters: $C = 0.58 \text{ fF}, G_L = 1/215 = 0.0046 \text{ S} \gg G_P, B_L = 0.02, P_{\text{in}} = 56 \text{ mW},$ and $S = 5.1 \text{ mA/W}$. The agreement between experiment and theory is even better than in the broadband case. The discrepancy in absolute power is less than 10% at 850 and 1050 GHz, but grows to 80% at 2700 THz.

D. Non-ideal effects

The agreement between experiment and theory in Secs. IV.B and C above is remarkably good considering the simplicity of assumptions used in the theoretical model. Four deleterious effects occur in photomixers that can significantly reduce the THz performance. The first effect is the possible anomalous thermal conductivity of the ultrafast photoductor and the dependence of thermal conductivity of the substrate on temperature. It is known that the thermal conductivity in LTG GaAs, for example, is significantly lower than in bulk crystalline GaAs.$^{31}$ This is assumed to be a result of the reduced phonon mean-free-path in the presence of the arsenic precipitates in this material. In addition to this problem, the thermal conductivity in the semi-insulating GaAs substrates is known to drop with temperature approximately as $T^{-1}$ because of increased phonon scattering (under ideal conditions this drop goes as $T^{5/4}$ — a dependence known as Bloch's law).

A second effect is the inequality of electron and hole drift velocities. Because of substantial differences in the band structure and effective mass of electrons and holes, the drift velocity is significantly different at most bias fields. The electron velocity is higher, particularly at the maximum point, where the electrons may “overshoot” the high-field saturation velocity by a large factor. So independent of how the averaging process is done in Sec. III, the average velocity of electrons will exceed that of holes by roughly a factor of two.

A third and more complicated effect is the inequality of electron and hole lifetimes. There is no reason that $\tau_e$ should equal $\tau_h$ except in the ideal case of a pure direct-band-gap semiconductor where the dominant recombination process is across the
gap. The ultrafast photoconductive materials used in photomixers are anything but pure semiconductors. For example, LTG GaAs is known to have many p-type trap states in addition to the mid-gap levels required for the sub-picosecond e-h recombination. The trapping time for a hole may be very short, perhaps less than a picosecond. But in the presence of the large bias electric field, the trap can be readily ionized by other free carriers, putting the hole back in the valence band for further drift to the electrodes.

A fourth and related effect is the dependence of photocarrier lifetime on the bias electric field. As discussed above in relation to the optical responsivity, at low bias the photocurrent is usually linear or quasi-linear, and the THz power varies quadratically. This occurs up to bias voltages high enough to quickly accelerate photocarriers deep within the device to the saturation velocity. However, the dielectric strength of LTG GaAs and, perhaps, ErAs:GaAs allows the application of such large fields that a second important effect occurs called field-dependent photocarrier lifetime. A detailed study has shown that at fields greater than roughly $10^5$ V/cm, ionization of traps begins to occur, so that photocarriers temporarily stopped from the drift process are released back into the closest band for transport. This effectively increases the lifetime, and related photoconductive gain, of the device. But by Eqns (24) or (31), it must also reduce the bandwidth. In other words, the $\eta B$ product remains the same, inversely proportional to the photocarrier transit time, not the lifetime. A similar effect occurs in avalanche photodetectors where one absorbed photon can generate more than one photocarrier, but the bandwidth must go down.

V. Improved Photomixers

A. Photomixer optimization: broadband load

The favorable agreement between experiment and the theory developed in Sec. III leads to the consideration of Eqn. (31) as the theoretical basis for optimization of interdigitated-electrode photomixers for maximum THz power. This can be done straightforwardly for two special cases: the broadband load, and the resonant load. For the broadband case we assume that $B_L = 0$ and $G_L$ is the appropriate value for a self-complementary antenna on a dielectric half-space, i.e., $G_L = (\epsilon_{eff})^{1/2}/60\pi$. For the resonant load, we assume that there exists a frequency for which the inductive portion of the load susceptance cancels the electrode capacitance (i.e., $B_L + \omega C = 0$) and that at this frequency the load conductance $G_L$ is known. In both cases the device impedance is accounted for through $G_P = I_0/V_b = SP_m/V_b$ and by writing the device capacitance according to Eqn (1). To include the thermal effects, we limit $P_{in}$ in Eqn (31) to the value $P_{in,max}$ of Eqn (28). Now Eqn. (31) becomes a parametric equation that can be evaluated numerically for a maximum. To minimize the number of free parameters, we make the practical assumption that the photomixer is square, so the active area can be written $[N_E(W_E) + (N_E-1)W_G]^2$. And we assume further that $W_E$ is limited by the fabrication technology to a fixed value of 0.2 $\mu$m So the four independent parameters become $N_E$, $W_G$, $\tau$, and $\Delta\omega$. 

178
Table I. Broadband photomixer optimization for $G_L = 0.014$ S.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>1.7</td>
<td>29</td>
<td>0.36</td>
<td>252</td>
<td>22.3</td>
<td>94</td>
<td>48</td>
<td>288</td>
<td>107</td>
<td>220</td>
</tr>
<tr>
<td>150</td>
<td>1.2</td>
<td>23</td>
<td>0.46</td>
<td>217</td>
<td>14.9</td>
<td>199</td>
<td>33</td>
<td>152</td>
<td>105</td>
<td>94</td>
</tr>
<tr>
<td>300</td>
<td>0.7</td>
<td>15</td>
<td>0.68</td>
<td>157</td>
<td>7.1</td>
<td>304</td>
<td>18</td>
<td>51</td>
<td>99</td>
<td>21</td>
</tr>
<tr>
<td>450</td>
<td>0.5</td>
<td>12</td>
<td>0.82</td>
<td>130</td>
<td>4.8</td>
<td>655</td>
<td>12</td>
<td>28</td>
<td>97</td>
<td>8.1</td>
</tr>
<tr>
<td>600</td>
<td>0.4</td>
<td>10</td>
<td>0.95</td>
<td>111</td>
<td>3.5</td>
<td>870</td>
<td>8.9</td>
<td>18</td>
<td>94</td>
<td>4.1</td>
</tr>
<tr>
<td>850</td>
<td>0.3</td>
<td>9</td>
<td>0.96</td>
<td>90</td>
<td>2.8</td>
<td>1288</td>
<td>6.7</td>
<td>13</td>
<td>91</td>
<td>1.7</td>
</tr>
<tr>
<td>1050</td>
<td>0.3</td>
<td>8</td>
<td>0.99</td>
<td>73</td>
<td>2.2</td>
<td>2122</td>
<td>6.6</td>
<td>11</td>
<td>81</td>
<td>1.0</td>
</tr>
<tr>
<td>1600</td>
<td>0.2</td>
<td>6</td>
<td>1.28</td>
<td>58</td>
<td>1.3</td>
<td>3.9</td>
<td>4.7</td>
<td>77</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>2700</td>
<td>0.2</td>
<td>5</td>
<td>1.20</td>
<td>34</td>
<td>0.8</td>
<td>4.0</td>
<td>4.0</td>
<td>59</td>
<td>0.08</td>
<td></td>
</tr>
</tbody>
</table>

Other variables, such as $h\nu$ and $\alpha$ clearly depend on the laser operating wavelength and on the photomixer material, but are not to be considered free parameters of the design since the availability of drive lasers, particularly semiconductor lasers, is limited to specific wavelength ranges.

The optimum photomixer parameters are listed in Table I and the corresponding maximum values of $P_L$ at several difference frequencies is plotted in Fig. 15. Note that the optimum $P_L$ falls by a factor over 2000 between 100 GHz and 2.4 THz. By inspection of Table I we deduce that this is caused by both the lifetime and capacitive-loading effects. Note that in all cases the optimum $G_P$ is much less than $G_L$, and the optimum RC 3-dB-down frequency $(2\pi C/G_L)^{-1}$ is surprisingly close to the design frequency $\Delta \omega$. In addition, the lifetime-related 3-dB-down frequency $(2\pi \tau)^{-1}$ tracks the design $\Delta \omega$ as well. As a result, the optimum parameters change rapidly with frequency. At the low-frequency end (100 GHz), the optimum $\tau$ is quite long, 1.9 ps, $N_E$ is large (21) and $W_G$ is very small (0.11 \mu m). At the high-frequency end (2400 GHz), the optimum $\tau$ is quite short, 0.1 ps, $N_E$ is small (4) and $W_G$ is larger (0.31 \mu m) to keep the capacitance tolerably low. Clearly, the trend at a given $\Delta \omega$ is to maximize the responsivity $S$ subject to the RC- and lifetime-related roll-off effects, and to the limitations on $P_m$ set by heating.

Of course a single broadband photomixer could never achieve the “optimum” curve of Fig. 16 because one of the key parameters ($\tau$) can not be changed after growth and two ($N_E$ and $W_G$) can not be changed after fabrication. So a more realistic approach is to fix the parameters for a chosen frequency and look at the plot of THz power vs $\Delta \omega$. In Fig. 16 this is done for the low-end (100 GHz) and high-end (2400 GHz) frequencies and compared to the optimum curve. Also shown are two reference roll-off curves having slopes of -6-dB/octave and -12-dB/octave, respectively. Note that if the parameters are chosen to achieve optimum performance at the low end, then the power at higher frequencies drops well below the optimum curve and approaches a roll-off slope of 12-dB/oct. This is consistent with the lifetime- and RC dependent terms of Eq. (31). On the other hand, if the parameters are chosen to achieve optimum performance at the high end, then the power at lower frequencies levels off to a value well below the
Fig. 15. Curves of the maximum attainable power from broadband photomixers as a function of \( \Delta \omega \) and four device parameters, whose value at each \( \Delta \omega \) is listed in Table I.

Table II: Resonant photomixer optimization for \( G_L = 0.004 \) S, \( B_L = 0.02 \) S

<table>
<thead>
<tr>
<th>Freq [GHz]</th>
<th>( \tau ) [ps]</th>
<th>( N_E )</th>
<th>( W_G ) [( \mu )m]</th>
<th>Area [( \mu )m(^2)]</th>
<th>( C ) [fF]</th>
<th>( (2\pi R_C C) ) (^{-1} ) [GHz]</th>
<th>( S ) [mA/W]</th>
<th>( G_P ) [( \mu )S]</th>
<th>( P_{in} ) [nW]</th>
<th>( P_L ) [( \mu )W]</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>1.5</td>
<td>33</td>
<td>0.48</td>
<td>482</td>
<td>31.8</td>
<td>94</td>
<td>45</td>
<td>249</td>
<td>134</td>
<td>2100</td>
</tr>
<tr>
<td>150</td>
<td>1.1</td>
<td>26</td>
<td>0.60</td>
<td>408</td>
<td>21.2</td>
<td>199</td>
<td>31</td>
<td>137</td>
<td>132</td>
<td>955</td>
</tr>
<tr>
<td>300</td>
<td>0.7</td>
<td>18</td>
<td>0.71</td>
<td>245</td>
<td>10.6</td>
<td>304</td>
<td>19</td>
<td>63</td>
<td>118</td>
<td>222</td>
</tr>
<tr>
<td>450</td>
<td>0.5</td>
<td>15</td>
<td>0.67</td>
<td>153</td>
<td>7.1</td>
<td>655</td>
<td>14</td>
<td>44</td>
<td>107</td>
<td>89</td>
</tr>
<tr>
<td>600</td>
<td>0.4</td>
<td>12</td>
<td>1.01</td>
<td>183</td>
<td>5.3</td>
<td>870</td>
<td>9.5</td>
<td>22</td>
<td>115</td>
<td>45</td>
</tr>
<tr>
<td>850</td>
<td>0.3</td>
<td>10</td>
<td>1.10</td>
<td>142</td>
<td>3.7</td>
<td>1288</td>
<td>6.9</td>
<td>14</td>
<td>109</td>
<td>20</td>
</tr>
<tr>
<td>1050</td>
<td>0.3</td>
<td>9</td>
<td>1.15</td>
<td>121</td>
<td>3.0</td>
<td>2122</td>
<td>6.7</td>
<td>12</td>
<td>100</td>
<td>12</td>
</tr>
<tr>
<td>1600</td>
<td>0.2</td>
<td>7</td>
<td>1.50</td>
<td>108</td>
<td>2.0</td>
<td>NA</td>
<td>3.9</td>
<td>5.3</td>
<td>102</td>
<td>3.9</td>
</tr>
<tr>
<td>2700</td>
<td>0.2</td>
<td>6</td>
<td>1.09</td>
<td>44</td>
<td>1.2</td>
<td>NA</td>
<td>4.6</td>
<td>5.7</td>
<td>67</td>
<td>1.0</td>
</tr>
</tbody>
</table>

The fastest roll-off for this case is approximately 6-dB/oct, consistent with a single frequency-dependent term.

**B. Photomixer optimization: resonant load**

Even more interesting is the optimization of photomixers coupled to resonant antennas. In this case we could start naively by setting \( B_L = 0 \) and then fix \( G_L \) to a value appropriate to a given antenna type. Application of Eqn (31) would then reveal a maximum, as should be expected, at \( G_P \approx G_L \), i.e., a large-signal impedance match between the photomixer and load. For practical values of \( G_L \), this procedure leads to unrealistically high values of \( P_L, N_E, C \), and \( P_{in} \). The photomixer area simply grows to
the value that can sustain the heat, and $P_{in}$ grows commensurately. So the optimization must be done with an additional constraint, which can be formulated in several ways. Perhaps the simplest constraint is on the imaginary part of $Y_L$, noting that practical resonant loads can only cancel the photomixer capacitance to a limited extent. A good example is the planar full-wave dipole antenna of Fig. 7. In the vicinity of its resonance, the full-wave dipole can be modeled as a parallel LCR resonator in which the $L_R$, $C_R$, and $R_R$ values can be determined uniquely from the center frequency, resonant resistance (or conductance), and full-width at half-maximum. On the GaAs half-space, a resonant frequency of 600 GHz is modeled by $R_R = 250 \ \Omega$, $L_R = 5.7 \ \mu\text{H}$, and $C_R = 1.1 \ \text{fF}$. So in the vicinity of the resonance, the inductive susceptance is $(\omega L)^{-1} = 0.046 \ \text{S}$.

Of course there are other possible resonant loads which could be designed to have higher reactance and thus lower inductive susceptance. A good example is a self-complementary antenna coupled to the photomixer through a transmission line containing a stub transformer. So for the purpose of optimization, we assume a value of $B_L = 0.02 \ \text{S}$. The resulting optimum photomixer parameters are listed in Table II. The corresponding maximum THz power at several difference frequencies is plotted in Fig. 16. Because of assumed cancellation of the electrode capacitance by inductive susceptance of the load, the resonant-load photomixer yields approximately ten times greater power than the broadband photomixer across the entire band from 0.1 to 3.0 THz. Interestingly, both curves have the same approximate slope of 6-dB/octave below 1 THz,
which gradually increases above 1 THz. At no point does the slope approach the fixed-
photomixer high-frequency rolloff of 12/dB octave.

C. Resonant-optical-cavity photomixers

The analysis in Sec. III suggests that the primary limitation on the interdigitated-
electrode photomixer performance is the low external quantum efficiency arising from
low photoconductive gain. A straightforward approach to mitigating this problem is to
enhance the fractional absorption near the surface where g is highest. In principle, this
can be accomplished by making the photomixer active layer thinner than \( \frac{1}{\alpha} \) and backing
it up with a dielectric mirror to form an optical cavity between the top semiconductor/air
interface and the dielectric mirror. Reducing electrode separation can then increase g
without causing more rapid RC rolloff if the photomixer is coupled to a resonant output,
as discussed in Sec. III.

To see the effect of an optical cavity, one can calculate the external quantum
efficiency as a function of LTG-GaAs thickness with the dielectric-mirror reflectivity \( \rho \)
as a parameter. The radiation in LTG-GaAs layer is assumed to propagate as a plane
wave so that matrix methods can be used to describe the intensity enhancement. The
resulting \( \eta \) is shown in Fig. 17 where it is observed to oscillate over the entire range of
thickness. The results depend critically on the absorption coefficient, which is assumed
to be \( 10,000 \text{ cm}^{-1} \) in these calculations. Each peak in Fig. 17 corresponds to the
constructive-interference condition \( D = n \frac{\lambda}{2} \) (\( n = 1,2,\ldots \)), and each valley corresponds to
the destructive-interference condition, \( D = n \frac{\lambda}{2} - \frac{\lambda}{4} \) (\( n = 1,2,\ldots \)). The peaks rise as the
bottom-mirror reflectance increases, but the valleys stay nearly fixed. The impact of
cavity resonance on \( \eta \) is shown in Fig. 17 where it appears that a substantial increase
over the single-pass \( \eta \) can be obtained with \( \rho \) as low as 0.3. For the easily realizable \( \rho \) of
0.9, the highest predicted value of \( \eta \) is 0.021 at \( D = 0.34 \mu\text{m} \). This is 2.7x greater than
the single pass value (0.0078) at 1.5 \( \mu\text{m} \), partly because of the greater fractional
absorption compared to the single-pass photomixer and partly because of the higher gain
of photocarriers absorbed close to the surface.

The impact of 2.7x improvement in \( \eta \) is significant. Because the output power
in Eqn (31) depends on \( \eta \) squared, these improvements may provide an increase in THz
output and O-E conversion efficiency of approximately 7.3 times. Of course, the
increased THz power does not come with impunity since the superior \( \eta \) also increases the
dc photocurrent and its contribution to the total thermal dissipation. The resulting
increase in junction temperature is a difficult issue that depends critically on the thermal
conductance of the dielectric mirror structure. On first analysis a standard dielectric
mirror, consisting of alternating layers of AlGaAs and AlAs, will have a much lower
thermal conductance than GaAs because of reduced phonon mean-free-path, so the
junction temperature may become prohibitively high. However, it is possible to use a
thick AlAs layer as the first material of the mirror below the photomixers. The thermal
conductivity of AlAs is higher than that of GaAs, and simulations have shown that if the
THz Generation by Photomixing in Ultrafast Photoconductors

Fig. 17. External quantum efficiency (left vertical scale) and photoconductive gain (right vertical scale) of a photomixer structure consisting of six 0.2-μm-wide electrodes and five 0.8-μm-wide gaps and driven at a λ of 0.85 μm. The cavity length represents a variable thickness of LTG-GaAs in the active layer, and the dielectric mirror below the LTG GaAs layer is parameterized by its power reflectivity ρ.

LTG GaAs layer is thin and the AlAs layer is relatively thick, then the latter will mitigate heating effects by acting like a heat spreader.35

D. Distributed Photomixers

In addition to the ROCP approach discussed above, another way to improve the bandwidth-efficiency product of photomixers is a distributed structure. The idea is to drive not a single interdigitated electrode structure but several in series, each integrated with a dielectric waveguide and a THz broadband transmission line, as shown in Fig. 18.

The dielectric waveguide must be made of a material, such as AlGaAs, that is transparent at the intended drive wavelength. The dielectric waveguide lies at the center of a coplanar-strip transmission line whose phase velocity is controlled by the periodic spacing and capacitance of the interdigitated-electrode structures. When the two frequency-offset laser tones propagate down the dielectric waveguide, a small fraction of the laser power is absorbed in each structure, creating a difference frequency current that excites the fundamental mode of the CPS transmission line. If the phase velocity of the dielectric waveguide and CPS are equal, then the THz photocurrents from each photomixer will add in phase, leading to enhancement of the total output power. This leads to a description of this device as the velocity-matched distributed photomixer (VMDP).36
The VMDP offers several interesting design issues relative to the vertically-driven photomixer structures. First, the interdigitated-electrode structures should be made very small in area so their capacitance C does not limit the THz conversion bandwidth. This condition can be stated as $\omega Z_0 C < 1$ where $Z_0$ is the characteristic impedance of the CPS line. Second, the electrode structures should be separated by a small enough distance $L$ that the THz bandwidth is not limited by the Bragg condition, $f_{\text{max}} = v/2L$, where $v$ is the velocity of the THz wave on the CPS line. But $L$ should be large enough to significantly reduce the thermal resistance compared to a single-element photomixer and thereby allow for higher laser drive power. In this way, one can improve the THz performance since the bandwidth of the VMDP is that of each interdigitated-electrode element, while the THz output power is increased through the cascading effect of the elements. Assuming $B$ is the design 3-dB bandwidth, it can be shown that the maximum drive power of the VMDP scales as $(B)^1$ instead of $(B)^2$.

### E. Photomixers for 1.55-μm

The primary reason for much of the research on LTG-GaAs is the fortuitous proximity of its band gap wavelength (~0.85 μm) to that of ultrafast (i.e., < 1 ps pulse width) mode locked lasers, such as Ti:Al₂O₃. Hence, LTG GaAs has been successfully used in a variety of pump-probe and electro-optic sampling techniques, which do not require spatial mixing of optical beams. Subsequently LTG GaAs was implemented in
Fig. 19. (a) Cross sectional view of material layers used in the first ErAs:InGaAs photomixer. (b) Room-temperature current-voltage characteristic of device.

the THz photomixer application, and promising results were obtained very quickly. However, photomixing is not as simple to implement as pump-probe experiments because of the complexity of efficiently combining two optical beams. With the lack of good fiber-optic components (particularly amplifiers), this is done in free space, which is sensitive to vibrations, requires a large volume for the optical components. Clearly, a better way to drive the photomixers would be with optical fiber in the heart of the telecommunications band from ~1.3 to 1.55 µm. In fact, LTG-GaAs displays significant absorption below the band edge. Unfortunately, the absorption coefficient around 1.5 µm is too weak to utilize in a sensitive photomixer.

A better approach is to fabricate photomixers on ultrafast In_{0.53}Ga_{0.47}As grown at low-temperature or with ErAs incorporation on InP substrates. The photon energy of 1.55-µm light (hv = 0.80 eV) is well above the cross-gap threshold so the absorption coefficient can be > 5000 cm\(^{-1}\). More importantly, 1.55-µm light provides ~ 80% more photons per watt as laser light just above the GaAs band edge, so that more THz output can be obtained per watt of laser drive. This is important for sub-picosecond-lifetime devices in which most of the thermal dissipation is from the absorbed laser power.

The problem with any In_{0.53}Ga_{0.47}As approach is that the narrow band-gap combined with a large background concentration of n-type impurities creates a high concentration of electrons, leading to high dark current and premature thermal breakdown of the photomixer structures. Recently, a growth technique has been developed that provides both short lifetime and high resistivity. It combines ErAs nanoparticle incorporation with Be compensation. The first successful photomixer made in this way...
Fig. 20. (a) Optical responsivity and (b) output power from ErAs:In_{0.53}Ga_{0.47}As photomixer.

contained 30 1.6-monolayer-thick ErAs layers placed 40 nm apart and embedded in an In_{0.53}Ga_{0.47}As matrix, as shown in cross section in Fig. 19(a). Eight 0.2-micron wide interdigitated electrodes with 1.0-micron gaps were deposited on the top surface to form the active area, which was coupled to a dipole antenna to radiate out the difference-frequency power into free space.

Electrical qualification of the photomixers was carried out by measurements of the I-V characteristics, resistivity, and breakdown field. The I-V curve shown in Fig. 19(b) is a direct indicator of the (dark) resistance of the device, which was previously too low to make useful photomixers. Through the advances at UCSB with ErAs incorporation and Be compensation, the resistivity has been increased to useful levels. From the quasi-static capacitance of our interdigitated electrodes, C ~ 2.5 fF and the measured zero-bias resistance R = 5.8×10^4 Ω of our I-V curve, we find RC = 145 ps. From electrostatics, we can set this equal to the dielectric relaxation time τ through RC = εr/ω = εeρ, leading to ρ ~ 118 Ω.cm. Note that the dielectric constant here is that of the ErAs:In_{0.53}Ga_{0.47}As, which we base on the dc permittivity εr = 13.94 of undoped In_{0.53}Ga_{0.47}As.24 This differs from the above calculation of antenna properties because the electromagnetic fields of the interdigitated electrodes are confined primarily to the ErAs:InGaAs epitaxial layer and the antenna fields extend mostly into the InP substrate below. Such a distinction is not necessary in GaAs photomixers.

The electric breakdown field E_B is estimated by taking the maximum voltage bias the photomixer and dividing by the gap width between interdigitated electrodes (0.9 μm). This results in E_B ~ 1×10^5 V/cm. Note that the choice of maximum bias is somewhat arbitrary since the breakdown in these devices is “soft”, similar to the impact-ionization breakdown of ErAs:GaAs. Nevertheless, this is comparable to the breakdown fields typically reported in In_{0.53}Ga_{0.47}As-based devices on InP substrates.

Optical qualification of the photomixers was made through measurements of the DC responsivity and optical-to-RF conversion. Results for the 1.55-μm dc responsivity
are shown in Fig. 20(a). At 30 volts bias, the responsivity peaked around 7.8 mA/W. Photomixing experiments were performed using two frequency-offset external cavity diode lasers (ECDLs) at 1.55 μm. The two ECDL beams were mixed in fiber via a 3-db fiber coupler and amplified by an EDFA. A fiber-coupled lens was used to focus the beam on to the device. Up to 26 GHz, difference frequency power was measured by directly connecting the device terminals to a RF spectrum analyzer via a coaxial cable. For frequencies greater than 100 GHz, a Golay cell was used to detect the difference frequency power coupled off-chip by a silicon lens through free space. Results of this measurement are shown in Fig. 20(b). The output power of 0.13 μW, is practically flat up to about 200 GHz and then drops rapidly. Based on these measurements, the upper limit on the photocarrier lifetime of the material is ≈ 1.0 ps. At the time of this composition, Be-compensated ErAs:In0.53Ga0.47As is being developed in which the separation between the ErAs nanoparticle layers is shorter so that lifetime below 1 ps can be obtained and THz photomixing can be realized.

F. Extension to Two-Dim Arrays

A promising means of substantially increasing the output power of photomixers is to combine the output of many elements configured as a two-dimensional array. It is known from RF phased-array antenna technology that a good architecture for the array is resonant antenna elements separated by approximately λ/2, where λ is the free-space wavelength. If the THz electric field at each element has the same amplitude and phase as that of its neighbors, then constructive interference will occur in space at a point equidistant from all elements, and destructive interference will tend to occur everywhere else. At a separation Z in the far field of the array (i.e., Z >> λ), the constructive-interference point is along the axis perpendicular to the plane of the array in the so-called “broad-side” direction. Under this condition the radiation pattern from the array, measured as a solid angle, will be narrower than the antenna pattern of a single element by approximately N². So the far-field broad-side intensity will be greater by N², even though the total radiated power (integrated over all angle) is greater by just N. This will, in general, make the THz output easier to collect and adapt to specific application. And if N is large enough that the array pattern becomes a “pencil” beam, the hyper-hemispherical coupling is no longer necessary. Even with a thick semiconductor substrate, such as that shown in Fig. 21(a), most of the radiation from the array will reach the semiconductor-air interface at angles less than the critical angle so can pass through with high transmission coefficient.

To understand the requirements on the array from the optical-drive side, we note from Eqn 8 and 9 in Sec. III that the THz amplitude depends on the product of the two laser powers, and the THz phase depends on the difference between the two laser phases. A simple way to make sure that the THz amplitudes at each element are the same is to feed the array using a 1:N optical power splitter. For photomixers amenable to fiber-optic drive, the splitter can be a “tree coupler”, as shown in Fig. 21(a). For VMDP
photomixers, the splitter can be made in dielectric waveguide using multi-mode interference (MMI) techniques. To make sure that the laser phase difference at each element is the same, a straightforward approach is to make the optical path between the power splitter and each photomixer element exactly equal. This is particularly simple in optical fiber since then one only needs to cut the individual feed fibers to an equal length. And note that the tolerance on this length is a small fraction of the THz wavelength, not the optical wavelength.

Important considerations for both array types in Fig. 21 are drive-power efficiency and bandwidth. In both the fiber-fed and waveguide-fed approaches, the power splitters are optically resonant elements. So a pair of single-frequency laser lines $\lambda_1 + \lambda_2$ can be split between the N output ports with insertion loss that is low for small N but tends to grow as N increases beyond roughly 8. And since photomixer elements require a significant optical drive (10s of mWs), it may not be feasible to split two laser drives N ways and have enough power, after inclusion of insertion loss, to drive photomixers. Fortunately, if the drive wavelengths are both around 1.55 micron, one can overcome both a large splitting factor and some insertion loss with an erbium-doped fiber amplifier located before the coupler, as shown in Fig. 21(a).\textsuperscript{41} So it is possible to use laser drivers that have high spectral purity (e.g., external cavity diode lasers) but relatively low output power. The resulting technique is similar to what is done in traditional microwave phased arrays where a pure RF tone (often synthesized) is split N ways and amplified, before or after the splitter, to driving the N elements of the array. The primary difference is that the optical feeding scheme must harbor two laser tones that may be frequency offset by 1 THz or more. Fortunately, the bandwidth of both fiber and
dielectric-waveguide components is ~10 THz or more, so that the two laser tones can be handled with near-ideal performance.

VI. Applications

A. Laboratory spectroscopy

An excellent example of the capabilities of the photomixer as a wideband sweep oscillator occur when it is applied as a wideband spectrometer for molecular spectroscopy. The first successful demonstration of such a spectrometer was conducted at NIST Gaithersburg. A spiral-antenna-coupled photomixer was driven by a fixed-frequency (standing-wave cavity) dye laser and a tunable (ring cavity) dye laser, each of which generated 50 to 100 mW of power near the peak gain of the rhodamine 6G dye at ~ 584 nm. Both lasers were jitter-stabilized to external confocal Fabry-Perot (CFP) etalons using the edge-lock servo control. This resulted in ring-laser jitter of ~ 1 MHz and standing-wave laser jitter of ~ 3 MHz. The absolute frequency of the standing-wave laser was established in comparison to a polarization-stabilized 633-nm HeNe laser, and the absolute frequency of the ring laser was monitored with a wavemeter and linearly varied to generate the THz difference frequency. To drive the photomixer the majority of power from the tunable laser was combined with approximately half of the output of the fixed laser and then focused on to the photomixer with a 25-mm-focal-length lens. The output of the photomixer hyperhemisphere was directed in to a 50-cm sample cell using Teflon lenses as windows, and then detected with a Si composite bolometer. The open-air path was kept < 5 cm to minimize water vapor absorption so prevalent in the THz region.

Fig. 22 shows a typical scan for 41 Pa of gaseous SO₂, an important molecule in atmospheric and environmental science, and a good control sample to test the accuracy of photomixer-based spectroscopy. This scan was made between ~855 and 900 GHz, and contained 6000 sample points, each covering 7.5 MHz of spectrum. Along the abscissa is shown the detector noise level obtained with the drive lasers blocked. Clearly the signal-to-noise ratio is very high. Because of this a large number of lines are measured in a scan time of a few minutes. Most of the lines belong to the Q₈ branch and their frequencies are in excellent agreement with the molecular constants known from theory and previous measurements. The rms deviation of the measured spectral lines from prediction is ~ 0.6 MHz, which was attributed primarily to the jitter and residual-scan drift of the drive lasers, which were free-running. Frequency or phase locking of the lasers would reduce the jitter considerably.

While the NIST demonstration established a convincing proof-of-concept, the dependence on dye lasers was clearly an impediment to the popularization of the photomixer approach. To alleviate this problem, a separate effort at Caltech demonstrated high-resolution spectroscopy using photomixers with semiconductor-diode (distributed Bragg reflector) drive lasers. For frequency stabilization, a small fraction of each laser
was directed to a folding mirror and then to a CFP cavity, so that the DBR lasers saw optical feedback only when they emitted in coincidence with a resonance of the cavity. Each CFP cavity consisted of two spherical mirrors bonded on quartz and piezoelectric tubes. The mechanical control of cavity length established the course setting of the resonant frequency. The electrical bias on the piezoelectric tubes established the fine setting over a range of ~1.5 GHz. Altogether this optical feedback technique provided a form of frequency locking free of electrical components and circuitry. To drive the photomixer, the majority of power from each DBR laser were combined and passed through an optical isolator to minimize the effect of optical feedback from the photomixer itself. The beams were focused on the photomixer chip, and the THz output beam was directed through a 20-cm gas sample cell and detected with a 4.2-K InSb hot-electron bolometer. In the optically-locked state, the time-integrated linewidth was found to be ~2 MHz compared to ~50 MHz in the free-running state. The 2 MHz was attributed to residual jitter and temperature-induced frequency drift in the DBR lasers, and represented the instrumental resolution. The all-solid-state spectrometer was demonstrated on a 150-mTorr sample of gaseous acetonitrile (CH$_3$CN) – a molecule of great interest to astrophysicists and chemists alike. Because of limitations on the optical self-locking range, the scan widths were limited to 115 MHz. Typical spectra are obtained by averaging 100 scans, each scan taking ~1 s. The averaging is made possible by the good long-term stability of the temperature-controlled DBR lasers. With the addition of 2$^{\text{nd}}$-derivative lock-in detection, the resulting signal-to-noise ratio is $> 10^4$. Because of this 4 lines were
fully resolved in just 100 seconds of measurement time. Most of the lines belong to the K stack of the J = 16 transitions of CH$_3$CN, and have linewidths of approximately 4 MHz.

B. **THz sensor and metrology applications**

Recent advances in superconducting THz receiver technology, particularly superconducting hot-electron-bolometer and SIS tunnel-junction mixers along with the utility of these receivers in astronomical and atmospheric science have created a compelling need for a compact, tunable local oscillator (LO) that can be integrated with airborne or spaceborne receiver platforms. The output power for this application is thought to be <10$^{-9}$ W from ~0.5 to 2.5 THz. The frequency stability must be comparable to that required for laboratory spectroscopy, ~1 MHz or better.

A key issue in the LO application is compactness. The photonic-drive side of the photomixer in the above implementations occupies too much volume and optical alignment to satisfy airborne or spaceborne applications. A first step in the compact direction is semiconductor-diode laser drives. A second step is constructing the photonic side in optical fiber. The photomixer can then be mounted on the same cryostat or refrigerator cold plate that holds the superconducting mixer. In the first such implementation, carried out at Lincoln Laboratory, the light from two laser diodes located external to the cryostat was combined in to a 4-m length of polarization-maintaining single-mode optical fiber. The fiber was coupled into the cryostat through a hermetic coupler, and the cleaved end was bonded over the photomixer active area using epoxy.

The demonstration of such a fiber-coupled-photomixer receiver has been made through a system integration effort at the Harvard Smithsonian Institute. Using an SIS receiver designed for frequencies above 500 GHz and a photomixer LO operating at 513 GHz, researchers obtained a double-sideband receiver noise temperature of 650 K. This result was limited by an insufficient LO power of ~10 nW usefully absorbed by the SIS mixer. Unfortunately, much of the photomixer output power was lost through diplexing and other coupling losses between the photomixer THz output and the SIS mixer. It was estimated that an increase in only 4 dB of coupled power would have reduced the receiver noise temperature to 200 K – close to the saturated-mixer value for this receiver.

A second technological application of the photomixer has been in metrology. There exist a family of atomic transitions (e.g., Cesium) for which the center frequency is very precisely known and can, therefore, be used as a frequency standard in the photonic regime. If one has a laser oscillating near one of the frequency standards, the laser frequency can be measured by beating it against the standard on a photodiode and measuring the difference frequency accurately in the RF region using a spectrum analyzer. This technique is easy only if the proximity of the unknown laser frequency to the standard is ~100 GHz or less where commercial spectrum analyzers exist. Unfortunately the standards are not very plentiful and there are many practical laser frequencies much more than 100 GHz away from the nearest standard.
One way to measure laser frequencies in this case is to extend the range of RF spectrum analysis using a wideband photomixer. The unknown laser tone and the standard are beat together to produce a difference frequency in the range of ~100 GHz to 1 THz. The photomixer output is then beat against a harmonic of an RF synthesizer on a Schottky diode mixer to produce a difference-frequency signal at 100 GHz or less for spectrum analysis. Because synthesizers are stabilized by digital phase-locked loops, they are even more precise than atomic standards, so the accuracy of this technique is outstanding.

VII. Conclusion and acknowledgements

This chapter has shown that the experimental THz performance achieved by broadband and resonant photomixers made from LTG-GaAs is already near theoretical limits imposed by electrical and thermal breakdown, recombination and RC time constants, and quantum-efficiency degradation imposed by the interdigitated electrodes. The only clear-cut means of improving the THz output power is through the improved structures and improved materials described in Sec. VI. Nevertheless, the interdigitated-electrode photomixer has already demonstrated several benefits as a new THz coherent source: (1) all-solid-state operation at room temperature, (2) simple fabrication by standard IC techniques on GaAs or InP substrates, (3) useful (> 1 μW) output for optical drive power (tens of mW) available from off-the-shelf GaAs semiconductor diode lasers, and (4) high tunability made possible through the wavelength variation of modern diode or solid-state lasers and the wide bandwidth of some planar antennas.

From a systems perspective, interest persists in cw THz generation for several scientific and technological applications, including terrestrial sensing, submillimeter-wave chemistry and astrophysics, and ultrafast analog and digital signal processing. To be useful in these applications, the radiation must be highly coherent (better than 1 part in $10^6$), minimally powerful (> 1 μW), and highly tunable (~ 10% or more, depending on the application). In addition it is desirable that the source display unimodal beam characteristics (TEM$_{00}$ Gaussian behavior preferred) for ease of coupling to other components. The photomixer meets all these requirements and does so under room-temperature operation - an issue sometimes overlooked by device researchers. Room temperature operation is very important to system integrators who traditionally are rather averse to cryogenically cooling their sources, even if receiver components (e.g., low-noise amplifier) have this requirement.

At this point in time, the primary disadvantage of photomixers is absolute THz power. The analysis carried out in Sec. III showed that the $\eta$-B product combined with the thermal-breakdown limit makes it difficult to increase the power of a single photomixer element beyond tens of microwatts at frequencies approaching 1 THz. This will preclude some applications, such as transmitters in active sensor systems. To increase the power into the mW range, two-dim arrays must be fabricated. This introduces packaging challenges such as the need to couple optical fibers. However,
with the addition of optical phase shifters, the two-dim array would offer the additional benefit of beam steering, which no existing single-element solid-state source can do.

Many people have contributed to the advancement of photomixer technology. The original research was carried out with the dedicated efforts of Alex McIntosh, Simon Verghese, Mike Manfra, Frank Smith, Karen Molvar, and Chris Dennis at Lincoln Laboratory. More recently, the author has relied heavily on the dedicated efforts of John Yeh and Mihail Sukhotin of UCLA; Andrew Jackson, Dan Driscoll, Micah Hanson, Christoph Kadow and Art Gossard of UCSB; and Rich Muller, Paul Maker, and Pierre Echternach of JPL. Valuable collaborations have been fostered by Drs. Rick Suenram and Allen Pine of NIST (Gaithersburg, MD), Prof. Ray Blundell of Harvard Smithsonian, Dr. Rob McGrath of JPL, and Prof. Geoff Blake of Caltech. Through the years, photomixer research has been supported by the Air Force Office of Scientific Research (Dr. Gerald Witt), NASA (Dr. Virendra Sarohia), DARPA (Dr. Edgar Martinez), and the U.S. Army Research Office (Dr. Dwight Woolard).

References


14 Experience indicates that this is approximately the maximum field that can sustained in planar LTG-GaAs devices.


21 see Infrared Laboratories, Inc., [www.irlabs.com](http://www.irlabs.com)


26 W. Shockley, “Currents to conductors induced by a moving point charge,” *J. Appl. Phys.* Vol. 9, p. 635 (1938)

27 Note that $\tau_e$ and $\tau_h$ can be much shorter than the cross-gap recombination time in the presence of deep levels or fast traps.


30 S.M. Duffy, S. Verghese, K.A. McIntosh, A. Jackson, A.C. Gossard, and S. Matsuura, “Accurate modeling of dual dipole and slot elements used with photomixers for coherent...
THz Generation by Photomixing in Ultrafast Photoconductors

45 Y.C.E. Tong and R. Blundell, Harvard-Smithsonian Center for Astrophysics.
46 S. Verghese, private communication.
This page is intentionally left blank
The prospects and advantages of silicon germanium quantum cascade lasers are discussed, from both physical and technological perspectives. A range of Si/SiGe intersubband laser configurations are discussed, for both edge and surface emission. Recent experimental activity on mid- and far-infrared devices is reviewed, and the value of detailed theoretical tools for heterostructure design is highlighted. Steps towards silicon optoelectronic integration are also considered.

Keywords: Silicon optoelectronics, intersubband, terahertz, virtual substrate.

1. Introduction

This chapter explores the techniques and prospects for Si-based SiGe/Si intersubband lasers that emit within the 3 to 10 THz spectral range, a region where photonic approaches to the creation of THz sources are appropriate. Today, SiGe intersubband lasers stand at the threshold of realization, and when these miniature solid state THz sources become actualized, they will open up a new Group IV-IV application area within quantum cascade laser (QCL) technology. Our laser-wavelength range of interest is 30 to 100 μm, also known as the far infrared (FIR) region. In this chapter we discuss innovative FIR device designs, simulation, theory, experiment, and the unique benefits of monolithic optoelectronic integration on a silicon substrate. The devices examined here are strained-layer superlattice and multi-quantum-well (MQW) heterostructures.

As discussed in the previous chapter of this book, QCLs in the InGaAs/InAlAs and GaAs/AlGaAs materials systems have been tremendously successful in the mid infrared (MIR) waveband and are presently moving towards the FIR or THz range. Generally, when the emission wavelength of a QCL is extended from MIR to FIR various difficulties are encountered, such as an increased free-carrier absorption loss which adds to the laser waveguide attenuation, thereby increasing the required gain and pumping. Also, the waveguide dimensions become larger and the modes become “leakier”; a problem which can be alleviated by substituting a metal-clad semiconductor structure for an all-dielectric THz waveguide. In addition, there is a perceived need to lower the operating temperature to approximately 20K, although, as we argue in this chapter,
operation at 77K or above appears possible for SiGe/Si lasers. We believe that the FIR technical hurdles can be surmounted in the IV-IV heterosystem.

2. Advantages and disadvantages of SiGe for intersubband lasers

Silicon is an excellent platform for monolithic integration of THz sources, THz detectors, and high-speed Si or SiGe electronic integrated circuits. Such optoelectronic (or "tera-electronic") integration could be used to create a miniaturized THz transceiver, for example. Silicon has a higher thermal conductivity than GaAs, thus providing better thermal management, which is a key issue in any active optoelectronic system. Silicon and the SiGe alloys also have the advantage of a relatively large radiative-to-nonradiative branching ratio for the intersubband laser transition. In the covalently bonded SiGe/Si heterosystem the non-polar optical phonon scattering is weaker than the polar optical phonon scattering found in III-V heterostructures. This implies stronger branching in IV-IV vis-a-vis III-V lasers. Between 4 and 77K, the alloy disorder scattering in SiGe dominates the optical phonon scattering, as discussed below. The indirect bandgap of SiGe, Si and Ge presents no disadvantage for quantum cascade laser operation because the radiation is generated by intersubband transitions, rather than interband recombination. The IV-IV heterostructures also offer a simple buried-mirror technique for vertical-cavity surface-emitting laser (VCSEL) construction. Unique to IV-IV is the fabrication of a WSi$_2$ or CoSi$_2$ layer buried beneath the surface of a Si substrate; a layer that is highly reflective at THz frequencies. The SiO$_2$ layer in a silicon-on-insulator (SOI) substrate may also be used as a buried THz reflector.

Unlike the conduction intersubband transitions employed in III-V QCLs, all current attempts to realize SiGe/Si quantum cascade lasers are based on valence intersubband transitions in p-type heterostructures. There are various reasons for this. For growth of pseudomorphic Si/SiGe heterostructures directly on silicon, substantial band offsets can only be achieved in the valence band. The standard method of obtaining usable conduction band offsets is to grow the heterostructure on a relaxed SiGe buffer (or 'virtual substrate'): even then the offsets obtained are generally not as large as those available in the valence band (up to 740meV for pure Si on pure Ge [1]). Realisation of a surface-normal emitting cascade laser (described in section 5 below) cannot be achieved in an n-type system (except by the use of a surface grating), but requires light hole to heavy hole transitions in a p-type heterostructure. That said, there is still potential for an n-type edge-emitting Si/SiGe cascade laser, as discussed in section 6 below.

One disadvantage of the p-type SiGe system is that the hole masses are significantly larger than the III-V electron masses. Consequently, the hole wavefunctions in SiGe heterostructures are quite well localized in individual quantum wells, hence interwell coupling is weak, and delocalised superlattice type states cannot be formed unless the barrier layers are very thin. For the same reason, the device currents in p-SiGe QCLs will be lower than in III-V devices for the same external bias. Also, both heavy hole and light
hole subbands are always present in p-SiGe heterostructures. Whilst this is an essential feature of surface-emitting QCLs, it is an added complication in the design of edge emitting lasers based on heavy hole to heavy hole transitions, and it makes the design of any interstitial injector regions more difficult.

Another difference between p-type Si/SiGe heterostructures and n-type III-V systems appears in plots of subband energy versus the wavevectors $k_x$ and $k_y$ (the in-plane dispersion). The valence subband dispersions are both anisotropic and non-parabolic, while the conduction subband dispersions in III-V materials are isotropic and almost parabolic (for the range of wavevectors typically encountered in cascade lasers). This feature of p-Si/SiGe systems clearly complicates device design; however, the non-parabolicity can be engineered to produce features which may be favourable for local-in-k-space population inversion, as discussed in section 5. The THz electric dipole matrix elements (the oscillator strength of the lasing transition) are comparable in the p-Si/SiGe and n-III-V cases.

3. **Advantages of the relaxed SiGe-on-Si virtual substrate**

The 4% lattice mismatch between Si and Ge can be viewed as a problem or as an opportunity for beneficial "strain engineering". Lattice matching to a silicon substrate is feasible for the special case of a SiGeC/SiGe superlattice wherein the Ge to C atomic ratio is chosen as 9 to 1 and the ternary SiGeC alloy layers are quantum wells. However, the carbon concentrations available with today's technology are < 2%, so the SiGeC strategy is limited to shallow quantum wells [2]. Generally, SiGe superlattice devices will require strained-layer epitaxy, unlike the unstrained III-V superlattices matched to GaAs or InP. Perhaps 98% of all the SiGe/Si photonic devices grown during the past decade have been grown directly upon an Si wafer and have been coherently but unsymmetrically strained; i.e., the SiGe quantum wells have had compressive in-plane strain with respect to the Si substrate, and the Si barriers have been unstrained. Because strain builds up across the asymmetrically strained MQW stack, the maximum permissible stack height is typically no more than 50 nm (the critical thickness for stable strain [3]), otherwise, dislocations will nucleate to relieve the strain in the stack. Therefore, the "unbalanced" strain approach is not a satisfactory solution to the problem of "thick" (0.1-10 μm) MQW growth. Strain symmetrization is the solution to this problem.

During the past three years, a strain-balanced MQW growth technique has been developed that allows stack thicknesses of several micrometers; beneficial for Si/SiGe lasers, detectors, modulators, resonant tunneling diodes, etc.[4]. First, a relaxed Si$_{1-y}$Ge$_y$ buffer or "virtual substrate" is grown upon silicon. The virtual substrate has a larger lattice parameter than Si and becomes a new "template" for subsequent SiGe/Si epitaxy. Next, the MQW is strain-engineered so that the compressive in-plane strain of the Si$_{1-y}$Ge$_y$ quantum wells opposes and balances the tensile in-plane strain of the Si$_{1-x}$Ge$_x$
barriers \((x > y > z)\). For sufficiently high precision growth, the net strain across the whole MQW stack is reduced to zero.

Experimental data for SiGe/Si virtual substrate heterostructures reported to date shows excellent interface flatness and smoothness. \cite{5}. The buffer concentration \(x\) offers another degree of freedom in device design that supplements the engineering of wavefunctions. SiGe buffer technology was pioneered at the Massachusetts Institute of Technology \cite{6}. In most approaches, the SiGe buffer is typically \(\sim 3\mu m\) thick, but new techniques are being developed (primarily for application in n-SiGe FETs) for growth of stable, strain-relaxed sub-\(0.1\mu m\) buffer layers \cite{7}. Two additional virtual substrate techniques are available today for device implementation. The first is the use of an SGOI substrate (SiGe on insulator) in which a thin crystalline layer of Si\(_{1-y}\)Ge\(_y\) rests atop a 500 nm layer of SiO\(_2\) upon silicon\cite{8-9}. The second is to base the devices upon a single-crystal Si\(_{1-y}\)Ge\(_y\) wafer (with \(0<x<0.2\)), although the present surface quality of such wafers needs improvement \cite{10}.

We believe that virtual substrate methods are likely to revolutionize the crystal-alloy layer quality and the performance of IV-IV MQW and superlattice devices. The virtual substrate appears essential for a silicon based cascade laser, since the number of periods usually required in such a device would require a total thickness of strained material well in excess of the critical thickness when grown directly on silicon. On the other hand, the success obtained so far with the virtual substrate approach implies that it may be possible to grow SiGe cascade heterostructures with a similar number of periods to the largest reported for III-V QCLs.

4. Theory and simulation techniques

Given the complexities of the Silicon-Germanium system, detailed modeling work is an important aid to understanding the electronic and optical processes involved in SiGe heterostructures, and to successful device design. The valence subbands in the SiGe system are strongly anisotropic, and significantly warped in in-plane \(k\) space due to mixing of heavy, light and spin-split-off hole states. Therefore, the one-band effective mass theory commonly used to model the electronic states of n-type cascade lasers is not sufficiently accurate here. However, for heterostructures involving reasonably low Germanium mole fractions \((< 0.5,\) which appears sufficient for THz photon emission), the \(k.p\) method has been shown to give results in very good agreement with pseudopotential bandstructure calculations \cite{11,12}, and hence provides a relatively fast means of calculating subband dispersions and carrier wavefunctions. The \(k.p\) method can be used for complex multiplayer structures, and a Fourier Transform based implementation provides substantial improvements in efficiency \cite{11}. The \(k.p\) approach can also be combined self-consistently with a solution of the Poisson equation to account for internal electric fields and space charge. Once the electronic states of a given cascade
structure have been obtained, the intersubband optical absorption spectra can then be calculated, for both edge and surface-normal propagating modes.

In both the III-V and IV-IV materials systems, the non-radiative intersubband transition rates are considerably faster than the radiative rates. Therefore, population inversion – and hence and laser gain – are determined primarily by the relative non-radiative transition rates in the system, and so accurate calculations of these rates are required for laser design and optimisation. In III-V heterostructures, polar optical phonon scattering is the dominant non-radiative intersubband transition process. In the non-polar SiGe system there is no polar carrier-phonon interaction, but there is a deformation potential optical phonon interaction which can give significant intersubband scattering. Several vibrational modes are present: Silicon-like phonons, Ge-like phonons, and new alloy phonon modes, all of which give intersubband scattering. For SiGe heterostructures designed for THz emission, the subband energy gaps are smaller than all the optical phonon energies (the 37.4meV Ge phonon being the smallest), and hence the intersubband phonon scattering rates are strongly suppressed at low temperatures ($\sim 10^9 \text{s}^{-1}$). At room temperature, the optical phonon scattering rates are typically $\sim 10^{11} \text{s}^{-1}$ – which is about an order of magnitude lower than the polar phonon scattering in III-V systems. The deformation potential interaction with acoustic phonons in SiGe is stronger than in n-type III-V systems, due to the larger effective masses involved, and the rates are generally comparable with those for optical phonon scattering [13]. Alloy disorder scattering in the SiGe quantum wells can also give rise to intersubband transitions. This process is virtually temperature independent, and hence is dominant in heterostructures with closely spaced subbands at low temperature [14]. Consequently, the intersubband lifetimes are relatively insensitive to temperature, in marked contrast to the case of polar III-V systems. This has been confirmed experimentally by pump-probe experiments on p-SiGe heterostructures using a free electron laser [15,16]. The measured intersubband lifetime for p-type modulation doped $\text{Si}_{0.72}\text{Ge}_{0.28}$ quantum wells grown on a $\text{Si}_{0.75}\text{Ge}_{0.25}$ virtual substrate was found to be $\sim 10^{-11} \text{s}$ throughout the temperature range 4.2-80K.

Another potentially important non-radiative scattering process is carrier-carrier scattering. In n-type III-V cascade lasers this process is important for transitions between closely spaced states, such as the lowest injector and highest active region states, and the lowest active region and highest collector states [17]. Relatively little theoretical work has been done on hole-hole scattering in quantum well structures. Our own calculations show that the LH1-HH1 hole-hole scattering rate in a $\text{Si}_{0.7}\text{Ge}_{0.3}/\text{Si}$ quantum well is about $10^{11} \text{s}$ at 4.2K for a carrier density of $10^{11} \text{cm}^{-2}$ (in otherwords, commensurate with the alloy disorder scattering rate), and is only weakly dependent on temperature [18].

More detailed studies of carrier dynamics require vertical transport models, using, for example, rate equation [17] or Monte Carlo methods [19]. Whilst the rate equation approach generally uses subband lifetimes which are averaged over a given carrier distribution, the Monte Carlo method can be programmed to account for the full k-space
dependence of scattering, and hence generate in-plane $k$-space carrier distributions which are essentially a solution of the semi-classical Boltzmann equation. This level of detail may be valuable, for example, in designs where local-in-$k$-space population inversion, rather than global subband population inversion, is sought, but obviously requires a much more extensive software development effort. We have developed a cellular Monte Carlo scheme in which scattering rates are tabulated on a relatively coarse 2-dimensional $k$-space mesh, and then interpolation is used to determine the exact transition rate between a specific pair of states[19]. This approach reduces the computer storage requirement to practical levels, and also ensures that the correct angular dependences for each scattering process are implicitly included in the model. The algorithm uses periodic boundary conditions to enable modeling of vertical transport through multi-period quantum cascade structures. Whichever of the rate equation or Monte Carlo methods is used, the task of modeling vertical carrier transport is simplified by the assumption that inter-well transitions are dominated by incoherent (intersubband scattering) processes, rather than coherent tunneling [20]. We have found that, for p-type heterostructures, the in-plane $k$ dependence of the inter-well scattering rates is an important factor in determining the device current[21]. The inter-well transition rates can vary strongly across $k$-space, and subbands in adjacent wells which appear resonant at the zone center are not necessarily resonant at finite $k_{||}$. Consequently, models based solely on the zone center transition rates, or on thermal equilibrium carrier distributions, can substantially underestimate the device current.

5. Surface emitting SiGe QCLs

All existing III-V n-type QCLs are edge-emitting devices, as a consequence of the symmetry rules which govern the radiative intersubband transitions. The optical matrix element is non-zero only for interaction with an electric field dipole oriented perpendicular to the quantum well layers (TM polarization), hence the emitted radiation propagates parallel to the layers – and thus emerges from the edges of the heterostructure. Collection of radiation from the surface of such a device can be achieved by fabricating a diffraction grating on the surface, although this is not a very efficient approach. Design of a QCL based on p-type Si/SiGe heterostructures allows the possibility of realizing a true surface-emitting device, without the aid of any grating. For transitions between pure heavy hole subbands, the symmetry rules restrict the interaction with radiation in the same manner as for electron intersubband transitions, resulting again in edge-emission. However, transitions between light hole and heavy hole subbands can also couple to in-plane electric field dipoles (TE polarization), resulting in surface-normal propagating radiation. For pairs of subbands in which the quantum well envelope functions have the same parity (e.g., LH1 and HH1), surface-normal emission only occurs for states with non-zero in-plane wavevectors (in the limit of zero electric field), but where the envelope functions have different parity (e.g., in the HH2 and LH1 subbands), strong surface emission can be obtained even at the Brillouin zone center.
The design of a QCL based on light hole – heavy hole transitions raises various challenges. For p-type Si/SiGe heterostructures the SiGe alloy layers represent the quantum wells, whilst the Si layers act as the confining barriers. In strain-balanced structures the energy gaps between different heavy hole subbands depend primarily on quantum well width, whereas the energy gaps between light hole and heavy hole subbands depend primarily on strain – and hence alloy composition – and are relatively invariant with well width. Therefore, engineering of heavy and light hole subband energies in a 3 or 4 level intersubband laser system requires simultaneous tuning of both layer widths and layer compositions.

For FIR laser design, the energy gaps between subbands are necessarily much smaller than in MIR devices (e.g., a 50μm (6THz) FIR device has a photon energy of 25meV, compared to 138meV for a 9μm (33THz) MIR device). Considering also that most MIR cascade lasers utilize optical phonon-mediated depopulation of the lower laser subband (with a phonon energy of ~36meV), the total energy difference between adjacent periods is at least 174meV for a 9μm device (there is also a small amount of energy loss in the injector region). The much lower energy drop per period in the FIR devices means that alternative design strategies must be considered. One attractively simple possibility is the ‘quantum-staircase’ design first proposed by Soref and co-workers [22,23]. Figure 1 shows a quantum staircase system based on LH1-HH1 intersubband transitions.

These transitions occur within each quantum well, and the carriers are transported from well to well by non-radiative HH1-LH1 transitions. The electric field is chosen to achieve near-resonance between the HH1 subband in one well, and the LH1 band in the following (‘downstream’) well. Clearly, non-radiative intra-well LH1-HH1 transitions will occur in addition to the desired radiative transitions and, in order to attain population inversion, the total inter-well HH1-LH1 transition rate must be faster than the total intra-well LH1-HH1 rate (which will be dominated, as in all intersubband lasers, by the non-radiative transition rate). Optimisation of the inter-well HH1-LH1 rate in such a system is rather difficult. As mentioned above, the interwell transition rates in QCLs are
generally dominated by the incoherent scattering processes, rather than coherent tunneling. This will certainly be the case in the p-type Si/SiGe system, where the effective masses in the Si barriers are much larger than those in the n-type III-V systems, leading to slow tunneling times. The large effective masses result in strong localization of the wavefunctions, which also reduces the matrix elements for all incoherent scattering processes. Obviously, reducing the barrier widths helps to increase the matrix elements, but this remedy is limited, in practice, by available growth technology. Furthermore, at the zone center, the heavy hole and light hole subbands are completely decoupled, therefore there is no transition matrix element between such states, even when they are exactly resonant. Thus, the interwell coupling is dependent on off-zone center states – which contain an admixture of heavy hole, light hole and spin-split off character (within the k.p representation).

One possible means of improving upon the basic LH1-HH1 quantum staircase design is by utilizing the inverted mass feature which can be engineered in the LH1 subband. An LH1-HH1 subband pair then form the basis of a 4 level unipolar laser, as shown in figure 2. Carriers are injected into the LH1 subband at the zone center, and then

![Figure 2. THz quantum staircase design based on HH2-LH1 optical transitions (wide arrows). The electric field is set such that carriers tunnel/scatter from HH1 into the HH2 subband in the following well (narrow arrow).](image)

relax to the (off-zone-center) subband minimum, from which radiative transitions to HH1 occur. These holes then relax towards the zone center in the HH1 subband, before being scattered into an adjacent well [22-24]. One advantage of such a system is that a 'global' population inversion between the LH1 and HH1 subbands is not necessary: it is sufficient to attain a local (in k-space) inversion, in the vicinity of the LH1 subband minimum, and this local inversion is can be more readily achieved due to the favourable band curvature of LH1 and HH1 respectively. Unfortunately, it is difficult to achieve a strong inverted mass feature in the LH1 subband in strain-balanced Si/SiGe structures:
the typical energy difference between the LH1 zone center and the subband minimum is of the order of 1-2meV. However, even for negligible energy differences, the fact that many of the LH1 carriers can access off-zone-center states is advantageous in itself, since the optical matrix element at the off-zone-center subband minimum is much larger than at the zone center. It should be mentioned that the inverted mass heterostructures are very difficult to realize in practice due to the sensitivity of the bandstructure to growth tolerances. Engineering of the inverted mass feature in the LH1 subband requires bringing LH1 into close proximity with HH2, at the zone center. This, in turn, requires accurate control of quantum well thickness, quantum well composition, and virtual substrate composition. Any error in any of these parameters can shift one or other of the subbands enough to destroy the inverted mass feature.

Another potential laser design based on the quantum staircase concept is the HH2-LH1 laser. A schematic diagram is shown in figure 3. In this device, there are three active subbands in each quantum well. The FIR optical transition is the intrawell HH2-LH1 transition. Depopulation of the lower laser subband (LH1 in this case) is achieved by non-radiative LH1-HH1 scattering. The electric field is chosen to attain near-resonance between HH1 and the HH2 subband in the following (downstream) well. This design offers two advantages over the LH1-HH1 quantum staircase. Firstly, the optical transition (HH2-LH1) is fully allowed at the zone center, whereas the LH1-HH1 zone-center optical transition is almost forbidden (the electric field serves to break the symmetry of the confining potential and hence allow weak coupling). Secondly, the inter-well coupling is much stronger, since it relies upon HH1-HH2 transitions rather
than HH1-LH1. True anticrossing behaviour can be observed between HH1 and HH2 in adjacent wells, for sufficiently thin barriers, with good delocalisation of wavefunctions across both quantum wells [18]. Although population inversion can be achieved between the HH2 and LH1 subbands, in the steady state a substantial percentage of the carriers reside in the HH1 subbands in each quantum well, which represents a loss in quantum efficiency. However, the situation is probably not much worse than that in typical MIR III-V cascades, where the total population in the many non-lasing states in the injector region, as well as that of the ground state in the active region, is quite substantial [19].

One way of improving the prospects of population inversion in SiGe quantum staircase devices is to design the structure to operate via inter-well, rather than intra-well, optical transitions. In such a device, the HH1 subband then becomes the upper laser level, whilst the lower laser level is an excited subband (LH1 in the simplest scheme) in the next ‘downstream’ quantum well, as shown in figure 4. The staircase must be biased such that

![Diagram of quantum staircase](image)

Figure 4. Quantum staircase designed for inter-well THz photon emission

the HH1 and the ‘downstream’ LH1 subbands are separated by the required THz photon energy at the zone center. Hence, the structure is readily field-tunable. In such a design, obtaining population inversion is relatively easy, since the majority of carriers generally occupy the HH1 subbands. A further advantage of this design is that the HH1-LH1 inter-well transition is fully allowed at the zone center, because the wavefunctions of the two states do not have the same parity relative to the two-well potential function. The main remaining design criterion is then to reduce the thickness of the Si barriers which separate the wells sufficiently to obtain a reasonably large optical matrix element for the inter-well transition. A barrier thickness of ~15Å is required to obtain optical matrix elements of a similar magnitude to those obtained for intrawell transitions [19].

A common problem for all the quantum staircase devices is the possibility of thermionic emission over the silicon barriers. The barrier height for low Ge composition wells with
Si barriers, grown on a silicon substrate, is relatively modest (~0.74x eV) where x is the Ge mole fraction. For strain-balanced heterostructures grown on a SiGe virtual substrate, the barrier height is somewhat lower for the same quantum well alloy composition, since the tensile strain (Si on relaxed Si$_{1-x}$Ge$_x$) results in a smaller shift in the bulk band edge than does the compressive strain (Si$_{1-x}$Ge$_x$ on Si or Si$_{1-y}$Ge$_y$ on Si$_{1-x}$Ge$_x$) [1] (the valence band offset for Si$_{0.7}$Ge$_{0.3}$ wells and Si barriers grown on a Si$_{0.8}$Ge$_{0.2}$ virtual substrate is approximately 120meV). The best way, in theory, to increase the valence band offset is to raise the quantum well Ge composition as high as possible, whilst keeping the virtual substrate composition as low as possible, but the strain balance criterion then dictates either very narrow wells – which will push the confined states closer to the top of the wells in any case – or very thick barriers – which will make inter-well coupling very weak. Of course, an alternative remedy for the thermionic emission problem is the design of a superlattice injector layer which acts as a DBR for the carrier wavefunctions, in a similar manner to that employed in n-type III-V cascades. However, injector design is more difficult for p-type cascades, since both light hole and heavy hole states are present; creating a `minigap’ in the energy spectrum therefore requires engineering of both layer thicknesses and compositions, within the limits of the strain-balance requirement.

Irrespective of the exact active region layer design, the generation of surface-normal propagating THz radiation from light hole – heavy hole transitions allows for the possibility of designing a vertical cavity surface emitting quantum cascade laser (VCSEQCL). As mentioned above, the Si/SiGe materials system offers an important advantage over III-V materials in this respect, in that the technology for buried reflector layers already exists. Si/SiGe multiple quantum well detectors which included silicon dioxide [25,26] or tungsten silicide [27] buried reflector layers, have both been demonstrated for operation at mid-infrared wavelengths. Cobalt disilicide is another possible choice of reflector material. New growth techniques are also emerging to enable preparation of a buried SiO$_2$ layer above a SiGe virtual substrate without the need for wafer bonding [28]. The envisaged VCSEQCL device geometry is shown in Figure 5.
Figure 5. Schematic diagram of the proposed SiGe vertical cavity surface emitting quantum cascade laser.

For the upper reflector, a distributed Bragg reflector (DBR) stack comprising alternate layers of amorphous silicon and SiO$_2$ is proposed. Of course, for operation at THz frequencies, the layer thicknesses required for the laser cavity and Bragg stack are unusually long. For 6THz operation, a $\lambda/2$ cavity requires a total Si/SiGe thickness of 7.6$\mu$m, which should be possible using chemical vapour deposition (CVD) for epitaxial growth, but not molecular beam epitaxy (MBE). Each silicon layer in the DBR must also be $\sim$3.8$\mu$m thick. Fortunately, there is a large refractive index between the silicon and SiO$_2$ layers, which means that a relatively small number of periods will be required in the DBR compared with those in visible-wavelength GaAs/AlGaAs VCSELs (at 6THz, only 6 periods are required for >99% reflectivity).

6. Edge emitting Si/SiGe QCLs

Edge-emission occurs for THz laser light polarized along the superlattice growth axis $Z$, the polarization that excites the TM$_0$ laser waveguide mode. The optical selection rules for SiGe/Si valence intersubband transitions say that $Z$ polarized radiation is strongly allowed for a transition between HH2 and HH1, hence this HH transition becomes the basis of edge-emitting QCLs. The University of Delaware and Sarnoff Corporation, under an Air Force contract funded by DARPA, are investigating edge-emitting SiGe/Si THz QCLs and have proposed a microdisk laser resonator that is discussed in another chapter of this book. This is a novel alternative to the ridge waveguide geometry which is used for MIR quantum cascade lasers and which should also be viable for FIR edge-emitting devices. The team is pursuing an FIR SiGe/Si QCL well-and-barrier design that
Silicon-Germanium Quantum-Cascade Lasers

contains MQW injector regions as well as 3-well or 4-well active regions, similar to the p-Si/SiGe MIR-emitting cascade structure reported by Dehlinger et al [29].

Simulations performed at AFRL and The University of Leeds have shown that there are two good possibilities for HH2-HH1 quantum staircase THz lasers [30]. The first technique is to engineer an inverted effective mass for the HH2 dispersion at \( k_x, k_y \) values of \( \sim 0.02 \text{ Å}^{-1} \), which allows the same type of local-in-k-space population inversion proposed for the LH1-HH1 laser. The second technique is to engineer the HH2 subband dispersion to be flat, or even parabolic near \( k_x = 0 \). When that is done, a zone center laser results. It has a global rather than local k-space population inversion. The zone center approach uses the anti-crossing of subband levels in adjacent quantum wells of the quantum staircase, wells coupled by a thin Si barrier. The level separation as a function of the applied electric field bias, known as the Stark ladder, reaches an anti-crossing condition at a field \( F = F_{ac} \). Slightly above \( F_{ac} \), HH1 and HH2 each form \( \sim 1 \text{ meV} \) doublets in most quantum wells of the staircase, yielding a 4-level laser scheme at the zone center with a photon energy smaller than the lowest optical phonon energy. Such a scheme was first discussed by Harrison and Soref for the CB2 and CB1 subbands of an GaAs/AlGaAs MQW [31] but the principle applies equally well to the HH2 and HH1 subbands in a p-type system.

The HH2 and LH1 dispersion shapes are governed by the proximity of the HH2 and LH1 subbands. We have observed the HH2-LH1 repulsion or anti-crossing in \( \mathbf{k.p} \) simulations. For example, we examined a series of eight related, unbiased, free-standing Si\(_{0.7}\)Ge\(_{0.3}\)/Si superlattices grown upon a (100) Si\(_{1-y}\)Ge\(_{y}\) buffer, where \( y \) is chosen in each case to give overall strain balance. Taking the barrier thickness \( l_b = 40\text{Å} \) with \( T = 77\text{K} \) and \( k_z = 0.5 \), we varied the quantum well thickness \( l_w \) from 80Å up to 150Å. As \( l_w \) was increased, we first found an inverted effective mass for LH1 at \( l_w \approx 100\text{Å} \). Then at \( l_w \approx 120\text{Å} \), the levels exchanged positions and an HH2 inverted mass appeared. This was followed by a nearly parabolic HH2 dispersion at \( l_w \approx 140\text{Å} \). So, the desired negative mass can be engineered generally by fixing \( x, y, l_b \), and varying \( l_w \).

Terahertz gain in a p-Si/SiGe quantum staircase design based on a negative-mass feature in the HH2 subband has recently been modeled in detail by Soref and Sun [32,33]. The subband dispersions for a structure comprising 90Å Si\(_{0.8}\)Ge\(_{0.2}\) quantum wells with 35Å Si barriers on a (100) Si\(_{0.9}\)Ge\(_{0.1}\) virtual substrate are shown in Figure 6. The desired radiative transition occurs between the two doublets in each well, as indicated in Figure 7, and has a calculated spontaneous emission lifetime of 77 μs at T=77K.
Figure 6. The (1,1) dispersion curves for a 5-layer SiGe/Si quantum staircase at the $F_0$ resonant tunneling bias.

Figure 7. Dispersion of the two doublet levels that appear in each central QW of an extended quantum staircase structure.

(The hole/ acoustic-phonon scattering lifetime at this temperature was 1.0 ns.) The dipole matrix element between the doublet states was 20Å. For an applied field of 30 kV/cm (which is 3.6 kV/cm above $F_{ac}$), and a hole current density $J = 1.7$ kA/cm$^2$ selectively injected into the upper HH doublet, a gain of 450 cm$^{-1}$ at a wavelength of 41 μm is predicted. The zone center laser with parabolic subbands has not yet been examined in detail.
There are also possibilities for n-type unipolar SiGe THz lasers, which await investigation. The conduction band offset in SiGe/Si superlattices (which determines the quantum well depth in the n-type laser) can be adjusted to some extent with the virtual substrate method. The quantum wells will be shallow for conduction subbands, yet deep enough for THz laser operation. Although the CB1, CB2...CBN subbands have their energy minima in the X-Δ region of the Brillouin zone, they have largely parallel $k_x$ dispersion curves, and strong THz radiative transitions are indeed expected between curves at the X valley; e.g., the X2 to X1 transition is allowed. The selection rules for the X subbands are affected by virtual substrate crystallographic orientations because band mixing comes into play [34]; thus a TE-polarized n-i-n VCSEL may be feasible for non-standard substrate orientations such as (110). One advantage of the n-type Si/SiGe system is that the Si layers act as the quantum wells for electron confinement, hence there is no alloy disorder scattering in the wells.

Low-loss waveguide design is an issue not-fully resolved in FIR edge-emitting lasers. The issue arises in the electrically pumped lasers discussed here, and in the phonon-pumped laser discussed in Section 7. Undoped semiconductor layers may be considered as dielectrics at THz frequencies, and an all-dielectric THz waveguide, either a ridge guide or a strip guide, will be feasible in a semiconductor heterostructure. However, the height of such waveguides is generally about $\lambda/2$, so for a typical wavelength of 60 μm (in air) this would imply semiconductor epitaxy $\sim$10 μm thick. As discussed above for the case of vertical cavity devices, epitaxial growth of this thickness is probably not practical using solid-source MBE, but should be achievable using CVD, and possibly gas-source MBE. An alternative solution is a metal-clad semiconductor waveguide (ridge or strip) in which the waveguide height is a small fraction of the wavelength such as $\lambda/20$. The results reported by Colombelli et al for III-V quantum cascade lasers operating at wavelengths of 21μm and 24μm [35] indicate that the metal-clad structure, known as a surface plasmon waveguide, should be practical at FIR wavelengths. The electromagnetic mode does not penetrate deeply into the metal film due to skin effect, ensuring low loss, and the TM0 mode can be supported for deep sub-wavelength waveguide thicknesses. Recently, THz laser action has been reported from an n-type GaAs/AlGaAs quantum cascade device [36,37]. A key feature of this device was apparently the low loss waveguide design. A ridge waveguide geometry was used, but incorporating adjacent n+ and undoped semiconductor layers such that the dielectric constant changes sign across the interface, providing confinement of surface plasmon modes without the need for a metal layer. This approach could also be used in the SiGe system.

7. Superlattice SiGe QCLs and quantum-parallel lasers

A 1997 paper from AFRL and more recent works [38-41] propose a SiGe/Si quantum-parallel laser (QPL); a device still awaiting implementation. The superlattice is in a nearly flat-band condition, and the laser transition is between superlattice minibands.
Recent papers from the Lucent group report innovative superlattice interminiband lasing in III-V heterostructures. The structure was a cascade of superlattice active regions in which interminiband lasing occurred "in parallel"[42]. Thus, this cascade was a series arrangement of parallel lasers. They investigated a cascade in which miniband transport was used within each superlattice injector in each period. They also studied an injectorless cascade in which the lower miniband of one active period resonated with the upper miniband of the next period[43].

We believe that analogous IV-IV cascades could be constructed using the valence minibands in a Si/SiGe superlattice. Rather than presenting those cascades designs explicity, we have chosen here propose a new and more speculative QPL: a flat band p-i-p SiGe/Si superlattice electrically pumped with holes. This QPL would have a very simple well-and-barrier structure. The ideas behind the QPL are: firstly, to use a low field $F$ that is below the Wannier-Stark localization field $F_{ws}$, implying that the shared miniband levels are not decoupled from one quantum well to the next, and secondly, to inject holes from a $p^+$ contact selectively into an upper miniband, giving rapid transport of carriers along that miniband to all quantum wells in the superlattice. Then the carriers in all quantum wells make simultaneous radiative transitions to the to a lower miniband, giving FIR emission. The quantum well pumping and emission are in parallel rather than in series.

The QPL lasing scheme - comprising, in effect, four levels - can be seen by looking at the valence subband curvatures as a function of superlattice normalized wavevector $Q_z$ over the superlattice minizone bounded by $k_z=0$ ($Q_z = 0$) and $k_z = \pi/P$ ($Q_z = 1$) where $P$ is the period (Figure 8). The upper and lower ends of a curve define the miniband width. We propose that a new THz QPL VCSEL could be created in SiGe/Si by utilizing a HH2 miniband to LH1 miniband radiative transition at the edge of the minizone. The dispersion diagram shown in figure 8, engineered at AFRL, was obtained after several $k.p$ trials; the final design being a strain balanced 70Å Si$_{0.7}$Ge$_{0.3}$/ 20Å Si superlattice on (100) SiGe virtual substrate. Selective pumping into HH2 is done at the superlattice zone center. Those holes relax rapidly via acoustic phonons to the HH2 minizone edge where the 22 meV TE-polarized vertical HH2-to-LH1 radiative emission occurs.
Note that Z-polarized photon emission from HH2 to HH1 is also feasible. For a QPL designed for TE emission, the cavity geometry would be engineered to suppress the TM mode (HH2-HH1) emission. However, an alternative QPL bandstructure design may also be conceived, in which HH2-HH1 is the desired radiative transition, with a ridge-waveguide geometry used as a resonator.

In the silicon rich superlattice of the QPL there are very few 37meV Ge-Ge mode optical phonons. Thus, in effect, the lowest-energy optical phonon modes are the Si-Ge alloy modes, with a phonon energy of approximately 52meV. We have engineered the superlattice so that the energy of the diagonal-in-\( Q_z \) optical phonon emission path from the populated HH2 band at \( Q_z = 1 \) to the relatively empty HH1 band at \( Q_z = 0 \) is 50.8 meV - less than the Si-Ge phonon energy. By that technique, the non-radiative phonon emission at \( Q_z = 1 \) that competes with the laser radiation is not energetically allowed, giving an enhanced ratio of radiative to nonradiative intersubband transitions, as desired. Hole/acoustic-phonon scattering is a very rapid intrasubband process that provides the necessary relaxation processes for this QPL, and acoustic phonons also help to empty the lower LH1 level, as needed for HH2-LH1 population inversion, by scattering holes from LH1 to HH1 at the zone center. More generally, the QPL design could also be implemented in III-V superlattices and in n-i-n devices that use the CB2 and CB1 minibands.

8. Other SiGe laser ideas: the phonon-pumped laser (PPL)

A revolutionary kind of SiGe/Si superlattice inter miniband THz laser - the "phonon-pumped laser" (PPL) - has been invented and analyzed by Sun, Soref, and Khurgin [44].
The PPL does not employ optical pumping or electrical pumping. Only a temperature difference across the superlattice, from 77K to 300K for example, is required to pump up the carriers from the lowest miniband into the excited miniband. The scheme has been described in recent publications [44,45]. A semiconductor layer called the heat buffer layer (a Ge-rich alloy) is deposited atop the superlattice to confine most of the optical phonons within the superlattice while acoustic phonons escape. Thus the optic phonons have a "hotter" temperature profile over a portion of the superlattice than the acoustic phonons as shown in Figure 9. This allows optical phonon absorption to create a hole population inversion between upper and lower minibands locally at the $Q_z=1$ zone edge. For example, the flat-band dispersion curves of a 68Å p-Si$_{0.94}$Ge$_{0.06}$/35Å Si superlattice strain balanced on an SiGe virtual substrate when plotted versus $Q_z$ look similar to those described above in Figure 9 for the QPL, and the opposed curvature of adjacent minibands is used in a 4-level phonon pumped scheme. Here the 52 meV Si-Ge phonon energy is greater than $E_{(HH2, Q_z=0)} - E_{(HH1, Q_z=0)}$, not less as in Figure 8. The quantum wells of the superlattice are p-doped, or remotely doped, to a density $N_a \sim 1 \times 10^{18}$ cm$^{-3}$ in order to populate the lowest miniband. It is also feasible to phonon-pump at the center of the Brillouin zone if the superlattice shows an inverted mass vs $k_x$ or $k_y$. In practice, the superlattice chip would be bonded to a cold finger of an optical dewar at 77K (or at 20K in closed-cycle cooler) that has a THz-transparent window situated near the chip end where the TM-polarized edge emission from the laser chip transmits through the window. To maintain the upper surface of the chip at about 300K, a thin-film NiCr electrical resistance heater (in the form of a stripe) would be deposited atop the superlattice as shown in Figure 10.

![Figure 9. End-view of PPL with its phonon-temperature distributions.](image-url)
Figure 10. Technique for maintaining $\Delta T$ across PPL chip.

This heater strip would have the same form as the line-shaped THz waveguide employed in the PPL. Then a dc current would be passed through the resistive strip and local Joule heating would give the desired temperature gradient across the superlattice stack. The required amount of electrical power dissipated in the strip depends upon the thermal conductivity of the superlattice, the superlattice thickness, and the width and length of the strip. For representative values, this power is about 0.2W.

9. Experimental progress to date

A consortium led by the University of Leeds (UK), and including Cambridge, Sheffield and Heriot-Watt universities, and the UK company QinetiQ, have been working towards the development of surface-emitting THz Si/SiGe quantum cascade lasers under DARPA/AFRL funding. P-type Si/SiGe heterostructures have been grown by low pressure CVD at QinetiQ using an industry standard Applied Materials Epi-Centura reactor. SiH4 and GeH4 sources were used, with a hydrogen carrier, whilst B2H6 was used to provide p-type dopants. The team have focused on strain-balanced heterostructure designs, as described in section 3, grown on SiGe virtual substrates, which typically comprise 3$\mu$m of linearly graded SiGe, followed by a 1$\mu$m strain-relaxed buffer at the target composition. Buffer compositions in the range 20-30% have been grown. Figure 11 shows a TEM image of a 10 period multiple

![TEM image of Si$_{0.72}$Ge$_{0.28}$ quantum wells with Si barriers and Si$_{0.78}$Ge$_{0.22}$ spacer layers grown on a Si$_{0.78}$Ge$_{0.22}$ virtual substrate [46].]
quantum well structure grown using this method. The buffer composition is 22% and the quantum well Ge composition is 28%. Each quantum well has Si barriers and the wells are separated by thin Si$_{0.78}$Ge$_{0.22}$ spacer layers. The TEM image clearly shows that, whilst misfit dislocations exist in the linear graded region of the sample, the vast majority terminate within the buffer layer, with no significant propagation into the quantum well layers.

The Leeds-led team has recorded a number of notable achievements in THz emission from Si/SiGe heterostructures. They have observed the first THz intersubband electroluminescence in Si/SiGe heterostructures (in both edge- and surface-propagation geometry)[46,47], and the first surface-normal electroluminescence in a quantum-cascade device in any materials system[5,48]. The quantum cascade devices grown by the team follow the quantum staircase design described in section 5. 30 period heterostructures comprising Si$_{0.7}$Ge$_{0.3}$ quantum wells and Si barriers have been grown, nominally strain balanced on a Si$_{0.77}$Ge$_{0.23}$ virtual substrate. Short (30nm) linearly graded (23-30% Ge composition) injector and collector layers were grown at either end of the quantum well stack to improve carrier injection and collection efficiencies. Figure 12 shows a TEM image of the 30 period quantum well stack, displaying the excellent uniformity and planarity of the layers, and again confirming the absence of any significant dislocation density in the heterostructure region.

Figure 12. TEM image of a 30 period Si/Si$_{0.7}$Ge$_{0.3}$ quantum cascade structure, nominally strain balanced on a Si$_{0.77}$Ge$_{0.23}$ virtual substrate [5].
Additionally, figure 13 shows the Germanium composition profile across the graded injector layer and the first few quantum well layers obtained using both electron energy loss spectroscopy (EELS) and energy dispersive X-ray analysis.

![Germanium Composition Profile](image)

**Figure 13.** Germanium composition profile across the graded injector layer and the first few quantum well layers of the quantum cascade structure shown in figure 12. The solid line shows data obtained from energy filtered TEM (electron energy loss spectroscopy), whilst the individual points (squares) mark data obtained from energy dispersive X-ray analysis [48].

Devices were fabricated by reactive ion etching into 180 μm x 180μm and 360 μm x 360μm mesas, with Al ohmic contacts, and were analysed by Fourier transform infrared spectroscopy using a Bruker 66V spectrometer with a liquid helium cooled Si bolometer detector. The spectrometer was used in step-scan mode with a lock-in amplifier, and the vertical bias applied to the cascade structures was pulsed (at 413Hz) to provide a lock-in reference. Figures 14 and 15 show typical spectra obtained for a sample mounted in edge-emission and surface-emission geometry respectively.
Both heavy hole – heavy hole and heavy hole – light hole optical intersubband transitions give rise to edge-propagating radiation, and figure 14 clearly shows features corresponding to the LH1-HH1 and HH2-HH1 transitions, as deduced from the agreement with the theoretically calculated intersubband absorption spectra, also shown. On the other hand, only the light hole – heavy hole transitions give rise to surface-normal emission, and, indeed, figure 15 shows that the HH2-HH1 emission is absent in surface-emission geometry, but the LH1-HH1 feature remains. THz power levels of up to 10nW have been measured in surface-emission mode at 4.2K, which is over 3 orders of magnitude higher than that reported for any other THz quantum cascade emission. When translated to a power efficiency, for the 15meV photon energy and 1mA current, this
device gives a figure of $2.4 \times 10^{-5}$, which compares remarkably favourably with the $4.4 \times 10^{-11}$ figure of merit obtained for spontaneous emission from a GaAs/AlGaAs THz cascade [49].

A consortium led by the Paul Scherrer Institute in Switzerland has reported electroluminescence from a prototype Si/SiGe quantum cascade structure at mid-infrared wavelengths [29]. The photon energy was $\sim 130$ meV, corresponding to HH2-HH1 transitions in their structure[50]. The MIR (edge-emission) power output was approximately 20 pW, for a current of 800 mA, giving a power efficiency of $\sim 10^{-11}$. The much higher efficiencies obtained for the SiGe THz cascade, compared to both the GaAs/AlGaAs THz cascade and the SiGe MIR cascade may be due, in part, to the reduced non-radiative scattering which results from use of a non-polar system (compared to the case of GaAs/AlGaAs) and from working at photon energies below the optical phonon energy (compared to the SiGe MIR device). Certainly, for a 130 meV subband separation, strong deformation potential optical phonon scattering may be expected, compared to the SiGe THz device. The high efficiencies obtained may also be due to a higher collection efficiency in surface emission mode, compared to the edge emission mode.

10. Prospects for Si-based optoelectronic THz chips

Optoelectronic integration is a key motivation for choosing SiGe/Si. We envision a wafer-scale or chip-scale Si or Si-virtual substrate platform serving as "motherboard" for a complete THz system including lasers, detectors, active switching components, Si and SiGe electronic drivers and amplifiers, and passive photonic components such as a planar network of THz waveguides, filters, couplers, splitters, combiners, add/drop wavelength multiplexers, and THz antennas. When THz radiation makes a transition from free space to-or-from the Si chip, Si gratings or undoped silicon lenses (transparent at THz) can aid this transition. (The assumption is that semi-insulating Si will be transparent because its free carrier absorption will be low.) SGOI and SOI are valuable THz substrates as well. We mentioned earlier the buried silicide and buried oxide possibilities for the lower mirror of a THz VCSEL. Another intriguing possibility being investigated at the University of Delaware under AFOSR sponsorship is THz waveguiding and switching in a photonic bandgap (PBG) lattice of posts or holes that are e-beam etched into an SOI substrate[51]. Simulations by FDTD indicate that a network of THz waveguide components and $2 \times 2$ electrically controlled THz switching components can be created in a 2D or 3D PBG SOI structure. Sharkawy et al [52] have modeled an SOI low-loss, low-crosstalk $2 \times 2$ PBG THz directional coupler switch in which the conductivity of the coupling region is controlled by carrier injection.
11. Summary

It is clear from the above description that much remains to be done in the area of SiGe quantum cascade lasers, and Si-based THz optoelectronics in general. However, there has certainly been an acceleration in activity. After several years existing as a theoretical proposal only, the SiGe quantum cascade laser is now being actively pursued in at least 4 experimental programmes worldwide. Although the SiGe materials system does pose some different challenges from those originally faced by the III-V quantum cascade laser community, none of these appear fundamental, and the example set by that community indicates that, with sufficient ingenuity, expertise and effort, a SiGe cascade laser can also become a reality.

Acknowledgments

The authors are most grateful to: D J Robbins and team at QinetiQ, Malvern, for growth of the Si/SiGe quantum cascade structures described in sections 5 and 9, D J Paul, S A Lynch and R Bates (Cambridge University) for processing and optical characterization of the same, D J Norris and A G Cullis (Sheffield University) for TEM and associated structural characterization, Z. Ikonic and P Harrison (Leeds University) for theoretical calculations and insight, and C R Pidgeon and E Martinez for discussions and support. This work was supported by the Air Force Office of Scientific Research and by Air Force contract F19628-99-C-0074 funded by DARPA/MTO.

References

18. Z Ikonic, unpublished
32. R. A. Soref and G. Sun, "Terahertz gain in a SiGe/Si quantum staircase utilizing the heavy-hole inverted effective mass" Appl. Phys. Lett.79 3639-3641 (2001)
37. A Tredicucci et al, unpublished
44. G. Sun and R. A. Soref, "Phonon-pumped terahertz gain in n-type GaAs/AlGaAs superlattices", Appl. Phys. Lett. 78 3520-3522 (2001)


50. MIR luminescence due to HH2-HH1 transitions in Si_{0.55}Ge_{0.45}/Si quantum cascade structures has also been reported by: I. Borman, K. Brunner, H. Riedhl, G. Abstreiter, S. Hackenbuchner, G. Zandler, P. Vogl, S. Schmult, W. Meindl and W. Wegscheider, "Mid-Infrared Intersubband Electroluminescence of Si/SiGe Quantum Cascade Structures", presented at the Sixth International Conference on Intersubband Transitions in Quantum Wells (ITQW'01), Asilomar CA, (10-14 Sept 2001).


This page is intentionally left blank
PLASMA WAVE ELECTRONICS

MICHAEL S. SHUR  
Electrical, Computer and Systems Engineering Department  
Rensselaer Polytechnic Institute, Troy, NY 12180-3590, USA  

VICTOR RYZHII  
Computer Solid State Physics Lab, University of Aizu, Aizu-Wakamatsu 965-8580, Japan

Plasma waves are oscillations of electron density in time and space. In deep submicron field effect transistors plasma wave frequencies lie in the terahertz range and can be tuned by applied gate bias. Since the plasma wave frequency is much larger than the inverse electron transit time in the device, it is easier to reach "ballistic" regimes for plasma waves than for electrons moving with drift velocities. In the ballistic regime, no collisions of electrons with impurities or lattice vibrations occur on a time scale on the order of the plasma oscillation period, and the device channel acts as a resonant cavity for the plasma waves, making possible tunable resonant detection or even emission of the electromagnetic radiation in the terahertz range. We review the theory of plasma waves in field effect transistors; discuss instabilities of these waves in different device structures and their applications for detection and generation of the terahertz radiation.

Keywords: Terahertz radiation, plasma waves, ballistic transport, resonant tunneling.

1. Introduction

The terahertz range of frequencies is often referred to as the "terahertz gap", since this frequency range lies in between the frequency ranges of electronic and photonic devices. The upper frequency limit of transistors operating in conventional regimes is limited by the transit time of carriers under the gate (for a field effect transistor) or across the base and collector depletion region (for a bipolar junction transistor). Scaling of feature sizes of silicon Metal Oxide Semiconductor Field Effect Transistors (MOSFETs) and compound semiconductor Heterostructure Field Effect Transistors (HFETs) and Heterojunction Bipolar Transistors has pushed the device parameters into the region, where the transistor operation at a few hundred gigahertz became feasible. However, device feature sizes approach the values, such that fundamental physics limitations lead to diminishing returns on investment in further scaling, and the transit time limited regimes face these limitations in trying to approach the terahertz range of frequencies. On the other hand, since the quanta of terahertz radiation are much smaller than the thermal energy at room or even at liquid nitrogen temperatures, photonic devices using interband or intersubband transitions have to operate at cryogenic temperatures (see Fig. 1).

Plasma waves are oscillations of electron density in time and space, and in deep submicron field effect transistors, typical plasma frequencies, $\omega_p$, lie in the terahertz range and do not involve any quantum transitions. Hence, using plasma wave excitation for detection and/or generation of terahertz oscillations is a very promising approach, and, as illustrated by Fig. 1, the terahertz gap in the electromagnetic spectrum between electronic and photonic devices can be closed by plasma wave electronics devices. We call this approach plasma wave electronics.

The properties of plasma waves depend on the electron density and on the dimensions and geometry of the electronic system. In a three dimensional case, the
plasma oscillation frequency is nearly independent of the wavelength, in a gated two-
dimensional electron gas (2DEG), the plasma wave have a linear dispersion law similar
to that of sound waves or light in vacuum. In this case, the plasma wave velocity, \( s \), is
proportional to the square root of the electron sheet density and can be easily tuned by the
gate bias that controls the 2DEG density.

In low electric fields, the effective electron mobility in short channel (submicron)
devices might be much smaller than the electron mobility in long channel devices. This
reduction was predicted for ballistic devices in \( ^{10,11,12} \), and, as was shown recently \(^{13} \), is
consistent with experimental data for submicron AlGaAs/GaAs High Electron Mobility
Transistors (HEMTs). This ballistic effect should be also important in short channel
devices implemented in silicon, silicon-germanium, AlN/GaN/InN or any other
semiconductors. The physical reasons for a drastic mobility reduction are related to the
ballistic motion first predicted in 1979. \(^{6} \) In ballistic field effect transistors, electrons
travel from the source to the drain without any collisions with impurities or phonons, and
electrons propagate in the device channel with a randomly oriented thermal velocity,

\[
v_{th} = \left( \frac{8k_B T}{\pi m} \right)^{1/2}
\]  

(1)

(where \( m \) is the effective mass, \( k_B \) is the Boltzmann constant, and \( T \) is temperature) or
with a Fermi velocity, \( v_F \), for a degenerate electron gas. Hence, electrons have only a
limited time to accelerate in the electric field and acquire a drift velocity. This
acceleration time is determined by $L/v_{th}$, where $L$ is the device length, (or by $L/v_F$ in a degenerate case). As a result, in low electric fields, the current is proportional to the electric field and to the electron concentration, just like in the collision-dominated case, but the electron mobility has to be substituted by an effective “ballistic” mobility, which (for a non-degenerate case) is given by

$$\mu_{\text{ballistic}} = \frac{2qL}{\pi m v_{th}}$$

In high electric fields, electron velocity in a ballistic device is expected to be higher or even much higher that in a collision dominated device, and the bulk plasma oscillations should lead to space oscillations of the electron density and resulting S-type current-voltage characteristic.

Plasma oscillations in a field effect transistor are affected by streamlines of the electric field directed from the channel toward the gate, and the velocity of plasma waves in a field effect transistor is typically much higher (by an order of magnitude or more) than the electron drift velocity. Hence, the characteristic transit time is much shorter, and the condition $\omega_p \tau_m > 1$ is much easier to meet that the condition $2\pi \tau_m / t_tr > 1$, where $t_tr$ is the electron transit time. When the condition $\omega_p \tau_m > 1$ is met, a channel of a field effect transistor can serve a resonant cavity for the plasma waves. As discussed below, the fundamental frequency of this cavity can be tuned by changing the gate bias, and a high mobility deep submicron FET can be used for resonant detection, mixing, multiplication, and even generation of terahertz radiation.

We will start from a general discussion of plasma waves for different geometries with a special emphasis on the plasma waves in a FET channel used for plasma wave electronics. We will then consider the analogy between plasma waves in gated two-dimensional electron gas (2DEG) and shallow water waves and the analogy between plasma waves in ungated 2DEG and deep water waves. This will be followed by the analysis of the instability of plasma waves in gated 2DEG channels. We will then review experimental results on resonant and non-resonant detection and new ideas on using resonant tunneling structures for sharpening surface plasmon resonances.

### 2. Plasma Wave Oscillations

Following, let us consider plasma wave oscillations in the systems of different dimensions by neglecting collisions and considering only the average drift velocity, $v$. The dispersion relations for plasma waves can be derived from the small signal equation of motion and the continuity equation, which, under these assumptions, are given by:

$$\frac{\partial \mathbf{j}}{\partial t} = \mathbf{E} \frac{e^2 n}{m}.$$  \hspace{1cm} (3)

$$\frac{\partial \rho}{\partial t} + \text{div} \mathbf{j} = 0.$$  \hspace{1cm} (4)

where $\mathbf{j} = qn\mathbf{v}$ is the current density, $e$ is the electronic charge, $n$ is the electron density, $m$ is the electronic effective mass, and $\rho$ is a small-signal charge density related to a deviation of $n$ from its equilibrium value, $\mathbf{E}$ is the small signal electric field. Eq. (3) follows from the Newton Second Law of Motion, where electron scattering is neglected. For a 3D geometry, the $\mathbf{j}$, $n$, and $\rho$ are current per unit area, electron concentration, and
electric charge per unit volume, respectively, whereas in the 2D case, the \( j, n, \) and \( \rho \) are current per unit length, electron concentration per unit area, and the electric charge per unit area, respectively.

Differentiating Eq. (4) with respect to time and using Eq. (3), we obtain:

\[
\frac{\partial^2 \rho}{\partial t^2} + \frac{q^2 n}{m} \text{div} E = 0. \tag{5}
\]

For the three dimensional case, the relation between \( E \) and \( \rho \) is given by:

\[
\text{div} E = \frac{\rho}{\varepsilon}. \tag{6}
\]

where \( \varepsilon \) is the dielectric permittivity. Substituting Eq. (6) into Eq. (5), we obtain the well-known expression for the 3D plasma frequency

\[
\omega_p = \sqrt{\frac{q^2 n}{\varepsilon m}}, \tag{7}
\]

where \( \varepsilon \) is the dielectric permittivity of a semiconductor.

For electrons under the gate of an FET, the electron sheet concentration in the channel is proportional to the voltage difference between the gate and channel potentials:

\[
qn = C U. \tag{8}
\]

Here \( U = U_g - U_c - U_T \), \( U_T \) is the threshold voltage, \( U_g - U_c \) is the potential difference between the gate and the channel, \( C = \varepsilon / d \) is the gate-to-channel capacitance per unit area, and \( d \) is the gate-to-channel separation. This equation is valid when \( U \) changes along the channel on the scale large compared to \( d \) (so-called gradual channel approximation).

From Eq. (8), we find the relationship between the electric field \( E = -\nabla U \) and two-dimensional charge density \( \rho = q(n - n_0) \)

\[
E = -\frac{1}{C} \nabla \rho. \tag{9}
\]

Substituting Eq. (7) into Eq. (3), we obtain the two-dimensional wave equation for the surface charge \( \rho \)

\[
\frac{\partial^2 \rho}{\partial t^2} - s^2 \Delta \rho = 0. \tag{10}
\]

where \( \Delta \) is the two-dimensional Laplace operator, and

\[
s = \sqrt{\frac{q^2 nd}{m\varepsilon}} = \sqrt{\frac{qU}{m}}. \tag{11}
\]

is the velocity of the surface plasma waves that have a linear dispersion law:

\[
\omega_p = sk. \tag{12}
\]

Here \( k \) is the wave vector of the plasma waves. Using Eq. (6) and (7), we can express the plasma wave velocity in terms of the gate voltage swing:

In a similar way, using Eq. (6) and the relation between the in-plane electric field and surface charge density, one can obtain the dispersion law for the plasma waves for ungated 2D electron gas
\( \omega_p = \sqrt{\frac{q^2n}{2m\varepsilon}} k \). \hspace{1cm} (13)

and for the plasma waves propagating along a one-dimensional wire:

\[ \omega_p = s_1 k \left( \ln \frac{1}{kr} \right)^{1/2}. \hspace{1cm} (14) \]

Here \( r \) is the radius of the wire, \( s_1 \) is the velocity of the one-dimensional plasma waves given by

\[ s_1 = \sqrt{\frac{nq^2}{4\pi m\varepsilon}}. \hspace{1cm} (15) \]

The case of a gated wire is very similar to that of a FET channel:

\[ \omega_p = s_{1d} k. \hspace{1cm} (16) \]

where

\[ s_{1d} = \sqrt{\frac{q^2n}{mC_{1d}}} = \sqrt{\frac{qU}{m}}, \hspace{1cm} (17) \]

\[ C_{1d} = 2\pi\varepsilon r + d_o. \hspace{1cm} (18) \]

The equations for the ungated 2D gas are valid when \( kd \ll 1 \) (in the opposite limiting case \( kd \gg 1 \) the existence of the metallic gate is irrelevant, and one obtains the dispersion relation given by Eq. (13)). For the 1D case, Eq. (16) holds when \( kr \ll 1 \).

Fig. 2 gives the summary of the plasma wave frequencies and dispersion relations for systems of different dimensions and geometry.

\[ 3D \]

\( \omega_p = \sqrt{\frac{q^2n}{\varepsilon\varepsilon_o m}} \)

\[ \text{ungated 2D} \]

\( \omega = \sqrt{\frac{q^m n}{2m\varepsilon\varepsilon_o k}} \)

\( s = \sqrt{\frac{q^m d_n}{m\varepsilon\varepsilon_o}} \); \( \omega = sk \)

\[ \text{gated 2D} \]

\[ s_1 = \sqrt{\frac{nq^2}{4\pi m\varepsilon\varepsilon_o}} \]

\[ \omega_p = s_1 k \left( \ln \frac{1}{kr} \right)^{1/2} \]

\[ 1D \]

Fig. 2. Plasma wave frequencies for different sample geometries.\(^3\)
Fig. 3 shows the dispersion relations for different geometries. (The dispersion curve calculations for Fig. 3 used the following parameters: dielectric constant, \( \varepsilon_r = 12.9 \), effective mass, \( m_r = 0.067 m_0 \); 3D electron concentration, \( n_{3d} = 5 \times 10^{24} \text{ m}^{-3} \); 2D electron concentration, \( n_{2d} = 1.5 \times 10^{12} \text{ m}^{-2} \); 1D electron concentration, \( n_{1d} = 1.41 \times 10^8 \text{ m}^{-1} \); gate-to-channel separation for the 1D gated structures, \( d_c = 3 \text{ nm} \); quantum wire radius, \( r = 3 \text{ nm} \).)

The similarity between the dispersion relations for the gated 2DEG case and gated and ungated 1DEG case implies that the results discussed below for the gated 2D EG should equally apply to gated and ungated quantum wires.

In ultra quantum limit in quantizing magnetic fields, it is possible to achieve a situation when electrons propagate as 1D particles but electrostatic interaction is three dimensional. This happens if the electron energy is less than the separation between the Landau subbands and optical phonon energy. In such a system, quasi-elastic collisions of electrons (with impurities and acoustic phonons) can only invert the direction of the electron propagation. This can minimize the effect of these collisions on plasma instabilities. As an example, Ivanov and Ryzhii considered the instability of the electron plasma optically generated in the lowest Landau subband.\(^\text{18}\) The dispersion equation for this system coincides with that for two colliding electron beams both in a classical case (for small wave numbers) and in a quantum case (when Wigner equation should be used). It is instructive that elastic scattering does not influence the instability threshold.

The same authors (jointly with Yu. Sigov and V. Kustov modeled the dynamics of this instability (which they called the collective relaxation) using an ensemble particle method and observed fast transformation of the distribution function due to interaction with the self-consistent electric field (unpublished). This instability led to a strong turbulence since its increment was close to the instability frequency.
3. Plasma Waves in FETs

Chaplik\textsuperscript{19} was the first to consider plasma waves in a gated 2D electron gas for electrons on the surface of liquid helium. Nakayama\textsuperscript{20} analyzed plasma waves for a FET. Allen et al.\textsuperscript{21} observed infrared absorption and Tsui et al.\textsuperscript{22} observed weak infrared emission related to such waves in silicon inversion layers. More recent studies by Burke et al.\textsuperscript{23} of high mobility AlGaAs/GaAs gated heterostructures revealed the resonance impedance peaks related to the plasma waves. Dyakonov and Shur used hydrodynamic equations to analyze plasma waves in 2DEG and predicted the instability of plasma waves in a high mobility field effect transistor.\textsuperscript{16} They also developed the theory of plasma wave electronics devices, such as detectors and mixers.\textsuperscript{15} Lu et al reported the evidence of the resonant plasma wave detection by a FET at the third harmonic.\textsuperscript{24} Recently, Knap et al.\textsuperscript{25} reported on the resonant plasma wave detection by a FET at the fundamental harmonic, and Peralta et al reported on a similar detection in multi gated periodic structures with 2DEG.\textsuperscript{26}

The gate electrode in a FET (see Fig. 4) is separated from the channel by the gate insulator (which is a doped or undoped wide band gap semiconductor, such as AlGaAs in a typical High Electron Mobility Transistors and SiO\textsubscript{2} in silicon MOSFETs).

![Fig. 4. Field Effect Transistor](image)

A FET operates in two different regimes. At gate biases, $V_G$, smaller than the FET threshold voltage, $V_T$, the source and drain contacts are separated by a potential barrier and the current between the source and the drain is very low. An increase in the gate voltage decreases the height of this barrier leading to an exponential rise in current. When $V_G > V_T$, electrons attracted by a more positive gate charge form a narrow sheet of mobile charge at the interface between the semiconductor and the gate insulator. This electron sheet and the gate contact form a capacitor, which capacitance per unit area is where $C = \varepsilon / d$, where $d$ is the gate-to-channel separation. Above the threshold, the surface concentration, $n_s$, in the FET channel is given by (see Eq. (8))

$$n_s = CU / q$$ (19)

As mentioned above, Eq. (19) represents the usual gradual channel approximation, which is valid when the characteristic scale of the potential variation in the channel is much greater than the gate-to-channel separation, $d$.

The equation of motion (the Euler equation) is
where \( \frac{\partial U}{\partial x} \) is the longitudinal electric field in the channel, \( v(x,t) \) is the local electron velocity, and \( \tau_m \) is the momentum relaxation time. The usual continuity equation can be written as:

\[
\frac{\partial n_s}{\partial t} + \frac{\partial(n_s v)}{\partial x} = 0 \tag{21}
\]

and re-written as

\[
\frac{\partial U}{\partial t} + \frac{\partial(Uv)}{\partial x} = 0 \tag{22}
\]

taking into account Eq. (19).

In the limiting case of a ballistic FET, \( \tau_m \) tends to infinity, and Eqs. (20) and (21) coincide with the hydrodynamic equations for shallow water (see, for example,\(^{27}\)). This means that the 2D-electron fluid in a Ballistic FET should behave like shallow water. In this hydrodynamic analogy, \( v \) corresponds to the fluid velocity, and \( qU/m \) corresponds to \( gh \), where \( h \) is the shallow water level and \( g \) is the free fall acceleration.

Hence, phenomena similar to wave and soliton propagation, hydraulic jump, and the "choking" effect\(^ {28} \) should take place in a ballistic hydrodynamic electron fluid. Effects of collisions, surface scattering, changes in the channel cross section, and others may be also understood using this analogy.

In the opposite limited case of a collision dominated transport, Eq. (20) becomes

\[
v = -\frac{q}{m} \tau_m \frac{\partial U}{\partial x} = \mu E, \tag{20a}
\]

which is usually replaced by a field-dependent velocity function

\[
v = v(E), \tag{23}
\]

which accounts for the electron velocity saturation in high electric fields and/or for a region of negative differential mobility. Examples of empirical velocity-field dependences are

\[
v = \frac{\mu E}{\sqrt{1 + \left(\frac{\mu E}{v_s} \right)^2}} \tag{24}
\]

for electrons in silicon, and

\[
v(E) = v_s \left[ 1 + \frac{E/E_s - 1}{1 + A(E/E_s)^t} \right] \tag{25}
\]

for electrons in compound semiconductor materials with a negative differential mobility region, such as GaAs. Here \( E_s = v_s/\mu \) is the velocity saturation field, \( v_s \) is the electron saturation velocity, and \( A \) and \( t \) are fitting parameters. The analytical model of a FET, which accounts for the saturation of the electron velocity and for the source and drain series resistances yields the following expressions for the device drain current.
Plasma Wave Electronics

\[ I_d \approx \frac{g_{ch} U_{ds}}{1 + (g_{ch} V_{ds} / I_{sat})^m} \]  \ \text{(26)}

where

\[ g_{ch} = \frac{g_{cho}}{1 + g_{cho} (R_s + R_d)} \]  \ \text{(27)}

\[ g_{cho} = C_i \mu_n \frac{W}{L} (U_{gs} - U_T) \]  \ \text{(28)}

\[ I_{sat} = \frac{g_{cho} (U_{gs} - U_T)}{1 + g_{cho} R_s + \sqrt{1 + 2 g_{cho} R_s + \left( \frac{U_{gs} - U_T}{U_L} \right)^2}} \]  \ \text{(29)}

\[ U_L = E_s L, \] and \[ R_s \] and \[ R_d \] are source and drain series resistances, respectively.

In the collision dominated regime, operating regime, the upper frequency of operation of a Field Effect Transistor (FET) is limited by the electron transit time, \( t_{tr} \). The transistor cutoff frequency, \( f_T = 1/(2\pi t_{tr}) \), see for example, 29. However, plasma effects become important in modern, short channel field effect transistors, where the sheet carrier density is very high. These effects should allow us to use FETs at much higher frequencies (in the terahertz range for deep submicron devices).

As explained above, plasma waves with a linear dispersion law, \( \omega = sk \). The velocity of the plasma waves, \( s \), is typically on the order of \( 10^8 \) cm/s, which is much larger than the drift velocity of the 2D electrons in the FET channel. This is why the propagation of plasma waves can be used for new regimes of FET operation, with a much higher frequency than for conventional, transit-time limited regimes. Under certain conditions, plasma oscillations can be excited in a FET by a dc current, and the FET can be used as an oscillator operating in the terahertz range. 16 Nonlinear properties of the plasma waves can be utilized for terahertz detectors, broad band detectors, mixers, and frequency multipliers. 30

Devices operating at terahertz frequencies should be considered to be inseparable from the circuit. Therefore, issues related to coupling plasma wave to electromagnetic radiation conversion, to antenna structures for electromagnetic radiation, and to device integration with submillimeter circuits will have to be addressed for practical implementation of plasma wave electronic devices. Since the plasma wave velocity is much smaller than the light velocity, and the device dimensions are much smaller than the electromagnetic wave length corresponding to the plasma frequency, antenna structures -- which are much larger than typical devices -- are needed for coupling plasma and electromagnetic waves. These issues and the antenna and circuit design will be similar to those currently investigated for deep submicron Schottky diodes operating in the terahertz range. 31
4. Plasma wave instability

Dyakonov and Shur showed that the surface plasma waves propagating in gated 2DEG might grow in a ballistic device due to reflections from the device boundaries. The linearized system of Eqs. (20) (in the limit of \( \tau \to \infty \)) and (22) predicts the dispersion law, \( k = \pm \omega / s \), which is the same as for the shallow water waves. The wave velocity \( s = (qU_0/m)^{1/2} \). If the electrons move with a drift velocity \( v_o \), the waves are carried along by the flow, and the dispersion relation becomes \( k = \omega \sqrt{(v_o \pm s)} \). (The drift velocity, \( v_o \), corresponds to the current per unit width \( j = qn_s v_o = CU_0 v_o / q \).)

Dyakonov and Shur assumed the following boundary conditions at the source and drain side the device channel: constant gate-to-source voltage, \( U_{gs} \), at the source side of the channel \( (x = 0) \) and constant drain current at the drain side of the channel \( (x = L) \). The realization of these boundary conditions is straightforward at low frequencies, i.e. one applies voltage between the gate and source using a voltage source, \( U_{gs} \), and connects a current source, \( I_{ds} \), to the drain. At high frequencies, the boundary conditions, such as short and open circuits, cannot be easily implemented. Also, the displacement current between the gate and the channel plays an important role. However, all microwave, millimeter, and submillimeter field effect transistors designed in such a way that the gate-to-source parasitic and fringing capacitances (which add to the input capacitance) is greater or even much greater than the gate-to-drain parasitic and fringing capacitances (which contribute to the Miller capacitance). The intrinsic gate-to-source and gate-to-drain capacitances are equal at zero drain bias. However, the intrinsic gate-to-source capacitance increases with an increase in the drain bias and the intrinsic gate-to-drain capacitance decreases with an increase in the drain bias (see Fig. 5).

Therefore, the boundary conditions at the source are closer to the short circuit conditions at high frequencies and the boundary conditions at the drain are closer to the open circuit. This asymmetry increases with an increase in the drain bias as shown in Fig. 5. We let \( v = v_o + v_1 \exp(-i\omega t), U = U_o + U_1 \exp(-i\omega t) \), linearize Eqs. (20), (22) with respect to \( v_1 \) and \( U_1 \), and use the boundary conditions \( U_1(0) = 0 \) and \( \Delta j(L) = 0 \) (i. e., \( U_0 v_1(L) + v_o U_1(L) = 0 \)) as discussed above. We then seek the solution as the sum of two waves propagating from the source to the drain and from the drain to the source, respectfully, with the wave vectors \( k_1 \) and \( k_2 \):

\[
\begin{align*}
  v_1 &= A \exp(ik_1 x) + B \exp(ik_2 x) \\
  U_1 &= C \exp(ik_1 x) + D \exp(ik_2 x)
\end{align*}
\]

where

\* The analogy between plasma waves in a FET and water waves goes even further. While for \( kd \ll 1 \) the plasma waves with \( \omega \sim k \) are similar to the shallow water waves, in the opposite limiting case \( kd \gg 1 \), the plasma waves have the same dispersion law \( \omega \sim k^{1/2} \) as the gravitational waves in deep water.

\† We should notice that capacitances shown in Fig. 5 are calculated assuming the collision dominated transport in a FET. However, the qualitative conclusions should apply to the ballistic regime as well.

234
This procedure leads to the following expressions for the real and imaginary parts of \( \omega = \omega' + i \omega'' \):

\[
\omega' = \left( \frac{\omega^2 - \nu_0^2}{2LS} \right) \ln \frac{s + \nu_0}{s - \nu_0}
\]

\[
\omega'' = \left( \frac{\omega^2 - \nu_0^2}{2LS} \right) \frac{-1}{\ln \frac{s + \nu_0}{s - \nu_0}}
\]

where \( n \) is an odd integer for \( |\nu_0| < s \) and an even integer for \( |\nu_0| > s \). Eq. (35) shows that for positive \( \nu_0 \), the steady flow is unstable if \( \nu_0 < s \) and stable if \( \nu_0 > s \). For \( \nu_0 / s < 1 \), \( \omega'' = \nu_0 / L \) which is the inverse electron transit time.

The boundary conditions play a very important role in determining the growth of this plasma instability. Different boundary conditions might be interpreted in terms of effective impedances connected between the source and gate and the gate and drain, respectively (see Fig. 6):

\[
Z_{gs} = \frac{U_1(0)}{j_1(0)W}
\]

\[
Z_{gd} = \frac{U_1(L)}{j_1(L)W}
\]

where \( j_1 = qn_{s1} \nu_0 + qn_{s0} \nu_1 \) and \( W \) is the gate width.
Our analysis showed that the instability might take place as long as $|Z_{gs}| \ll |Z_{gd}|$. A more detailed analysis of different boundary conditions for this system was given by Crowne.\textsuperscript{33}

The above analysis is based on the solution of hydrodynamic equations, which are valid when electron-electron collisions are very frequent. Dmitriev et al.\textsuperscript{34} presented the solution of the Boltzmann equation for a relatively low density gated two-dimensional electron gas, where electron-electron collisions are not significant. This solution revealed that the plasma waves have the same dispersion law as for a high electron sheet density. In both cases, the plasma waves become unstable in a short channel field effect transistor with asymmetric boundary conditions at small channel currents. They also investigated the role of the Landau damping. The Landau damping is the damping of a plasma wave caused by a transfer of energy from the wave to particles, whose velocity equals the phase velocity of the wave. In this system, the electron velocity is $v_F$ (or $v_{th}$ in a non-degenerate case) and the plasma wave velocity is $s$. Dmitriev et al showed that the Landau damping is small when $s \gg v_F$, which is quite understandable. The inequality $s \gg v_F$ can be presented as

$$d \gg r_s/2$$

(38)

where $r_s = \frac{4\pi e^2 \hbar^2}{m q^2}$ is the Bohr radius. For GaAs, $e = 1.14 \times 10^{-10}$ F/m, $m = 0.067 m_o$, where $m_o$ is the free electron mass, and $r_s = 10$ nm. Typical values of $d$ are on the order of 30 nm, so that Eq. (38) is valid, even though this estimate shows that the Landau damping might be important in devices with a smaller separation between the gate and the channel.

5. Instability conditions

Three decay mechanisms oppose plasma wave growth caused by the instability described in Section 4: electron scattering by phonons or impurities, the effect of the finite electron transit time in the channel leading to the "ballistic mobility" given by Eq. (2), and internal friction caused by the viscosity of the electron fluid.

The electron scattering can be accounted for by retaining the term $-v/\tau_m$ in the right-hand side of Eq. (20). This leads to the addition of the $-1/(2\tau_m)$ term to the wave
increment. Hence, the wave grows only if the number of scattering events during the transit time is small.

The viscosity, \( \nu \), of the electron fluid causes an additional damping with the decrement of \( \nu k^2 \) where \( k \) is the wave vector. Hence, the viscosity is especially effective in damping higher order modes. Comparing \( \omega'' \) with \( \nu k^2 \) for the first mode, we find that the effect of the viscosity is small when the Reynolds number \( \text{Re} = L \nu_o / \nu \) is much greater than unity. In a highly non-ideal electron gas, where the Bohr energy, thermal energy, and Fermi energy are of the same order (which roughly corresponds to the surface electron concentration of \( 10^{12} \text{ cm}^{-2} \) at 77 K in GaAs), the viscosity of the electron fluid is on the order of \( \hbar / m \), which is approximately \( 15 \text{ cm}^2/\text{s} \) (comparable to that of castor oil or glycerin at room temperature). The Reynolds number may be estimated as \( \text{Re} = m \nu_o L / \hbar - 12 \) for \( \nu_o = 10^7 \text{ cm/s} \) and \( L = 0.2 \mu \text{m} \) (see Fig. 7).

For a sample with \( L = 0.2 \mu \text{m} \) at 77 K, assuming \( \tau_m \sim 10^{-11} \text{ s} \), the increment \( \nu_o/L \) exceeds the decrement \( 1/(2\tau_m) \) caused by the collisions when \( \nu_o > 10^6 \text{ cm/s} \). For the same sample, the decrement caused by viscosity, \( \nu(2\pi/L)^2/16 \), is smaller than the increment \( \nu_o/L \) when \( \nu = \pi^2 \nu / (4L) \sim 1.8 \times 10^6 \text{ cm/s} \). Hence, the threshold velocity for the instability is well below the peak velocity in GaAs (~2.10^7 \text{ cm/s} \).

The condition \( \nu_o/L < 1/(2\tau_m) \) corresponds to the requirement of a ballistic transport in a transistor channel. In this regime,

\[
\nu_o = \mu_{\text{bal}} U_{ds}/L
\]

(39)

where \( U_{ds} \) is the intrinsic drain-to-source voltage drop across the channel (excluding the voltage drop across the source and drain series resistances) and \( \mu_{\text{bal}} \) is given by Eq. (2) for a non-degenerate case. Hence, the inequality \( \nu_o/L < 1/(2\tau_m) \) can be re-written as
Fig. 8 shows the values of $U_{cr}$ and $v_{ocr} = L/(2\tau_m)$ for GaAs-based HEMTs versus the gate length at 300 K and 77 K. As seen from these figures, the instability condition can be achieved in deep submicron (less than 0.1 micron or so) devices at room temperature and in one to two micron size (or shorter) devices at 77 K.

The first measurement results by Cheremisin point out that such instability might indeed occur in 0.1 micron GaAs-based HEMTs. However, more experimental studies are needed to prove the existence of this instability.

Recently Ryzhii and Shur proposed to combine this approach with using resonant tunneling structure inserted between the gate and a channel of HEMT or a related NERFET-type device (see Fig. 9). Their theory predicts a very large increase of the plasma wave instability increment. Ryzhii et al. also proposed to use a similar structure to enhance the plasma wave detector sensitivity (see Section 6).

$$U_{ds} > U_{cr} = \frac{1.25L}{\mu} \left( \frac{k_B T}{m_e} \right)^{1/2}$$

(40)
For a FET with a resonant tunneling structure between the gate and the channel, Eq. (21) should be re-written as

\[
\frac{\partial n_s}{\partial t} + \frac{\partial (n_s v)}{\partial x} = \frac{n_s}{\tau_e} \Lambda \left( \frac{E_{RT}}{\Gamma} \right)
\]  

where \( \tau_e^{-1} \) is the product of the try-to-escape frequency and the maximum transmission, \( \Lambda(z) = \frac{1}{1 + z^2} \) is the resonance tunneling form-factor, \( E_{RT} = E_{RT} - qU/2 \) and \( \Gamma \) are the energy and position of the resonant tunneling level, and \( a \sim 1 \) is a geometrical factor.

The boundary conditions for this structure were chosen as

\[
U(0) = U(L) = 0
\]  

As was shown in Reference 33, the plasma wave frequencies in such system are given by

\[
\omega_n \sim sk_n
\]  

where \( k_n = \pi n/L \), and \( n = 1, 2, 3 \ldots \)

Fig. 10 shows the increment of the instability calculated in Reference 33 for 0.5 micron gate devices. As seen, this instability should take place in submicron GaAs-based HEMTs with a resonant tunneling structure at 77 K, when the electron mobility is large enough.

Ryzhii and Shur also suggested that a forward gate current in FETs can enhance the plasma instability increment. Electrons constituting this current experience the time delay that typically might be on the order of \( d/v_{eff} \), where \( v_{eff} \) is the velocity of the electrons crossing from the channel into the gate. Such delay might lead to a dynamic negative conductance at plasma frequencies, which should result in the excitation of plasma waves in the transistor channel.

Once the electron velocity exceeds the threshold, the plasma waves grow. Since no other steady states exist for \( v_0 < s \), this growth should lead to oscillations for which the plasma wave amplitude is limited by non-linearity. Nonlinear plasma oscillations in FETs have been studied in References 40, 41, 33.

Let us now discuss possible applications of this instability. The plasma oscillations result in a periodic variation of the channel charge and the mirror image charge in the

---

\( ^{16} \) We should notice that for \( |v_0| > s \), in addition to the uniform flow, there is another steady state solution corresponding to a hydraulic jump in shallow water.
gate contact, i.e. to the periodic variation of the dipole moment. This variation should lead to electromagnetic radiation. The device length is much smaller than the wavelength of the electromagnetic radiation, $\lambda_R$, at the plasma wave frequency. (The transverse dimension, $W$, may be made comparable to $\lambda_R$). Hence, the Ballistic FET operates as a point or linear source of electromagnetic radiation. Many such devices can be placed into a quasi-optical array for power combining. This device de-couples the operating frequency (which can be tuned in a wide frequency range by varying $U_0$) from the electron transit time limitation. The maximum modulation frequency is still limited by the transit time.

![Graph](image)

**Fig. 10.** Increment of fundamental mode versus channel mobility for resonant tunneling HETM structures. 33

### 6. Detectors and mixers of terahertz radiation

The excitation of surface plasma waves can be also used for detection, multiplication, and mixing of terahertz radiation. 15 Weikle et al. 42 and Lu et al. 43 reported on the detection of the microwave and terahertz radiations, respectively, using AlGaAs/GaAs HEMTs. Fig. 11 shows a schematic diagram and an equivalent circuit of a HEMT subjected to terahertz radiation.
The incoming electromagnetic radiation induces an ac voltage at the source side of the channel. The drain side of the channel is an open circuit. (A coupling with a terahertz radiation might be enhanced using an appropriate antenna structure.) This ac voltage excites the plasma waves. As mentioned above, the dc gate-to-source voltage determines the velocity, $s$, of these waves. Because of the nonlinear properties of the electron fluid and the asymmetry in the boundary conditions, a FET biased only by the gate-to-source voltage and subjected to electromagnetic radiation develops a constant drain-to-source voltage, which has a resonant dependence on the radiation frequency with maxima at the plasma oscillation frequency and its odd harmonics. The induced drain voltage, $\Delta U$, is given by

$$\frac{\Delta U}{U_o} = \frac{1}{4} \left( \frac{U_a}{U_o} \right)^2 f(\omega)$$

where

$$f(\omega) = 1 + \beta - \frac{1 + \beta \cos(2k_o'L)}{\sinh^2(k_o'L) + \cos^2(k_o'L)}$$

Here
\[
\beta = \frac{2\omega \tau}{\sqrt{1 + (\omega \tau)^2}}, \tag{46}
\]

and \( k^r_0 \) and \( k^i_0 \) are the real and imaginary parts of the wave vector, \( k_0 \):  

\[
k^r_0 = \frac{\omega}{s} \left( \frac{1 + \omega^{-2} \tau^{-2}}{2} \right)^{1/2}, \tag{47}
\]

\[
k^i_0 = \frac{\omega}{s} \left( \frac{1 + \omega^{-2} \tau^{-2}}{2} \right)^{1/2}, \tag{48}
\]

Here where \( n = 1, 3, 5, 7..., \) \( U_o \) is the gate-to-source gate voltage swing, and \( U_a \) is the amplitude of the ac source-to-gate voltage induced by the incoming radiation. The half width of the resonance curve is determined by the damping of the plasma oscillations caused by the electron momentum relaxation and/or the electron fluid viscosity. Thus, the FET acts as a tunable resonance quadratic detector of electromagnetic radiation.

Lu et al reported on the implementation of a terahertz detector utilizing 2D electronic fluid in a High Electron Mobility Transistor (HEMT). To our knowledge, this is the first plasma wave terahertz detector ever demonstrated. The detector was fabricated using a Fujitsu FHR20X HEMT mounted on a quartz substrate. The device operated at a frequency of 2.5 THz, which is about 30 times higher than the transistor cutoff frequency. A CO\(_2\)-pumped far-infrared gas laser served as a source of terahertz radiation. The laser beam was chopped and focused on the sample with the electric field polarization oriented in the drain-to-source direction. The drain was open, thus the drain current is zero. The gate current is in the order of nano-ampere for the measurement range of gate bias. In agreement with the predictions of our terahertz detector theory, the radiation induced a negative DC drain-to-source bias proportional to the radiation intensity. This voltage was measured using a lock-in amplifier. The measured dependencies of the detector responsivity on the gate bias are in good agreement with the gate bias dependence of the normalized responsivity predicted by the detector theory (see Fig. 12). The responsivity increases at smaller gate voltage swings as predicted by the theory. The dimensionless responsivity function displays a rather broad resonant peak corresponding to the third harmonic of the surface plasma frequency.

More recently, Knap et al demonstrated the resonant detection of sub-terahertz radiation by the two-dimensional electron plasma confined in a submicron gate GaAs/AlGaAs field-effect transistors. Their results confirmed that the critical parameter governing the sensitivity of the resonant detection is \( \omega \tau_m \), where \(\omega \) is the radiation frequency and \(\tau_m \) is the momentum scattering time. By lowering the temperature and hence increasing \(\tau_m \) and increasing the detection frequency \(\omega \), they reached the \(\omega \tau_m \sim 1\) condition and observed a resonant detection of 600 GHz radiation in the 0.15 \(\mu\)m gate length GaAs field-effect transistor (see Fig. 13). The evolution of the observed photoresponse signal with temperature and frequency is well reproduced within the framework of a theoretical model developed earlier.

Let us now consider an HFET detector with a longer channel, such that \( s \tau_m / L \ll 1 \). In this case, the plasma waves are excited by the incoming radiation near the source, provided that \(\omega \tau >> 1\) but they decay before ever reaching the drain side of the channel.
Fig. 12. Measured (symbols) response versus incident power at zero gate bias at 2.5 THz and the laser power of 5.7 mW. The solid line is normalized for comparison. The peak corresponds to the third harmonic of the surface plasma frequency, $\omega_3$. © IEEE, 1998

Fig. 13 a) Experimental drain response, $U_{ds}$, for different temperatures. Inset shows the resonant signal $\Delta U_{ds}$ at 8 K obtained after subtraction of the $1/U_0$ like background. b) Resonant plasma frequency as a function of the gate voltage. The vertical arrow shows the cross point with 600GHz horizontal line – the voltage of the expected resonant detection.
The analysis of this case, based on the solution of Eqs. (44) - (48), shows that the dc voltage will still be developed between the drain and source, and the device will operate as a broad band detector of electromagnetic radiation with the output signal

\[
\frac{\Delta U}{U_o} = \frac{1}{4} \left( \frac{U_m}{U_o} \right)^2 \left[ 1 + \frac{2\omega\tau_m}{\sqrt{1 + \omega^2\tau_m^2}} \right]
\]

(49)

As can be seen from Eq. (49), a long HFET should act as a broad band detector of electromagnetic radiation. The highest frequency of detection is on the order of \(s/d\), where \(d\) is the gate-to-source spacing, since the gradual channel approximation is not valid for the plasma wavelengths shorter than \(d\). For \(s \sim 10^8\) cm/s and \(d \sim 100\) Å, this corresponds to 100 THz.

Weikle et al fabricated a prototype non-resonant detector (operating in the microwave range) using an AlGaAs/GaAs 0.15 micron gate HEMT. The measured dependencies of the detector responsivity on the gate bias and frequency were in good agreement with our theory. AlGaN/GaN HFETs exhibited a non-resonant detection at frequencies much higher than the HEMT cutoff frequency. The measured detector responsivity was in good agreement with the theory and reaches 300 V/W, which is comparable to the responsivity of Schottky diode detectors at these frequencies. This value can be greatly improved by using a proper antenna structure.

More recently, Knap et al reported on an experimental and theoretical study of nonresonant detection of sub-terahertz radiation in GaAs/AlGaAs and GaN/AlGaN heterostructure field effect transistors (see Fig. 14). The experiments were performed in a wide range of temperatures (8-300K) and for frequencies ranging from 100GHz to 600GHz. The photoresponse measured as a function of the gate voltage exhibited a maximum near the threshold voltage. The results were interpreted using a new theoretical model that shows that the maximum in photoresponse and showed explained by the combined effect of exponential decrease of the electron density and the gate leakage current.

![Fig.14. Response at 600 GHz, measured dots, calculated solid lines (a) and drain current versus gate voltage (b) measured in three experiments T1, T2, T3. Results marked as T1 corresponds to the transistor with the threshold voltage \(U_{th} = -0.55\) V measured at 300K. Results marked as T2 corresponds to the transistor with \(U_{th} = -0.42\) V. Results marked as T3 correspond to the same device but measured at temperature of 10 K, for which the threshold voltage was lower \((U_{th} = -0.22\) V).](image-url)
These detectors operated at frequencies well above the cutoff frequency. Knap et al. obtained similar experimental results for the broadband sub-terahertz detectors fabricated using AlGaN/GaN HFETs. Their results showed that the 2D electron plasma effects are indeed universal, not dependent upon a particular material system (see also)

7. Terahertz photomixing using resonant excitation of plasma oscillations

Alternative approaches to generate THz radiation are associated with optical techniques, that use a coherent output at the difference frequency (equal to the difference between the frequencies of radiation emitted by two lasers) or a response of photoconductive structures to femtosecond optical pulses. Fast quantum well infrared photodetectors (QWIPs) utilizing intersubband transitions can also be used for the generation of THz radiation by mixing infrared laser beams. Indeed, as shown theoretically, QWIPs can exhibit a marked response to infrared signals in the THz range of modulation frequencies if the electron transit time is short enough (as can be in single QWIPs) or if electrons reveal a pronounced velocity overshoot after their photoexcitation from QWs. The THz signals produced by coherent plasma oscillations of the photogenerated carriers have been observed in p-i-n structures by Sha et al. and Kersting et al. However, due to a strong damping of the plasma oscillations excited by short optical pulses, only few-cycle THz signals have been observed.

Recently, Ryzhii et al. predicted that modulated infrared radiation can cause the resonant excitation of plasma oscillations in quantum well diode and transistor structures (see Fig. 15 and 16). This effect provides a new mechanism for the generation of tunable terahertz radiation using photomixing of infrared signals. They developed a device model for a Quantum Well Photo Mixer (QWPM) and calculated its high-frequency performance. It was shown that the proposed device can significantly surpass photomixers utilizing standard quantum well infrared photodetectors.

Fig. 15. Schematic view of diode-type (a) and transistor-type (b) QWPM structures without doping of the gate barrier and a band diagram of transistor-type QWPM (c). Arrows indicate electron trajectories (bound-to-continuum photoexcitation followed by transport over collector barrier).
Ryzhii et al.\(^6\) also proposed a QWPM based on QW diode and transistor structures utilizing bound-to-continuum electron transitions, which can be used for photomixing, and evaluated the device operation using an analytical model. It was shown that the QWPM with high electron mobility in the QW channel can exhibit a resonant response to modulated infrared radiation when the modulation frequency is close to one of the plasma frequencies which can be in the THz range. High values of the responsivity in the THz range of signal frequencies indicated that the proposed QWPM and arrays of such devices can be used for efficient generation of THz radiation using photomixing of infrared signals.

![Graph showing characteristic photoelectric gain vs signal frequency for QWPMs with different electron mobilities (different Q factors) in quantum well channel.](image)

**Conclusions**

Plasma wave excitation in submicron field effect transistors and related device structures should allow us to develop a new generation of solid-state terahertz tunable terahertz devices that will find numerous applications in industry, defense, and biotechnology. Recent experimental results demonstrated both resonant and non-resonant detection of terahertz radiation by plasma waves. These results are quite encouraging and indicate that proposed improved structures will be viable. New improved structures will use an enhancement of the plasma wave growth due to static or dynamic negative differential conductivity induced by the negative differential conductivity caused by the gate current.

**Acknowledgment**

This work has been partially supported by ARO (Project Monitor Dr. Dwight Woolard), by ONR (Project Monitor Dr. John Zolper), and by DARPA (Project Monitor, Dr. Edgar Martinez). We are grateful to Professor M. Dyakonov for useful comments and suggestions.
References


T-RAY SENSING AND IMAGING

S. P. MICKAN* and X.-C. ZHANG†

Department of Physics, Applied Physics & Astronomy, and
Department of Electrical, Computer & System Engineering,
Rensselaer Polytechnic Institute, Troy NY 12180, USA.
†Corresponding author. Email: zhangxc@rpi.edu

Terahertz (THz) radiation occupies part of the electromagnetic spectrum between the infrared and microwave bands. Until recently, technology at THz frequencies was underdeveloped compared to the rest of the electromagnetic spectrum, leaving a gap between millimeter waves and the far-infrared (FIR). In the past decade, interest in the THz gap has been increased by the development of ultrafast laser-based T-ray systems and their demonstration of diffraction-limited spatial resolution, picosecond temporal resolution, DC-THz spectral bandwidth and signal-to-noise ratios above $10^4$.

This chapter reviews the development, the state of the art and the applications of T-ray spectrometers. Continuous-wave (CW) THz-frequency sources and detectors are briefly introduced in comparison to ultrafast pulsed THz systems. An emphasis is placed on experimental applications of T-rays to sensing and imaging, with a view to the continuing advance of technologies and applications in the THz band.

Keywords: Ultrafast; Terahertz, THz; T-rays; sensing; imaging.

1. Introduction

Terahertz (THz) radiation can be generated and detected by various systems, each with different output powers, sensitivities, bandwidths and technological implementations. Recently the terahertz band has become more accessible due to ultra-sensitive pulsed sources and detectors based on pulsed laser excitation, which generate and detect picosecond (ps) pulses of free-space THz radiation, dubbed T-rays.¹

T-ray spectrometers are wide-spread in the optics community and are being used in an increasing number of research and industrial applications. T-rays are complemented by a number of other techniques for generating and detecting THz radiation, specifically CW systems, which use electronic or optical methods, as

*Permanent address: Centre for Biomedical Engineering (CBME), and Department of Electrical and Electronic Engineering, The University of Adelaide 5005 Australia. Email: sp-mickan@eleceng.adelaide.edu.au.
discussed in 2. In this chapter we will generally refer to ultrafast pulsed THz as T-rays.

As shown in Fig. 1, THz radiation lies at the boundary between electronics and photonics. The THz bandwidth centers on the frequencies 0.3 to 3 THz, overlapping with high-frequency mm-waves and the long wavelengths of the far-infrared. At this boundary, THz has improved far-field spatial resolution, \( \approx 300 \mu m \), over mm-waves, and reduced Rayleigh scattering, \( \propto \lambda^{-4} \), compared to infrared light.

THz radiation has a spatial resolution limited by wavelength, \( \lambda \), to approximately 0.3 mm at 1 THz, which is better than millimeter (mm)-waves, and THz Rayleigh scattering is less than for IR light due to its \( \lambda^{-4} \) dependence.

Research interest in T-rays stems from the broadband pulsed nature of the radiation and the THz-frequency response of materials, ranging from semiconductors to human tissue. The wavelength of the radiation corresponds to a photon energy, and thus to certain energy transitions in materials. The THz regime is populated by the rotational and vibrational energy states of polar molecules, either in liquid or gas form. Larger molecules show many collective vibrational and librational (bending) modes at THz frequencies. Polar molecules interact strongly with T-rays; water molecules absorb THz very strongly, on the one hand limiting penetration of the radiation into moist substances, and on the other making it readily detectable even in very low concentrations.

T-ray photons have energy quanta corresponding to many discrete energy levels in matter. In molecules, vibrational states are typically separated by energy transitions of approximately 0.1 eV and rotational states are separated by approximately 0.001 eV. Multi-atomic molecules have many vibrational and rotational modes, resulting in a very large number of T-ray transitions for large molecules. In larger systems, the relative vibration of sub-domains in a molecule and the vibration of hydrogen bonds between molecules interact with T-ray photons. In condensed matter systems, typical T-ray energies can excite polaritons, phonons and plasmons in semiconductors, and energy gaps in superconductors.

Most molecules have dense and distinctive absorption spectra in the far-infrared, which has led to much interest in THz spectroscopy. Using T-ray transmission or reflection spectroscopy, samples from gas to the solid-state can be completely characterized at THz frequencies. The density of ro-vibrational modes in the THz
T-Ray Sensing and Imaging

bandwidth provides a wealth of information about the composition and state of samples.

In terms of applications, THz radiation will penetrate non-polar substances such as fats, cardboard, cloth and plastic with little attenuation, and can be used for detecting low concentrations of polar gases, conceivably for pollution control. Apart from detection, materials can also be differentiated spectroscopically using T-rays. The identification of a group of rotational or vibrational lines in the molecular spectrum leads to unique classification of the molecular substance itself.

Table 1. Comparison of different notations used to describe where the T-ray frequencies, from 0.3 to 3 THz, lie on the electromagnetic spectrum.

<table>
<thead>
<tr>
<th>Wavenumber</th>
<th>Wavelength</th>
<th>Frequency</th>
<th>Photon Energy</th>
<th>Blackbody Temp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 cm⁻¹</td>
<td>10 mm</td>
<td>30 GHz</td>
<td>120 μeV</td>
<td>1.4 K</td>
</tr>
<tr>
<td>10 cm⁻¹</td>
<td>1 mm</td>
<td>300 GHz</td>
<td>1.2 meV</td>
<td>14 K</td>
</tr>
<tr>
<td>33 cm⁻¹</td>
<td>300 μm</td>
<td>1 THz</td>
<td>4.1 meV</td>
<td>48 K</td>
</tr>
<tr>
<td>100 cm⁻¹</td>
<td>100 μm</td>
<td>3 THz</td>
<td>12 meV</td>
<td>140 K</td>
</tr>
<tr>
<td>200 cm⁻¹</td>
<td>50 μm</td>
<td>6 THz</td>
<td>25 meV</td>
<td>290 K</td>
</tr>
<tr>
<td>670 cm⁻¹</td>
<td>15 μm</td>
<td>20 THz</td>
<td>83 meV</td>
<td>960 K</td>
</tr>
</tbody>
</table>

T-ray science has expanded rapidly during the past decade, exploring the THz gap in the electromagnetic spectrum. Two broad THz classes of T-ray generation and detection, both relying on ultrafast laser pulses, have developed over the past fifteen years. The first, using photoconductive antennas, was developed at Bell Labs and the IBM Watson Research Center and is now available in a commercial product from Picometrix Inc, MI.9 The second, using the nonlinear effects of optical rectification and electro-optic sampling, is marketed by Zomega Technology Corp, NY.10 Large-scale stand-alone imaging systems became available in 2002 from TeraView11 and Nikon.12 THz radiation is being used for an increasing range of sensing and imaging applications, discussed in Secs. 2 and 5.

T-ray spectrometers are well suited to complement existing chemical sensing and analysis tools, such as Fourier transform infrared (FTIR) and nuclear magnetic resonance spectroscopy. Pulsed THz radiation consists of ultrashort pulses, each with a bandwidth spanning the range from approximately 0.1 to 10 THz. The electric field of a typical T-ray pulse is shown in Fig. 2. The vertical axis shows the magnitude of the pulse, which is proportional to the THz electric field, and the horizontal axis shows the time-delay between the optical generation and detection pulses. The operation of a T-ray spectrometer is explained in Sec. 3. The important characteristic of a T-ray is its short pulse length. This picosecond pulse of electromagnetic radiation, that is 10⁻¹² seconds, has a short time resolution and a broad spectral bandwidth, and time-gated detection is sensitive to its high peak power. Pulsed T-ray systems have demonstrated signal-to-noise ratios (SNRs) of over 10,000:1 using lock-in detection,13 with only small pulse energies. Having very low average power, T-rays are particularly attractive for medical applications, where they are currently...
considered harmless (see Sec. 4.3.4). Looking to future applications, T-ray imaging faces the hurdles of spatial resolution, size, cost, output power, SNR, bandwidth, depth penetration, water sensitivity and speed of data acquisition.

2. Continuous-wave Terahertz Systems

2.1. Introduction

In the THz region of the electromagnetic spectrum, there are five primary classes of CW sources and three classes of detectors. The simplest THz sources and detectors are thermal in nature; hot emitters, such as globars, and cooled detectors are used in THz FTIR spectroscopy systems. The third and fourth most common THz sources have arisen from opposite sides of the THz spectral band; electronic sources are based on microwave-style resonators and circuits, whereas nonlinear optical sources rely on optical laser technology. Three classes of THz laser have been introduced in the last thirty years, and remain in various stages of development: free-electron lasers (FELs), gas-vapor lasers and semiconductor lasers, including quantum cascade lasers.

THz radiation has historically been difficult to generate and detect. Thermal sources have weak emission at long wavelengths and gas vapor lasers are cumbersome. Bolometric detectors only operate under vacuum conditions at liquid-helium temperatures, and although FTIR spectroscopy offers spectroscopic information down to approximately 1 THz, lower frequencies are difficult to access. This section reviews current research on THz sources and detectors that do not rely
on ultrafast generation and detection.

2.2. **Thermal sources**

FTIR spectroscopy is a powerful and well-developed technique for studying resonances from the visible to the mid-infrared. FTIR relies on incoherent thermal sources and detectors for THz-frequency measurements. Despite the low power and poor sensitivity of thermal sources and detectors in the THz band, FTIR is used widely and extensively in material and biological studies.\(^{19}\)

THz FTIR spectroscopy relies on THz radiation generated by a mercury arc lamp, or SiC rod (globar), which is directed into two arms of an optical interferometer. The sample to be studied is placed in one arm of the interferometer, and a characteristic interference pattern, influenced by the THz properties of the sample, is measured by scanning the length of one arm of the interferometer. The actual spectral response of the sample can be calculated from the interferogram using Fourier theory and the numerical Fast Fourier Transform (FFT), with consideration of zero-padding, apodization and aliasing.\(^{17}\)

With regard to speed, frequency resolution and reliance on the FFT, FTIR is comparable to T-ray spectroscopy, introduced in Sec. 3. The main advantage of FTIR spectroscopy is its sensitivity spanning many frequencies, from the far-infrared up to the visible, whereas T-ray spectrometers are more sensitive below 3 THz.\(^{18}\)

<table>
<thead>
<tr>
<th>Technique</th>
<th>Freq. range</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ellipsometry</td>
<td>20 THz–800 THz</td>
<td>optical technique(^{20})</td>
</tr>
<tr>
<td>FTIR</td>
<td>5 THz–80 THz</td>
<td>for bulk and sheets(^{18})</td>
</tr>
<tr>
<td>Resonator</td>
<td>35 GHz–144 GHz</td>
<td>for sheets</td>
</tr>
<tr>
<td>Vector Network Analyzer</td>
<td>50 MHz–100 GHz</td>
<td>requires electrode patterning</td>
</tr>
<tr>
<td>RF Impedance Analyzer</td>
<td>1 MHz–100 GHz</td>
<td>for circuits</td>
</tr>
<tr>
<td>LCR Meter</td>
<td>10 Hz–1 MHz</td>
<td>for circuits</td>
</tr>
<tr>
<td>C-V technique</td>
<td>DC &amp; low freq.</td>
<td>requires Hg probe contacts(^{21})</td>
</tr>
</tbody>
</table>

2.3. **Photonic sources**

The primary difficulty in developing a THz laser is finding a cheap and convenient gain medium that can be pumped efficiently, with high gain and high output power.

2.3.1. **Free-electron lasers**

The gain medium of the free-electron laser (FEL) is an oscillating beam of electron bunches, which is modulated by a spatially periodic magnetic ‘wiggler’ field. The emitted radiation can be tuned by changing the frequency of the wiggler and can in principle reach from mm to X-ray wavelengths. The FEL is the most powerful
source of THz radiation available, producing a CW or pulsed beam of coherent, diffraction-limited THz radiation with an efficiency potentially close to unity.\textsuperscript{22}

Despite requiring large dedicated installations, FELs are important tools for studying THz radiation and its interaction with, for example, condensed matter and biological materials.\textsuperscript{23} Current research aims to produce more compact FEL systems\textsuperscript{24} and up to 50 W of average THz power has been generated at a 75 MHz pulse rate.\textsuperscript{25}

2.3.2. Gas-vapor lasers

As mentioned in the introduction, simple molecules such as water have many rotational and vibrational resonances within the THz region. These molecules can be used as a laser gain medium if efficiently pumped into a long-lived excited state. The first obstacle in creating a THz source is finding a suitable radiative transition, since rotational and vibrational energy levels of large molecules are difficult to use in a laser.\textsuperscript{15} A number of solutions have been found, including methanol and HCN vapor, pumped either optically or electronically. Methanol lasers, for example, are designed around optically exciting the rocking and asymmetric deformation modes of methanol with a CO\textsubscript{2} laser.\textsuperscript{26} HCN lasers are the most popular gas-vapor lasers, with CW or pulsed lasing possible on many modes in the FIR. HCN lasers require large cavities (2 m or more), large voltages and currents (kV with up to 100s of amps), and active temperature stabilization.\textsuperscript{27}

Gas-vapor lasers are not continuously tunable, but have been shown to operate at over 2000 wavelengths from 26 to 254 \textmu m, with power levels ranging from \mu W to 10 mW.\textsuperscript{28} Gas-vapor lasers have been used to characterize many materials in the FIR, including liquids, bulk semiconductors, thin-film semiconductors and thin film superconductors.\textsuperscript{29} Gas-vapor laser have been reviewed by Blake \textit{et al.}\textsuperscript{30} and current developments continue to reduce the size and increase the tunability of gas-vapor lasers.

2.3.3. Semiconductor lasers

The offer of small, cheap and tunable solid-state THz sources has led to two main approaches in developing a THz semiconductor laser: p-germanium (Ge) lasers and semiconductor heterostructure (quantum cascade) lasers. Other solid-state THz lasers are proposed using narrow bandgap lead salts, intersubband quantum fountain oscillations\textsuperscript{31} and impurity centers in silicon.\textsuperscript{32,33}

The Ge laser operates through the electrical excitation of hot holes in p-doped Ge. The laser cavity is created by polishing the surfaces of the Ge crystal since Ge has a high THz refractive index, \textit{n}=4.0.\textsuperscript{34} Ge lasers doped with beryllium have exhibited up to 5\% duty cycles, with up to 1 W of output power, continuously-tunable from 1 to 4 THz. With advances in slab design and cooling, Bründemann \textit{et al.} expect to achieve CW lasing in the near future.\textsuperscript{35-37} Current devices operate at 20–30 K, with 10–20 W of refrigeration. A commercial p-Ge laser was first available in 2002.\textsuperscript{38}

THz radiation emission from quantum heterostructures in semiconductors is based around intersubband transitions.\textsuperscript{39} The first quantum THz laser was reported by Kohler \textit{et al} in 2002,\textsuperscript{40} using a quantum cascade gain medium,\textsuperscript{41} after
many groups had demonstrated THz emission from quantum structures. These devices are pumped with low voltages and mA currents, and generate narrow bandwidth radiation between 1 and 5 THz from intersubband transitions and resonant tunneling processes.\textsuperscript{42-44} Quantum gain media are currently operational only at low temperatures (≈4.2 K) and suffer from fast depopulation of the excited states,\textsuperscript{45} although room-temperature emission is predicted from 4-level asymmetric quantum well structures.\textsuperscript{46,47} Quantum cascade structures have been used to generate THz in a nonlinear mixer configuration.\textsuperscript{48} THz-frequency quantum cascade laser research has been reviewed by Tredicucci \textit{et al.}\textsuperscript{49}

\subsection*{2.4. Nonlinear optical processes}

Nonlinear optical materials with an appreciable $\chi^{(2)}$ nonlinear optical coefficient provide a medium for generation of radiation at the beat frequency between two pump sources.\textsuperscript{50,51} For pump sources with wavelengths offset by THz frequencies, the $\chi^{(2)}$ material can act as a source of THz.\textsuperscript{52} Difference frequency generation (DFG), or three-wave mixing, was an early source of THz, but was hampered by low conversion efficiencies and phase matching requirements.\textsuperscript{53-57} DFG has been revisited as a THz source using the patented organic crystal DAST (4-dimethylamino-N-methyl-4-stilbazolium-tosylate), designed to have a very high nonlinear electro-optic (EO) coefficient and low dielectric constant in the FIR.\textsuperscript{58} DAST has been used to mix quasi-CW pulses from a tunable dual-wavelength Ti:sapphire laser to generate 32-ns THz pulses with over 2 $\mu$W of THz power.\textsuperscript{59} A more compact and lower power source has been demonstrated using a Nd:YAG pump and dual wavelengths generated in a periodically-poled lithium niobate (PPLN) optical parametric oscillator;\textsuperscript{60} the THz output wavelength is tuned from 120 to 160 $\mu$m using temperature control of the PPLN crystal. THz beat frequencies are now available from dual-frequency laser diodes\textsuperscript{61} and injection-locked semiconductor lasers.\textsuperscript{62} DFG is closely related to optical rectification, described in Sec. 3.3.2, where THz is generated as the difference frequency between Fourier components of the pulse.\textsuperscript{52}

A nonlinear optical crystal can also be used to generate THz radiation via a parametric effect. In optical parametric generation (OPG), a near-infrared (NIR) pump beam generates a second NIR idler beam in a nonlinear crystal, and THz radiation is generated from the beating of the pump and the idler.\textsuperscript{52} If the nonlinear crystal is placed in a cavity for the idler beam, feedback amplifies the idler beam, creating an optical parametric oscillator (OPO). OPG was used to generate THz radiation in the early 1970s, using a Nd:YAG pump beam, polariton scattering in LiNbO$_3$ to generate the Stokes (down-shifted) idler, and angular rotation of the parametric oscillator cavity to tune the THz output wavelength.\textsuperscript{63} Since the mid-1990’s, Kawase \textit{et al.} have made a number of improvements to what they call the THz parametric oscillator (TPO), thereby decreasing system size, enhancing wavelength selectivity and increasing the output power by several orders of magnitude. A grating coupler on the LiTaO$_3$ crystal was used to improve the coupling of the THz radiation into free space,\textsuperscript{66} and a prism coupler was used to remove the dependence of the THz output direction on the angle of the parametric cavity (and thus the output wavelength).\textsuperscript{66} Rapid wavelength tuning was achieved by scanning the angle of the pump beam with a fast mirror.\textsuperscript{67} The TPO has also been implemented in
a ring cavity configuration and in cryogenically-cooled crystals, which improves the THz output power by over 100 times and reduces the gain threshold. In 2001, THz parametric generation (TPG), with an injection seeded idler beam, was experimentally shown to be superior to TPO. TPG shows a reduced line width ($\Delta \nu / \nu \approx 10^{-4}$) and a peak THz power of over 100 mW for 3.4-ns pulses. TPG has also been demonstrated in a compact form, admittedly with limited tunability (1.2–2.4 THz) and power (30 $\mu$W).

Nonlinear optical methods of generating CW THz radiation are attractive for their simplicity, availability of solid-state pump sources, linewidth and tunability. Although more cumbersome than proposed electrically-pumped solid state sources, discussed in Sec. 2.3.3 above, DFG and OPG systems are significantly smaller than free-electron and gas-vapor lasers.

2.5. Photomixing in biased semiconductors

In a manner similar to DFG, biased semiconductors with very short carrier lifetimes and high carrier mobility can be used to convert a THz beat frequency between two optical beams into THz radiation.

Brown et al. initially developed GHz and THz photomixers, or optical-heterodyne converters, using electrodes and antennas deposited on low-temperature-grown (LT)-GaAs. LT-GaAs is characterized by sub-picosecond electron-hole recombination rates, a high carrier mobility ($\mu \approx 200$ cm$^2$/V-s) and a high breakdown field ($>4 \times 10^5$ V/cm). The photomixer was an array of interdigitated electrodes with micron spacing, connected either to a strip line, for GHz generation, or a spiral antenna for free-space radiation of THz-frequency radiation. Initially limited to 200 GHz, the bandwidth was increased to above 5 THz of coherent, tunable THz radiation, and demonstrated for use in a spectrometer. The biased photomixer is very similar to a photo-conductive antenna (Sec. 3.3.1), which is used with pulsed optical excitation. Photomixing of fiber-coupled beams has been demonstrated with dipole, log-spiral and slot antennas at different temperatures, and analyzed with theoretical modeling.

A number of optical sources have been used to generate the two pump beams for mixing, including two fiber-coupled Ti:sapphire lasers, two distributed-Bragg-reflector (DBR) laser diodes, two longitudinal modes in laser diodes, microchip lasers (for mm-waves) and iodine lasers (up to 14 GHz). The advantage of using two modes in a single laser diode is the compact nature of the source, and a very narrow linewidth. Other dual-mode lasers have been used to generate THz-frequency amplitude beats, for example a Ti:sapphire ring cavity, a Ti:sapphire $\alpha$-cavity and a laser diode array with a bandwidth reaching 7 THz. A novel method of tuning the THz beat frequency is to use two linearly chirped pulses, and control the phase difference between them at the photomixer with, for example, a Michelson interferometer. Biased photoconductors are also very important in generating THz pulses from ultrashort optical pump pulses, as discussed in Sec. 3.3.1 below.

CW photomixing systems have been used for scanning imaging, analogous to T-ray imaging systems, as detailed in Sec. 5 below. With CW THz power output levels of approximately 1 mW, photomixing systems are useful in applications.
requiring excellent tunability, small system size and narrow linewidth.

2.6. Electronic sources

The frequency of electronic sources has increased steadily for many years, from MHz to the GHz frequencies. State-of-the-art electronic sources remain limited to frequencies in the hundreds of GHz and sub-mW powers. High-speed electronic devices are suited to low-power CW operation, although a non-contact system based on complex impedance bridges has been used to accurately characterize the dielectric constant of thin films between 30 GHz and 1 THz. Nonlinear transmission lines (NLTLs) have also been used to generate electronically-tunable sub-ps transients, which can drive antennas to generate THz radiation. NLTL THz sources have demonstrated very narrow linewidths in room-temperature spectroscopy applications.

Backward wave oscillators (BWOs) are electrically-driven microwave generators that can generate CW radiation from mm-waves up to 2 THz. BWOs are large table-top devices, requiring powerful magnetic fields and water cooling, but provide up to 300 mW of polarized, narrowband (Δν/ν = 10⁻⁵), tunable THz radiation. Each BWO is tunable within approximately ±30% of its central value, therefore a number of devices are required for full frequency coverage. BWOs have been used in THz spectroscopy of gases, thin film superconductors, and biomolecules.

Electronic CW generation techniques are primarily of interest for their simplicity and integratability with other electronic systems. The BWO provides a high quality source of CW THz at power levels between electronic and large laser systems.

2.7. Detection

As with sources, THz radiation can be detected using thermal, optical or electronic effects. Detecting THz signals is difficult because blackbody radiation at room temperature is strong in the FIR, as shown in Table 1. The most common THz detectors are thermal detectors, for example the helium-cooled bolometer, which is desensitized to ambient temperature and registers only the heating effect of the THz radiation. The bolometer is an incoherent detector, registering only incident power, which is appropriate for CW detection, but limits the information available in broadband systems. FTIR, for example, uses an interferometric technique to resolve the different frequency components of the broadband source. In an analogous manner to nonlinear optical mixing and photomixing in biased semiconductors, CW THz radiation can be coherently detected using homodyne mixing in nonlinear crystals and photoconductive switches. For CW detection with THz local oscillators, electronic detectors have been developed based on cooled photodetectors, electron plasmas in semiconductors, semiconductor superlattices, mesoscopic quantum devices, resonant quantum well infrared photodetectors (QWIPS) and high-electron-mobility transistors (HEMTs). Purely electronic systems incorporate symmetrical sources and detectors.

THz imaging with CW systems is widely relevant in FIR astronomy, as reviewed by Siegel. CW imaging has been demonstrated with QWIP arrays and in laboratory conditions using photomixing.
3. Pulsed T-ray Spectrometers

3.1. Introduction

Pulsed THz, or T-ray, spectrometers are based on ultrafast pulsed generation and detection using mode-locked lasers. Mode-locked Ti:sapphire ultrafast lasers became commonly available in the 1980s, triggering increased research into ultrafast sources and detectors of THz radiation. The cost and size of T-ray spectrometers is dominated by the cost and size of the ultrafast laser, a technology that is continuously becoming more convenient, compact and economical.

T-ray spectrometers have come out of the lab in recent years and are now commercially available as entire systems. The designs of T-ray systems continue to increase in variety as they are improved and applied to many fields in science, medicine and industry.

3.2. Pulsed systems

The development of pulsed THz radiation (T-rays) generated using ultrafast optical lasers has provided a new method for accessing the THz frequency regime. Pulsed THz techniques were initially developed for waveguide and circuit characterization. Free-space THz time-domain spectroscopy grew from the development of both photoconductive dipole antennas and EO crystals as sources and detectors. The first complete T-ray spectrometers were used for spectroscopy of bulk dielectrics at THz frequencies. Many varieties of T-ray spectrometers now exist, detailed in the following sections. The most common is the commercially available spectrometer using an PCA emitter and detector.

T-ray spectrometers are inherently broadband systems, a result of using ultrafast optical pulses to generate pulses of THz radiation. The broadband nature of T-rays means they are able to probe a wide region of the FIR, depending on the actual system, from DC to the mid-infrared (>30 THz).

3.3. Generation

Many methods have been exploited to develop sources of pulsed THz. Apart from the most common techniques of photoconductive switching, optical rectification and semiconductor surface current generation, T-rays have been generated using the inverse Franz-Keldysh effect, coherent control concepts and plasma oscillations. These are methods of producing an ultrafast current or polarization transient, which acts as a broadband source as predicted in Hertz’s equations. The radiation is then directly emitted into free space or coupled out with an antenna.

T-ray generation has been reviewed in many publications, for example by Gornik and Kersting. The state-of-the-art in 1988 is well covered by Auston and Nuss and nonlinear generation is developed in a text by Shen.

3.3.1. Photoconductive antennas

A photoconductive antenna (PCA) has two important features: an antenna structure and a photoconductive substrate. The antenna structure is designed to radiate sub-mm-wavelength electromagnetic waves into free space. The antenna is driven
by an ultrafast current transient in the switch consisting of photocarriers swept by an applied DC bias field. The photocarriers are generated in the photoconductive substrate, which acts as a switch, by an ultrafast laser pulse. A PCA can be designed with different substrates and different actual antenna designs, which in turn influence the possible T-ray bandwidth and output power. When an ultrafast optical pulse is absorbed by an appropriate substrate, carriers are generated and swept apart by the DC bias field, which creates the current transient to drive the antenna. The substrate must be able to have sufficiently fast carrier recombination time and high carrier mobility to support an ultrafast transient. A sub-picosecond transient will generate a free-space pulse with a duration of a few picoseconds, which corresponds to a THz bandwidth.

![Elementary T-ray beam system with PCA emitter and detector](image1)

![PCA output power dependence on optical pump power, with DC bias field as a parameter](image2)

Fig. 3. A T-ray beam is generated and detected using two PCAs. The ultrafast probe beam pulse shorts the biased electrodes and the resulting current transient is coupled into free space. Gold or Al-coated parabolic mirrors collimate and focus the T-rays onto a second PCA. The PCA detector, described in Sec. 3.4.1, is gated by the probe pulse; when the PCA is conducting, a current can be measured across the electrodes that is directly proportional to the amplitude of the T-ray electric field. Fig. 3(a) shows how the emitted T-ray electric field depends both on the power in the optical pump pulse and the DC bias field applied to the PCA electrodes. The T-ray output saturates for high levels of optical fluence. The DC bias is limited by breakdown in the photoconductive switch.

The first PCA was demonstrated by Mourou et al. at microwave frequencies in 1981. PCA sources used in T-ray spectrometers, or THz time-domain spectroscopy (THz-TDS) systems, were developed primarily by Lucent Technologies and IBM in 1988 and 1989, although other groups were actively developing sources and detectors, as reviewed by Auston et al. A schematic of the first T-ray spectrometer, based on PCAs, is shown in Fig. 3(a). The basic elements of a PCA, the antenna geometry, the photoconductive switch, the electric bias and the optical pulse have been subsequently varied in many ways to study and improve the generated T-ray power, bandwidth, radiation pattern and pulse characteristics.

The high SNR of T-ray spectrometers is due to high emitter power and sensitive
detection (Sec. 3.4). The average T-ray power in a focused beam has developed from nW to μW levels, with current systems achieving 30–40 μW of average THz power. The availability of high power, high repetition rate femtosecond laser sources has enabled PCAs to be driven into saturation, and photoconductive substrates with high breakdown voltages permit large bias fields.

The maximum power emitted from a PCA depends on the photoconductive substrate and the coupling efficiency of the antenna. THz power scales with both the optical pulse power and the DC bias field, as indicated in Fig. 3(b). A photoconductive switch material should have a high breakdown voltage, a low optical refractive index, low bandgap, low carrier lifetime, high optical absorption and high carrier mobility. The polarization of the normally-incident optical pulse is perpendicular to the bias electrodes. High power PCAs were demonstrated initially on implanted silicon-on-sapphire and low-temperature-grown (LT) GaAs, and since with semi-insulating GaAs, ion-implanted GaAs and InGaAs. Antenna design and coupling influences the efficiency, bandwidth and radiation pattern of T-ray emission. Many antenna geometries have been explored, however high power T-rays are still generated with coplanar strip lines and large aperture emitters.

A broad bandwidth is the second most desirable characteristic of a T-ray source, which relies on a short THz pulse duration. PCAs in GaAs typically have a useful bandwidth extending from less than 100 GHz to 2 THz, which can be extended to 4 THz by injecting carriers close to the band edge. Bandwidths up to 6 THz have been reported for PCAs, but pulse duration continues to be limited by carrier mobility, leading to interest in optical rectification emitters (Sec. 3.3.2). T-ray bandwidth can be controlled by shaping the optical pulse, to either tune the output frequency or increase the overall output power.

The radiation pattern of a PCA is important for designing beam-steering optics and imaging systems. Generally, THz radiation is coupled out from the antenna and collimated into a beam with a hemispherical lens. The radiation pattern for the common dipole antenna is essentially dipolar, with a weak quadrupole component perpendicular to the bias field, weak elliptical polarization, and propagates as a Gaussian into the far-field with high frequency components concentrated in the center of the beam. Near-field effects are important for understanding screening of the bias field, pulse propagation and applications in near-field imaging, and are detailed in Sec. 5.6.

PCA emitters have been integrated into commercial optical fiber-coupled T-ray systems, for their convenience and high power. Miniature PCAs have been demonstrated on optical fibers for potential endoscopic applications. T-ray emission has been observed from air as the photoconductive switch and PCAs have been incorporated directly into ultrafast laser cavities to produce up to 7 μW of pulsed THz power.

### 3.3.2. Optical rectification

Optical rectification (OR) is a nonlinear optical effect that can generate T-ray pulses with very large bandwidth from ultrafast optical pulses. OR is a second-order effect that occurs in materials with a non-zero χ(2), or EO, coefficient, and is referred
Fig 4. OR is a second-order nonlinear effect, whereby an ultrafast electric field pulse is rectified in a $\chi^{(2)}$ medium. The ultrafast pump pulses induce an transient polarization, $P(t)$, which in turn emits a THz-bandwidth pulse. The time evolution of the THz pulse is given by the second time derivative of the polarization transient.

$E_{\text{THz}}(t) \propto \partial^2 P(t)/\partial^2 t$

$\Delta \tau \Delta \omega$

$pump beam$

$\chi^{(2)}$

emitted T-rays

to as the inverse EO, or Pockel’s effect.\textsuperscript{52,179,180} OR was first observed in 1962, with a high intensity 100-ns laser pulse causing a DC polarization in certain crystals.\textsuperscript{181} With the development of ps and fs laser systems, faster polarization transients were induced, with bandwidths reaching to THz frequencies.\textsuperscript{182-186} The ultrafast ‘shock wave’ propagating through a nonlinear medium was observed to generate Čerenkov THz radiation, but was confined to the medium due to total internal reflection.\textsuperscript{187-189}

After the development of the PCA for free-space THz generation, the Čerenkov radiation was coupled out of LiTaO$_3$ into THz beams.\textsuperscript{129} The power and bandwidth of T-rays generated by OR are determined by the driving optical pulse, the nonlinear $\chi^{(2)}$ coefficient of the crystal, the crystal damage threshold, the output coupling constraints and the visible-to-THz phase matching.\textsuperscript{177,190}

Unlike a PCA, the THz power generated by OR is due only to the incident laser power.\textsuperscript{191} The maximum optical power is limited by available ultrafast laser sources and by damage to the medium.\textsuperscript{192} With the development of high-power solid-state mode-locked lasers, sub-100-fs pulse trains can be generated at 10s of MHz repetition rates and up to 2 W of average power. The generation efficiency depends on the magnitude of the $\chi^{(2)}$ coefficient, THz absorption in the material and phase matching between the optical and THz pulses.\textsuperscript{193,194} The magnitude of the $\chi^{(2)}$ coefficient varies with the crystal cut and orientation.\textsuperscript{176,195-197} Although DAST has a very large EO coefficient,\textsuperscript{198,199} it is hygroscopic. The most popular EO material is ZnTe, because of its physical durability and excellent phase matching.\textsuperscript{177} With a broadly focused optical pump beam, to avoid damaging the medium, and a ZnTe crystal, nW T-ray average power can be generated by OR. OR saturation due to second harmonic generation of the pump beam at high optical fluence has been studied in ZnTe.\textsuperscript{192} OR saturation has also been studied in LiTaO$_3$, LiNbO$_3$ and DAST.\textsuperscript{191}

OR owes its intrinsically broad bandwidth to Pockel’s effect, which operates on a fs time scale.\textsuperscript{179} The actual bandwidth depends on the duration of the optical pump pulse. With ultrashort optical pulses, down to 15 fs in duration, THz pulses have been generated with spectra extending to the mid-infrared.\textsuperscript{200,201}

OR effects have been observed in air,\textsuperscript{202} in biased quantum wells (QWs),\textsuperscript{203}
periodically-poled LiNbO₃ (PPLN), in poled polymers, in polymer thin films, in superconducting thin films, and from beating in coupled quantum wells. T-rays generated from OR in PPLN can be tuned by the poling period, the temperature or the crystal orientation. Typically, OR is driven by ultrafast pulses from a solid state Ti:sapphire mode-locked laser, although 1.55-μm light has been used in optical fiber-based systems.

3.3.3. Pulsed photomixing

Pulsed photomixing in biased semiconductors is related to THz generation in PCAs (Sec. 3.3.1), CW photomixing (Sec. 2.5) and CW nonlinear DFG (Sec. 2.4). Photomixing in a PCA uses two pulses from an ultrafast laser, split in an interferometer with consequent phase control, which provides control over the output T-ray pulse shape. Shaping of the optical near-infrared pump pulse enables T-ray pulse shaping generated from nonlinear optical mixing effects. Mixing pulses with an equal frequency chirp creates a difference frequency that depends on the phase delay of the two pulses in the arms of the interferometer. Chirped pulse mixing therefore generates relatively narrowband, frequency-tunable T-rays. Photomixing has been observed in quantum well structures using resonant excitation of plasmons.

3.3.4. Photoexcited semiconductors and superconductors

![THz generation from surface fields in unbiased semiconductors](image)

Fig 5. T-ray generation from surface fields in unbiased semiconductors. The intrinsic bias at a semiconductor surface is shown in Fig. 5(b). This bias field sweeps ultrafast photo-generated carriers, and for high-mobility semiconductors, the resulting current transient acts as a source of THz. Fig. 5(a) shows THz generation from a semiconductor surface either in transmission or reflection. \( \theta \) is set at Brewster's angle to improve coupling of the optical and THz beams at the air-semiconductor interface.
T-ray emission has been observed from unbiased semiconductors, and attributed to a number of effects. Notably among these are OR, photocarriers accelerated in the semiconductor surface field, coherent polarization oscillations,\textsuperscript{222} coherent phonons,\textsuperscript{223–226} coherent plasmon oscillations,\textsuperscript{227,228} coherently controlled photocurrents,\textsuperscript{135} transitions in coupled quantum wells,\textsuperscript{229} intersubband transitions in quantum wells,\textsuperscript{230} coherent charge oscillations in quantum wells,\textsuperscript{231} Rabi oscillations,\textsuperscript{232} Stark wave packets,\textsuperscript{233} coherent Bloch oscillations in superlattices,\textsuperscript{234–238} photo-Dember fields\textsuperscript{239,240} and superluminal ionization fronts.\textsuperscript{136}

T-ray emission has been observed in high-$T$ superconductors due to ultrafast modulation of the superconductivity.\textsuperscript{241}

T-ray emission from unbiased semiconductor surfaces excited with ultrafast optical pulses was first observed in 1990 by Zhang et al.\textsuperscript{220,221} Fig. 5 shows a schematic of THz generation from an unbiased semiconductor surface and the intrinsic bias field at the surface. The radiated THz field is proportional to the optical pump power and is due to a combination of OR,\textsuperscript{176,242} bulk DFG\textsuperscript{243} and photocarrier acceleration in the surface field of the semiconductor.\textsuperscript{244,245} T-ray emission from semiconductor surfaces is influenced by the magnetic field surrounding the emitter,\textsuperscript{246,247} which can be used to switch or enhance the generation efficiency.\textsuperscript{248–252} T-ray emission from an unbiased GaAs wafer in a switchable magnetic field is shown in Fig. 6, indicating enhancement and phase reversal of the generated THz pulse. This enhancement has been explained by considering the alignment of the induced radiating dipoles and total internal reflection at the semiconductor-air interface.\textsuperscript{253,254}

Optical pulse shaping and magnetic field control enables both enhancements and spectral control of the emitted T-rays.\textsuperscript{255,256} T-ray emission from semiconductor heterostructures has been coherently controlled by shaping the optical pump pulse.\textsuperscript{257}
3.4. Detection

Although CW detection techniques can be used for T-ray detection, the high SNR of T-ray spectrometers is due to time-gated detection (pump-probe) techniques. Gated detection relies on an optically switched detector, and part of the pulse split from a mode-locked laser, as illustrated in Fig. 7.

A feature of PCA and EO detection is the large DC background in the probe compared to the modulation due to the THz field. A common method to observe a T-ray signal is to AC modulate the free-space THz beam at a certain frequency, indicated by the chopper in Fig. 7, then electronically filter out any signals not at this modulation frequency. This modulation frequency is pushed as high as possible because laser noise is greater at low frequencies (it has a $1/f$ characteristic), up to around 3.5 MHz, where the noise floor is set by instrumentation and photon noise.

Gated detection systems acquire a time-domain response by scanning the time delay of the detector across the generated waveform. For T-ray generation, the detector operates on a fs time scale, which is orders of magnitude faster than the ps duration of the THz pulse that is being sampled. This delay is achieved with a variable path length, represented by the delay stage in Fig. 7. The total length of the delay stage movement determines the duration of the sampled pulse. The speed at which the waveform is acquired is limited by the desired SNR. A faster scan leaves less time for the signal at each point to be averaged with, for example, a lock-in amplifier, thus increasing the noise level.
The two major methods of T-ray detection, as with generation, use PCAs or EO crystals.

3.4.1. Photoconductive sampling

Photoconductive sampling was developed in conjunction with PCA emitters. For T-ray detection, an unbiased PCA is placed in the T-ray beam path and gated with an optical probe pulse. A PCA in detection configuration is shown in Fig. 8(a). The gating pulse allows current to flow in the PCA, which is connected to an ammeter. The THz electric field biases the PCA, and the current is therefore proportional to the T-ray field. The optical probe pulse has a far shorter duration than the T-ray pulse, so the T-ray waveform is sampled in time by changing the time delay of the two optical pulses. The detected T-ray signal is a convolution of the incident T-ray waveform and the response of the PCA. In spectroscopy experiments, the detector and emitter responses are accounted for by deconvolution, or signal normalization with a reference pulse, as discussed in Sec. 4.2.

Initially fabricated on LT-GaAs, PCA detectors achieved a maximum bandwidth
of \approx 2\,\text{THz}.^{127,139} Recent experiments with ultrafast gating pulses of 15-fs duration have extended the detection bandwidth to 40\,\text{THz}.^{263,264} LT-GaAs PCAs can be gated with 1.55-\mu\text{m} wavelength light through a two-photon absorption process.^{265}

3.4.2. Electro-optic sampling

Fig. 9. Crystal thickness determines signal sensitivity and bandwidth in EOS. The two figures above show a T-ray pulse sampled with two ZnTe detectors. Typically a 2-mm-thick ZnTe crystal provides a high signal, but a bandwidth limited to under 10\,\text{THz}. ZnTe is a good EO sensor because of phase matching between the THz and optical pulses. Ultra-broadband phase matching is, however, not possible, so a thinner crystal detects higher frequencies. Using a 10-\mu\text{m}-thick crystal, frequencies up to 70\,\text{THz} have been measured.^{266}

Electro-optic sampling (EOS) is a broadband T-ray sampling technique. Like photoconductive sampling, the detector is gated by a time-delayed probe pulse. The detector is an EO crystal, placed between crossed polarizers, and the optical and THz beams propagate collinearly through it, shown in Fig. 8(b). The incident T-ray pulse modulates the birefringence of the crystal, through Pockel's effect.^{52,179,195} The induced birefringence in the crystal is proportional to the electric field strength of the T-ray pulse, and causes a rotation of the polarization of the optical pulse. The crossed polarizers effect an amplitude modulation of the optical pulse, which is proportional to the polarization modulation and thus the THz field strength.

Initially used in the characterization of high-speed electronic circuits,^{120,121,269-271} EOS was first used for free-space THz detection in 1995 by Wu and Zhang.^{130} Although more difficult to align experimentally than PCA detectors,^{164,197,272-275} EOS soon demonstrated a wide bandwidth and high sensitivity.^{193,194,276-279} The sensitivity of EOS depends on the EO \chi^{(2)} nonlinear coefficient of the detection crystal, which must be transparent to the probe beam, and the quality of phase matching between the T-ray and optical pulses.^{164}

The central trade-off for EOS is the thickness of the crystal. A thicker crystal provides a longer interaction length for the coupled pulses, hence a larger signal. However, since perfect phase matching between the optical and THz pulses is not possible, a thinner crystal with less dispersion and provides a broader detection
Fig 10. The chirped pulse single-shot T-ray pulse detection technique. This is similar to a typical T-ray detection system, using OR and EOS, except the probe pulse is stretched to a ps duration with a grating pair. When the T-ray and probe pulses propagate collinearly through the EO detector, they both have approximately the same duration, thus different parts of the probe pulse experience different intensity modulation in the EOS. As the pulse is chirped, these different parts are separated with a spectral grating, and detected with a linear photodiode (PD) array. Apart from a background measurement of the unmodulated chirped pulse required to extract the T-ray waveform, the technique can measure the entire T-ray waveform in a single shot. The T-rays are focused from emitter to detector in this scheme with a polyethylene lens, which had a very low dielectric constant for THz. For single-shot imaging, discussed in Sec. 5.4, the T-ray beam is spatially expanded in 1D and the PD array is replaced with a CCD.

Fig 11. These results demonstrate the linearity of ZnTe as a T-ray sensor, to both the probe beam power and the T-ray electric field. The linearity is (<2%) for over more than 5 orders of magnitude.
bandwidth. Fig. 9 shows T-ray waveforms detected with an EO crystal, indicating the increased bandwidth but reduced signal amplitude for a thinner detector. Using sub-100-\(\mu\)m-thick crystals, mid-infrared pulses are detectable. The most popular crystal used for THz EOS is ZnTe due to its high EO coefficient, low group velocity mismatch between THz pulses and optical probe pulses, and good mechanical properties. GaP, GaAs, multilayered EO polymer films, poled polymers and DAST are among alternative materials used for EOS.

A primary advantage of EOS is its extension to 2D imaging. With a larger EO crystal and an expanded optical beam, the intensity modulation across the transverse T-ray beam can be imaged with a CCD. Rapid 2D, 3D and single-shot T-ray imaging is discussed in Sec. 5.4.

**Single-shot** detection is fundamentally different to normal gated detection in that the entire T-ray waveform is measured by each probe beam pulse, and then detected spatially. The ps time duration of the T-ray pulse is transferred to the spatial domain using non-collinear propagation of the T-ray and probe beams in an EO crystal or collinear propagation of a chirped probe beam. In chirped probe detection, the spatial distribution of the T-ray pulse along the propagation axis in the EO crystal is mapped onto the different wavelengths of a chirped probe pulse, and then separated with an optical grating or streak camera. A schematic of chirped pulse detection is shown in Fig. 10. Chirped pulse detection can be designed with automatic background cancelling with dual photodetector arrays, or extended to 1D imaging with a CCD. Single-shot detection is used for electron bunch measurements in FELs and rapid imaging applications (Sec. 5.4).

### 3.4.3. Magneto-optic detection

Magneto-optic detection is analogous to EOS in that the transient magnetic component of a propagating free-space THz pulse can be detected using crossed polarizers and a magneto-optic crystal. The THz magnetic field is detected using the Faraday effect in an optical medium.

### 3.5. T-ray propagation

#### 3.5.1. Propagation & filters

With the development of free-space T-ray emitters and detectors, interest increased in the propagation of THz-bandwidth pulses in a vacuum and through quasi-optical elements. T-ray systems use quasi-optical collimation and focusing elements, including gold-coated off-axis parabolic mirrors, silicon hyperhemispherical lenses, anti-reflection coatings, variable phase polarization compensators, reflective gratings, teflon prisms, and transmissive optics fabricated from silicon or polyethylene. T-rays have been used to study single cycle pulse propagation, the Gouy phase shift at a focus and THz whispering-gallery modes in cylinders. Spatial filters are fabricated using metal, which has a very high absorption at THz frequencies, including high-pass filters with metal slits, apertures, band-pass filters, low pass filters of glass beads in polyethylene, filter cascades and dichoric filters. Fresnel lenses have been modeled and constructed from silicon to allow frequency-dependant focusing of the broadband T-ray beams. A silicon binary lens for THz wavelengths is pictured in Fig. 12(a) and the sharp THz
focus is shown in Fig. 12(b). T-ray beam propagation has been studied in scattering media\textsuperscript{317} and simulated with finite-difference time-domain techniques.\textsuperscript{318}

![Fresnel lens](image)

(a) 8-level silicon Fresnel lens

![Focused T-ray electric field](image)

(b) Focused T-ray electric field

Fig 12. A Fresnel lens constructed from high-resistivity silicon demonstrates sharp focusing of the terahertz beam. This is comparable to a conventional refractive THz lens.\textsuperscript{316}

The study of T-rays propagating through air and optical elements has been greatly enhanced by techniques for imaging the T-ray emission patterns and beam profiles using either scanned PCAs (Sec. 5.2) or EO detection (Sec. 5.6).

Dynamic T-ray filters have been demonstrated using free carrier generation in a semiconductor to block, or reflect, the THz pulse. This transient mirror can be used to slice up very short duration sections of a T-ray pulse,\textsuperscript{319} or reflect a subwavelength diameter cross-section of the beam for near-field studies, as discussed in Sec. 5.6.

### 3.5.2. Transmission and reflection

T-ray spectrometers can be used either in transmission or in reflection.\textsuperscript{320} The T-ray pulses are detected after transmission through a sample, or reflection from its surface. The mode used depends on the type of sample being studied. For samples with very high absorption, no transmitted signal is detectable. However, normalizing the system response is more difficult with a reflective system, because a reference pulse must be taken without the sample present, and it is difficult to maintain exactly the same free-space path length between the emitter and detector. Alternative geometries used for T-ray spectroscopy and analysis techniques are discussed in Sec. 4.2.
3.5.3. Transceivers

In an extension of the standard T-ray reflection system, the roles of emitter and detector can be combined into one device, either a PCA or an EO crystal. In a T-ray transceiver, the pump and probe pulses travel collinearly with opposite polarization and an adjustable time-delay to reflect off a sample, as shown in the schematic in Fig. 13. The EO transceiver extends into free space the concepts of early spectroscopy with THz confined to a single crystal by total internal reflection.

3.5.4. Waveguides

THz propagation in waveguides is related to microwave technology in the same way that focusing and filtering in free space are related to optical techniques. Waveguide propagation is important in studying near-field T-ray devices, THz interconnects, THz cavities and ultra-sensitive T-ray spectroscopy.

4. Sensing

4.1. Introduction

T-ray sensing involves applying T-ray techniques to the study of materials by monitoring transmitted or reflected radiation. The THz spectrum is populated by energy transitions of less than 0.1 eV, which correspond to different rotational and torsional states of a whole molecule. For larger molecules, the FIR has fewer transitions.
than the near-infrared, which can be very densely populated. The THz bandwidth corresponds to energy transitions in superconductors, plasma states, lattice vibrations and other resonances in crystals.\textsuperscript{337}

Pulsed THz spectroscopy is a coherent technique, where both the amplitude and phase of the THz pulse are measured. Coherent detection enables direct calculation of absorption and refraction profiles without using the Kramers-Kronig relations. T-ray spectrometers provide very high SNRs and a broad bandwidth, making them attractive to sensitive spectroscopic studies on the THz regime. T-ray spectroscopy builds upon a rich history of sub-mm and FIR spectroscopy.\textsuperscript{338}

![Fig 14](image)

**Fig 14.** Geometry for TIR THz-TDS. T-rays are generated in the EO material and detected by a probe pulse after interacting with a reference or sample by frustrated TIR.\textsuperscript{124}

![Fig 15](image)

**Fig 15.** Transmission through a dielectric slab.

![Fig 16](image)

**Fig 16.** Notation for the reflection and transmission coefficients at a dielectric interface.\textsuperscript{179, 339}

### 4.2. T-ray time-domain spectroscopy

The first pulsed THz spectroscopy measurements were performed in reflection, with the THz confined by total internal reflection to an EO crystal that acted as both
emitter and detector. THz spectroscopy could only be performed on the EO crystal itself, or on materials placed in direct contact with the EO crystal, into which the radiated pulses were coupled by frustrated total internal reflection. Fig. 14 is a sketch of TIR THz time-domain spectroscopy (TDS). The basic elements of THz-TDS are present: a pump pulse to generate the THz radiation and a time-delayed probe pulse to sample the THz pulses, both with a sample and without.

Time-domain spectroscopy is a long-established method in the study of electronic circuits, but THz-TDS was its first application to the quasi-optical study of dielectrics. THz-TDS is markedly different from optical spectroscopic techniques that rely primarily on incoherent detectors. Any sample can be characterized by a complex dielectric constant \( \varepsilon(\omega) \), which describes the attenuation and delay of transmitted radiation at a given frequency \( \omega \). A coherent detector is able to determine the delay, or phase relationship, between incident radiation with and without a sample present, and thus directly measure both the real and imaginary parts of \( \varepsilon \). With an incoherent detector, however, additional processing is required to estimate the phase delay caused by a material, using the Kramers-Kronig relationships, integrals that relate the real and imaginary parts of the complex dielectric constant.

In THz-TDS, we are interested in measuring \( \varepsilon(\omega) = \varepsilon'(\omega) - j \cdot \varepsilon''(\omega) \), or equivalently \( \tilde{n}(\omega) = n(\omega) - j \cdot \kappa(\omega) \), which is the complex index of refraction of the material, where \( (\tilde{n})^2 = \varepsilon \). The dielectric constant is typically referred to in the field of high-speed electronics and the refractive index in THz optics. Estimating these material properties from THz-TDS measurements requires calculations that depend on the system configuration. Essentially, an expression is derived for the expected delay and attenuation of the pulse due to passing through or reflecting from the material of interest. This expression is in terms of the complex material properties and the thickness of the sample(s), and can be written as a product of factors in the frequency domain. The interaction between a sample and the spectral components \( S(\omega) \) of the T-ray radiation can be expressed as a transfer function \( \tilde{H}(\omega) \), which is a product of reflection \( r \), propagation \( p \) and transmission \( t \) coefficients, depending on the geometry of the system, as sketched in Fig. 15,

\[
\tilde{S}_{\text{final}}(\omega) = \tilde{H}(\omega) \cdot \tilde{S}_{\text{initial}}(\omega). \tag{1}
\]

The reflection and transmission at each material interface depend on \( \varepsilon \) (or equivalently \( \tilde{n} \)) of the adjacent dielectrics, and the polarization of the incident light, as shown in Fig. 16. For TE polarization, the complex frequency-dependent coefficients of transmission \( t(\omega) \) and reflection \( r(\omega) \) are

\[
r_{ab}^s(\omega) = \frac{\tilde{n}_a \cos \theta_i - \tilde{n}_b \cos \theta_R}{\tilde{n}_a \cos \theta_i + \tilde{n}_b \cos \theta_R}, \tag{2}
\]
\[
t_{ab}^s(\omega) = 1 + r_{ab}^s, \tag{3}
\]

and for TM polarization,

\[
r_{ab}^p(\omega) = \frac{\tilde{n}_b \cos \theta_i - \tilde{n}_a \cos \theta_R}{\tilde{n}_b \cos \theta_i + \tilde{n}_a \cos \theta_R}, \tag{4}
\]
\[
t_{ab}^p(\omega) = \frac{\tilde{n}_a}{\tilde{n}_b} (1 + r_{ab}^p), \tag{5}
\]
where $n_a$ and $n_b$ are the complex refractive indicies of the homogeneous materials before and after the interface, $\theta_i$ is the incident angle and $\theta_R$ is the refracted angle. With knowledge of the refractive indicies $n_a$ and $n_b$, $\theta_R$ can be determined from $\theta_i$ and Snell's Law, $n_a \sin \theta_i = n_b \sin \theta_R$. For a sample oriented normal to the radiation path, $\theta_i = 90^\circ$, the propagation equations simplify to

$$
t_{ab}(\omega) = \frac{2n_a}{(n_a + n_b)},
$$

$$
r_{ab}(\omega) = \frac{n_a - n_b}{n_a + n_b}.
$$

Radiation propagating a distance $d$ through a linear medium, $n_a$, is delayed and attenuated according to the factor

$$
p(\omega) = e^{-j2n_ao\omega d/c_0},
$$

where $c_0$ is the speed of light in a vacuum.

A final linear propagation effect occurs due to multiply-reflected radiation between two plane parallel interfaces, shown in Fig. 15. These are Fabry-Pérot reflections, or etalon effects, and are described by a sum of reflections,

$$
FP(\omega) = \sum_{k=0}^{\infty} \left\{ r_{23} \cdot p_2^2 \cdot r_{21} \right\}^k,
$$

where $k$ is the number of reflections, and depends on the time duration of the measured waveform and the delay caused by propagation between the interfaces. For a very large number of multi-reflections, $k \rightarrow \infty$, the Fabry-Pérot factor can be approximated by

$$
FP(\omega) = \frac{1}{1 - \left( \frac{n_2 - n_1}{n_2 + n_3} \right) \left( \frac{n_2 - n_1}{n_2 + n_3} \right) e^{-j2n_2\omega d/c_0}}.
$$

The Fabry-Pérot factor introduces frequency-domain interference fringes into transmitted or reflected radiation from thin samples, and is unavoidable in very thin samples.

Typically in THz-TDS experiments, it is simplest to measure two time-domain pulses, a reference $y_r$ and a sample pulse $y_s$. The characteristic response of the entire system, which depends on many factors including the emitter and detector, is canceled out by normalizing the sample pulse with the reference pulse in the frequency domain. This is deconvolution. The complex frequency spectra $\tilde{S}_r$ and $\tilde{S}_s$ are calculated using numerical Fourier transforms from the time-domain waveforms sampled in the experiment,

$$
\tilde{S}_r = FT(y_r),
$$

$$
\tilde{S}_s = FT(y_s).
$$

The effect of the sample on the propagating pulses can be modeled using Eqs. (6)-(10), which provide a theoretical expression relating $\tilde{S}_r$ and $\tilde{S}_s$ to the material properties $\varepsilon$. The material properties are then estimated by comparing the model to the
measured spectra. Depending on the complexity of the model expression, it may be possible to solve it analytically, otherwise an iterative curve-fitting procedure is used. The method described in this section is used for most THz-TDS measurements, with the model expressions for \( \tilde{S}_r \) and \( \tilde{S}_s \) varying with sample geometry. The expression of interest is the ratio of the complex sample spectrum to the complex reference spectrum, which enables the responses (transfer functions) of all invariant system components (for example, the emitter, detector, free-space propagation and mirror surfaces) to be canceled out. This ratio will be referred to as the deconvolved sample spectrum.

4.2.1. Transmission

\[
\tilde{S}_r = A(\omega) \cdot t_{12} \cdot p_2 \cdot t_{23}, \\
\tilde{S}_s = A(\omega) \cdot t_{1s} \cdot p_s \cdot t_{s3},
\]

The simplest and most common geometry for free-space THz-TDS is transmission through an orthogonally-positioned slab of homogeneous material, characterized by \( \tilde{n}_s \) (Fig. 15). This slab may be free-standing in air, or either deposited on or implanted into a substrate or holder. The model for the normalized sample spectrum depends on each individual experiment, but three main classes can be discerned: 1) thick samples, 2) thin samples and 3) dual-thickness samples.

A thick sample is a sample that causes sufficient delay so that the transmitted pulse can be measured without any overlap with the first Fabry-Perot (FP) reflection. The exact requirements for the delay will depend on the desired total scan length (see Sec. 4.2.8). A diagram of a thick sample and the reference and sample paths is shown in Fig. 17. Note that the substrate material both before and after the sample must also be sufficiently thick to avoid any FP reflections in the measured pulses. Thin substrates are discussed below with thin samples.

In the thick sample geometry, the experimentally-measured spectral components of the reference and sample pulses can be modeled by

\[
\tilde{S}_r(\omega) = A(\omega) \cdot t_{12} \cdot p_2 \cdot t_{23}, \\
\tilde{S}_s(\omega) = A(\omega) \cdot t_{1s} \cdot p_s \cdot t_{s3},
\]

Fig 17. Notation for transmission geometry. Fig 18. Notation for a dual-thickness geometry.
where $A(\omega)$ is the product of all other system responses that remain constant between the sample and reference measurements. Using Eqs. (6) to (8), the ratio of transmission spectra can be entirely determined in terms of refractive indices and the sample thickness,

$$\frac{\tilde{S}_s}{\tilde{S}_r} = \frac{\tilde{n}_s(\tilde{n}_1 + \tilde{n}_2)(\tilde{n}_2 + \tilde{n}_3)}{\tilde{n}_2(\tilde{n}_1 + \tilde{n}_s)(\tilde{n}_s + \tilde{n}_3)} \cdot e^{-j(\tilde{n}_s - \tilde{n}_2)d/\omega_0}. \quad (15)$$

This set of equations can be solved for the real and imaginary parts of $\tilde{n}_s$ using iterative techniques. For the common case where the sample $\tilde{n}_s$ is placed in a vacuum, $\tilde{n}_1 = \tilde{n}_2 = \tilde{n}_3 = 1.0$, a simplified expression can be determined for the transmitted pulse (and any time-separated FP pulses). For a sample with very low THz absorption, that is $\kappa_s \ll n_s$, analytic expressions for $n_s$ and $\kappa_s$ can be written in terms of the magnitude $\rho$ and phase $\phi$ of the deconvolved sample spectrum,

$$\frac{\tilde{S}_s(\omega)}{\tilde{S}_r(\omega)} = \rho(\omega) \cdot e^{-j\phi(\omega)}, \quad (16)$$

$$n_s(\omega) = \phi(\omega) \cdot \frac{c_0}{\omega d} + 1, \quad (17)$$

$$\kappa_s(\omega) = \ln \left( \frac{\rho(\omega) \cdot (n_s(\omega) + 1)^2}{\rho(\omega) \cdot (n_s(\omega) + 1)^2} \right) \cdot \frac{c_0}{\omega d}. \quad (18)$$

A thin sample is one where a number of Fabry-Pérot reflections overlap in the measured pulse train. Theoretically, this geometry can be modeled by multiplying Eq. (15) by an FP factor from Eq. (9) or (10), and then solving iteratively. In practice, however, this method amplifies errors from the measurement of $d$ and from alignment; it is difficult, for example, to maintain collinearity between the multi-reflections in sample positioning.

Thin substrates introduce a further level of complexity, as FP reflections from the substrate are overlaid on the detected pulse. The FP reflections depend on $n_s$ and so cannot be deconvolved without approximations. An expression in the form of Eq. (15) can be written and solved iteratively, although not aligning all the FP reflections collinearly can introduce errors.

A dual-thickness measurement greatly simplifies modeling, and involves the reference pulse passing through the sample material, where the thickness is different from the sample pulse measurement, as shown in Fig. 18. The advantage of dual-thickness measurements is that both reference and sample pulses pass through the same interfaces, so for a thick sample, the normalized transmission is

$$\frac{\tilde{S}_s}{\tilde{S}_r} = \frac{p_s}{p_r}. \quad (19)$$

This results in two simple expressions for $n_s$ and $\kappa_s$ in terms of the difference between the two thicknesses. For the case where the sample thickness $d_s$ is greater than the reference thickness $d_r$,

$$n_s(\omega) = \phi(\omega) \cdot \frac{c_0}{\omega(d_s - d_r)} + 1, \quad (20)$$

$$\kappa_s(\omega) = \ln \left( \frac{1}{\rho(\omega)} \right) \cdot \frac{c_0}{\omega(d_s - d_r)}. \quad (21)$$
For thin dual-thickness samples, where multiple FP reflections are present, the expression for the normalized transmission is more complex and must be solved iteratively. The accuracy of these estimates can be calculated from the partial derivatives of Eqs. 20 and 21, using multiple measurements to obtain the experimental fluctuation of the normalized transmission.\textsuperscript{342}

One of the major sources of error in material constant estimation is the measurement of the thickness $d$.\textsuperscript{340} To increase the accuracy of measuring $d$ and integrate it with T-ray spectroscopy, a number of schemes have been proposed to use a train of time-separated FP pulses from a sufficiently thick sample to simultaneously estimate $\tilde{n}_s$ and $d$.\textsuperscript{340,343}

4.2.2. Reflection

![Fig 19. TIR spectroscopy using a first surface reflection as the reference pulse (Sec. 4.2.2).](image1)

![Fig 20. The dithered sample used in DTDS (Sec. 4.2.5).](image2)

Reflection spectroscopy is used for large or highly-absorbing samples. As with transmission spectroscopy, the structure of the sample plays a role in the analysis and accuracy of the parameter estimations.

The geometry of first surface THz spectroscopy is the same as the TIR geometry in Fig. 14, except the initial medium is typically air or vacuum. The measured spectral components of the reference and sample pulses are given by

\begin{align}
\tilde{S}_r(\omega) &= A(\omega) \cdot r_{12}, \\
\tilde{S}_s(\omega) &= A(\omega) \cdot r_{1s},
\end{align}

where $r$, the reflection coefficient, is found from the Fresnel equations above, and $A(\omega)$ represents the common factors between reference and sample. The reference interface is created with a material of known $\tilde{n}$. It is critical that the reference and sample path lengths are identical, which means the sample interface must be placed at the same position as the reference interface. This geometry allows the deconvolved spectrum $\tilde{S}_s/\tilde{S}_r$, which is calculated from the measured data, to be modeled in terms of the reflectivity of the reference $r_{12}$,

\begin{equation}
\frac{\tilde{S}_s(\omega)}{\tilde{S}_r(\omega)} = \frac{r_{1s}(\omega)}{r_{12}(\omega)}. \tag{24}
\end{equation}
...can be calculated from Eq. (4), or (2), depending on the T-ray polarization, since \( n_1, n_2 \) and the incident angle \( \theta_i \) are known. Once \( r_{12} \) is known, Eq. (24) can be rearranged to estimate \( r_{1s} \). The dielectric constant of the sample, \( \tilde{\epsilon}_s = n_s^2 \), is estimated by re-arranging Eq. (4) (or (2)),\(^\text{124}\)

\[
\tilde{\epsilon}_s(\omega) = \left[ \sin^2 \theta_i + \cos^2 \theta_i \left( \frac{1 - r(\omega)}{1 + r(\omega)} \right)^2 \right] \tilde{\epsilon}_1(\omega). \tag{25}
\]

For reflection spectroscopy at or near normal incidence, which includes T-ray transceiver applications (Sec. 3.5.3), the above expressions simplify by setting \( \theta_i \approx 90^\circ \).

For measurements using total internal reflection (TIR), the position of the interface is set by the edge of the TIR prism. The difficulty is knowing accurately the refractive indices of the prism and the reference material.\(^\text{124}\) One method to avoid making a separate reference measurement is to use the pulse reflected from the top of a TIR prism, as shown in Fig. 19.\(^\text{344}\) The reference pulse is detected from the air-silicon window interface, and the sample pulse comes from the silicon window-sample interface. The spectral reflection ratio of sample to reference can be written

\[
\frac{S_s(\omega)}{S_r(\omega)} = \rho e^{-j\phi} = \frac{l_{12} \cdot r_{2s} \cdot t_{21}}{r_{12}} \cdot e^{-j2n_2d_2\omega/c_0}, \tag{26}
\]

for a deconvolved sample spectrum \( \rho \cdot e^{-j\phi} \). The equations can then be solved iteratively.

If \( n_1 \gg \kappa_1 \) and \( n_2 \gg \kappa_2 \), for a Si TIR prism in air, the sample properties can be estimated by

\[
n_s(\omega) = \frac{n_2(1 - \rho^2)}{1 + \rho^2 + 2\rho \cos \phi}, \tag{28}
\]

\[
\kappa_s(\omega) = \frac{2n_2\rho \sin \phi}{1 + \rho^2 + 2\rho \cos \phi}. \tag{29}
\]

For samples with internal surface reflections, for example multiple quantum wells, the expression for reflection is complicated by FP factors.\(^\text{324}\)

An integrated method for determining the thickness \( d \) of a thin surface layer, analyzed in reflection, is to observe frequencies where destructive interference occurs. This technique relies on the broadband nature of the T-ray pulse, requiring a bandwidth of 12 THz to measure films of 1-\(\mu\)m thickness.\(^\text{345}\)

4.2.3. Goniometric

Measuring reflectance as a function of incident angle, specifically near Brewster's angle, involves a reflection geometry where the angle of incidence is variable.\(^\text{8,346}\) These measurements require rotation of the sample by an angle \( \theta \) and the detector by \( 2\theta \) in a goniometer. With a series of measurements, the amplitude and phase
shift of the T-ray pulse can be observed around Brewster’s angle, specifically the 180° phase shift for p-polarized radiation.

The refractive index of the thin film, \( n_s \), can be estimated by measuring the reflectance at different angles and comparing it to a model based on the Drude and Fresnel equations. The Drude equation predicts the complex reflectance \( \tilde{r}_D \) from a thin film,

\[
\tilde{r}_D = \frac{r_{1s} + r_{2}p_{s}}{1 + r_{1s}r_{2}p_{s}},
\]

where \( r \) and \( p \) are given by Eqs. (7) and (8). The angles \( \theta_s \) and \( \theta_2 \) can be expressed in terms of \( \theta_1, \tilde{n}_1, \tilde{n}_s \) and \( \tilde{n}_2 \) with Snell’s law. The values of \( d, \tilde{n}_2, \tilde{n}_1, \lambda \) and \( \theta_1 \) are known for each measurement, so the reflectance can be modeled using estimates for \( \tilde{n}_s \), and compared to experimental measurements. For two angles of \( \theta_1, \theta_1 A \) and \( \theta_1 B \), the T-ray spectra are \( \tilde{S}_A(\omega) = A(\omega) \tilde{r}_{DA} \) and \( \tilde{S}_B(\omega) = A(\omega) \tilde{r}_{DB} \), where \( A(\omega) \) is a product of the transfer functions of the system which remain constant as the angle is varied. Thus the ratio of reflectances

\[
\frac{\tilde{r}_{DB}}{\tilde{r}_{DA}} = \frac{\tilde{S}_B(\omega)}{\tilde{S}_A(\omega)}.
\]

The value of \( \tilde{n}_s \) that fits the model can be estimated by iteratively fitting the experimental reflectance ratio to the modeled reflectance ratio, using Eq. (30). Specifically, when measurements are made close to Brewster’s angle, the large variation in phase enables more reliable estimates of \( \tilde{n}_s \) than normal reflection or transmission measurements.

4.2.4. Ellipsometry

THz ellipsometry is a technique to estimate a sample’s dielectric constant by measuring two reflected pulses with s and p polarized T-rays. The advantages of this technique over simple reflection spectroscopy is that there is no need to position a reference mirror. The equipment is simpler than for goniometric measurements.

4.2.5. Differential

THz thin film characterization requires very high sensitivity. Even with an SNR of 10^7, it is difficult to see phase shifts that are less than the coherence length of the radiation. Picosecond T-ray pulses with a bandwidth of approximately 2 THz and a center frequency of 0.8 THz have a coherence length of approximately 150 \( \mu m \). The phase shift caused by a dielectric sample in the T-ray radiation path is proportional to \((\tilde{n} - 1)d/\lambda\), where \( \tilde{n} \) is the complex refractive index of the medium, \( d \) is the sample thickness and \( \lambda \) is the T-ray wavelength. For small \( \tilde{n} \) and \( d \), this phase change is very difficult to detect in background noise. A differential technique may be used to reduce background noise.

Differential THz-TDS (DTDS) involves modulating the T-ray signal using only the thin film, and detecting the magnitude of this modulation using a lock-in amplifier (LIA). The sample is a substrate half-covered with the film and half bare, as shown in Fig. 20. The thin film, characterized by the complex refractive index \( \tilde{n}_s \), is supported by a substrate, \( \tilde{n}_1 \).
For a typical T-ray spectrometer, the major source of noise is the pump laser, which is very sensitive to slow fluctuations in temperature. These fluctuations can cause larger changes in the detected T-ray signal than the thin film sample itself. The advantage of differential spectroscopy is that the signal transmitted through the film is compared to the signal through the substrate at each point of the delay stage, thus normalizing the laser-based fluctuations. A more complete discussion of the noise present in DTDS and how it can be further reduced by double modulation is found in Mickan et al.\textsuperscript{353–355} The differential waveform is equivalent to the difference between the reference and sample waveforms, $y_d = y_r - y_s$. The material parameters can be simply estimated by measuring $y_r$ and $y_d$, calculating $y_s$, and using the transmission equations above for thin samples. For a thin film with thickness $d$ and $n_s \gg \kappa_s$, deposited (so $n_2 = n_1$) on a substrate $n_3$, where $n_1 \gg \kappa_1$, $n_2 \gg \kappa_2$ and $n_3 \gg \kappa_3$, an analytic expression for the refractive index can be derived:\textsuperscript{351}

$$n_s(\omega) = \sqrt{\frac{c_0}{\omega d} \left| \frac{S_d}{S_r} \right| (n_1 + n_3) - n_1 n_3},$$

(32)

where $S_d = \mathcal{F}T(y_d)$. 

---

Fig 21. Double-modulated DTDS schematic. The pump and probe beams are split from a femtosecond laser and act to generate and detect the T-rays. The T-rays, shaded in grey, are collimated and focused with gold-plated parabolic mirrors. The T-ray electric field is converted to an electronic signal with the EO crystal, a quarter wave plate ($\lambda/4$) and balanced photodiodes. Two lock-in amplifiers (LIAs), with the shaker and an optical modulator, implement the double modulation scheme. The DTDS output signal depends on the differences between the films on opposite ends of the slide.\textsuperscript{348,349}
4.2.6. Interferometry

Another method to increase the sensitivity of T-ray spectrometers to thin films is interferometry. T-ray interferometric techniques typically induce a 180° phase shift between two arms of an interferometer, then detect phase changes introduced by a thin sample into one arm. Enhanced depth and spatial resolution have been achieved with T-rays focused to a point on a reflective sample in one arm of an interferometer, where the Gouy shift occurs at the focal point on the sample. The peak amplitude showed a 20% change for a 12.5-μm-thick air gap in Teflon. Other interferometers have been constructed using silicon prisms as reflectors, where the 180° phase shift was induced by a fixed end reflection from one of the faces, and using counter-propagating OR to generate T-ray pulses with an opposing sign.

The signal at the detector in a T-ray interferometer is the sum of the pulses in the two arms, \( y = y_1 + y_2 \). For a system with identical pulses in both arms, differing only by a (small) phase difference \( \phi(\omega) \),

\[
\begin{align*}
\tilde{S}_1(\omega) &= A(\omega), \\
\tilde{S}_r(\omega) &= A(\omega)(1 - e^{j\phi(\omega)}), \\
\tilde{S}_s(\omega) &= A(\omega)(1 - t_s p_s e^{j\phi}),
\end{align*}
\]

where \( \tilde{S}_1(\omega) \) is the spectral response of arm 1 of the interferometer when arm 2 is blocked, and \( t \) and \( p \) are the transmission and propagation coefficients for the sample placed in arm 2 of the interferometer. The phase difference \( \phi(\omega) \) can be calculated by normalizing the reference spectrum to the spectrum in arm 1 alone

\[
\frac{\tilde{S}_r(\omega)}{\tilde{S}_1(\omega)} = 1 - e^{j\phi(\omega)}. \tag{36}
\]

Typically, interferometric precision is required for samples with very low refractive index, so the transmission coefficients can be approximated \( t \approx 1 \). For very thin samples, where \( k_0(n_s - 1)d \ll 1 \), the sample spectrum can be normalized to the spectrum measured by arm 1 alone, giving an approximate expression for the sample’s properties

\[
\frac{\tilde{S}_s(\omega)}{\tilde{S}_1(\omega)} = \rho e^{-j\phi} \approx k_0 \kappa(\omega)d + jk_0(n(\omega) - 1)d + j\phi(\omega). \tag{38}
\]

4.2.7. Waveguide resonators

An important new demonstration of T-ray spectroscopy is in micro-stripline resonators. In this geometry, T-rays propagate along a micro-stripline rather than through free space, returning to the original experiments with THz radiation in circuits. The presence of a different dielectric in the resonator cavity causes a frequency and amplitude shift in the transmitted THz.

Currently, this geometry has only demonstrated qualitative estimates of the sample dielectric, although it is potentially a method for measuring \( n_s \) of extremely small sample sizes (\( \approx 250 \times 50 \) μm²).
4.2.8. **Numerical Fourier spectra**

The FFT algorithm is used for all analysis in time-domain spectroscopy. A number of points need to be made regarding the relationship between the sampled time-domain waveform data and the FFT spectra. These issues are typical of systems using numerical Fourier transforms, including well-established Fourier Transform spectroscopy.\(^{360}\)

Of primary concern in spectroscopy are the frequency range and resolution of measurements. From basic Fourier considerations, the time between data points, \(\Delta t\), determines the maximum frequency observable in a spectrum \(f_{\text{max}} = 1/(2\Delta t)\), and the total duration of the data \(T\) determines the frequency resolution, and thus minimum frequency, \(\Delta f = 1/T\). However, when a time-domain signal is represented as discrete data points, high-frequency signals above \(f_{\text{max}}\) can be aliased to frequencies below \(f_{\text{max}}\), thus it is important to ensure that there are no signals present above the \(f_{\text{max}}\) set by \(\Delta t\). For T-rays, this upper signal cut-off is typically determined by the emission or detection bandwidth of between 5 and 10 THz.

The sample duration \(T\) is often limited in T-ray spectroscopy by physical considerations or by FP reflections. Spectral resolution can be improved by artificially adding zeros to the time-domain data. This provides smoother spectra, but if the time duration is more than doubled, no new information is actually present. The extra points in the spectra are just sinusoidal interpolations. FP reflections can arise in the sample, the emitter and the detector. When an EO crystal is used, FP reflections can be removed by bonding a refractive index-matched material to the crystal itself, thereby increasing its thickness. For example, the (110) ZnTe emitter or sensor can be bonded to a (100) ZnTe crystal, which has a null transverse EO coefficient.\(^{193}\)

One of the advantages of T-ray spectroscopy over other THz optical techniques is the acquisition of phase information. Unfortunately, the phase spectrum from the FFT output needs to be unwrapped before it is useful. Common unwrapping algorithms can distort the unwrapped phase due to noise at low frequencies. One method to overcome this is to take phase information only at frequencies where the SNR is high, then artificially extrapolate back to DC.\(^{6}\)

4.3. **Materials studied with T-rays**

Many materials, molecules and mixtures have been studied with T-rays, as outlined in the sections below. T-ray experiments have provided information on the THz response of samples, enhancing theoretical models and previous studies, including CW methods reviewed in Sec. 2.

4.3.1. **Gases and vapors**

T-rays were first used for accurate spectroscopy of water vapor in the air, measuring the frequencies and strengths of absorption due to rotational water molecule transitions.\(^{5}\) Gas and vapor transmission (absorption) experiments typically use PCA emitters and detectors with an enclosed metal gas sample cell with thick (> 10-mm) high resistivity (\(\approx 10\cdot k\Omega/cm\)) silicon windows, which are designed to minimize T-ray attenuation and avoid multiple Fabry-Pérot reflections inside the windows. The length of the sample cell is chosen depending on the strength of the absorption
for the pressures of interest. The entire T-ray path must be dried or evacuated to remove absorption from residue polar molecules, such as water, in the beam path.

T-ray gas spectrometers can accurately measure absorption lines and collision broadening for molecules that have a permanent dipole moment, as demonstrated with H$_2$O, SO$_2$, methyl halides, ammonia and CH. T-ray measurements allow absorption, dispersion and line shape data to support theoretical models of these molecules. Although demonstrating low average power, the ps time scale of the T-ray pulses results in very high peak powers, therefore T-rays can be used for spectroscopy of samples with large average background THz radiation, such as flames. Gases with no permanent dipole moments, such as N$_2$, O$_2$ and CO$_2$, show no T-ray absorption, whereas the concentrations of H$_2$O, CH and NH$_3$ in flames can be estimated from the T-ray spectra. T-ray spectra of flames can be used to estimate flame temperature and study the higher rotational absorption states of hot water molecules.

Detecting gases in a mixture is an important task for gas spectrometers. Gas identification, and classification using a linear predictive coding algorithm, has been demonstrated from the T-ray power spectra of NH$_3$ and H$_2$O. An accurate gas filter correlation (GFC) system for detecting specific gas species has been demonstrated for H$_2$S. The GFC system uses a calibrated sample cell in one arm of an interferometer and the unidentified gas mixture in the other, and was implemented using a pulsed photomixer emitter. A detection sensitivity of 30 ppm was achieved using GFC.

An important characteristic of polar gases and vapors at room temperatures and atmospheric pressures is line broadening. The broadened absorption lines from most gases severely overlap, making identification difficult. To avoid line broadening, the sample cell (or the calibration cell in GFC) must be held at a reduced pressure. A review of collision-broadened rotational lines in gases studied with T-rays was prepared by Harde, Cheville and Grischkowsky.

Vapors have been shown to re-emit THz pulses after excitation with a primary THz pulse, for example, N$_2$O and methyl chloride. These experiments have been used to characterize rotational and vibrational constants, and to study line shape broadening, which shows excellent agreement with linear dispersion theory.

4.3.2. **Liquids**

Liquid studies at THz frequencies are concerned with characteristic relaxation times of permanent or induced molecular dipole moments. Keiding *et al.* have studied THz spectra of liquid water as a function of temperature, modeled using a Debye relaxation model and numerical simulation. These experiments are performed using a reflective TIR method, described in Sec. 4.2.2 above. The temperature dependence of liquid water has also been compared to D$_2$O. An alternative model of the THz response of liquid water, describing molecular plasma oscillations in an ice-like crystalline lattice, has been proposed from transmission data.

Liquid water shows a very high THz absorption, greater than 200 cm$^{-1}$ at 1 THz, whereas non-polar liquids have coefficients around 100x smaller (for example, benzene, carbon tetra-chloride and cyclohexane). Keiding *et al.* have studied the temperature dependence of the solvents benzene and toluene, observing rotational
and librational bands. Schmuttenmaer et al. have studied numerous liquids of varying polarity with a dual-thickness sample cell, based on a polyethylene bag held between two movable silicon or polyethylene windows, including water, methanol, 1-propanol and liquid ammonia. All results have been matched to Debye relaxation dielectric models with good accuracy. A number of optical-pump and T-ray-probe studies have explored the response of solvents and dyes to photoexcitation, and linked these results to finite difference time-domain models. The solvation dynamics of polar and non-polar liquids, including acetonitrile and water, acetone, acetonitrile and methanol and water, have been related to molecular dynamics simulations to explore the relationship of decreasing THz absorption with increasing liquid structure. The solvation dynamics of lithium salts in water, methanol and propylene carbonate have been explained with Debye relaxation models.

Mittleman et al. have measured and modeled the THz response of inverse surfactant micelles of water in heptane compared to bulk water. The reduced dielectric constant of the micelles is attributed to confinement effects on the water molecules.

4.3.3. Solids

The earliest studies using pulsed THz spectroscopy were on LiTaO₃, generating and detecting the THz radiation inside the same crystal. The importance of EO materials in OR and EOS has led to further studies, characterizing the complex dielectric of ZnTe, GaAs, LiTaO₃ and organic crystals, and the temperature-dependent power absorption spectrum of two-phonon processes in ZnTe and CdTe.

Table 3. Dielectric constants of select solids at 1 THz, as measured using THz-TDS. \( n_o \) refers to the ordinary refractive index and \( n_e \) refers to the extraordinary refractive index for birefringent materials. High-resistivity (10-kΩ-cm) silicon demonstrates the lowest dispersion, with \( n \) almost constant across the spectrum from 0.2 to 2 THz. The accuracy of the first 6 measurements is better than 0.0004.

<table>
<thead>
<tr>
<th>Solid</th>
<th>Refractive index</th>
<th>Power absorption</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>sapphire</td>
<td>( n_o = 3.070 ), ( n_e = 3.415 )</td>
<td>( \alpha \approx 1 \text{ cm}^{-1} )</td>
<td>133</td>
</tr>
<tr>
<td>crystalline quartz</td>
<td>( n_o = 2.108 ), ( n_e = 2.156 )</td>
<td>( \alpha = 0.1 \text{ cm}^{-1} )</td>
<td></td>
</tr>
<tr>
<td>fused silica</td>
<td>( n = 1.952 )</td>
<td>( \alpha = 1.5 \text{ cm}^{-1} )</td>
<td></td>
</tr>
<tr>
<td>intrinsic Ge</td>
<td>( n = 4.002 )</td>
<td>( \alpha = 0.5 \text{ cm}^{-1} )</td>
<td></td>
</tr>
<tr>
<td>high-( R ) GaAs</td>
<td>( n = 3.595 )</td>
<td>( \alpha = 0.5 \text{ cm}^{-1} )</td>
<td></td>
</tr>
<tr>
<td>high-( R ) Si</td>
<td>( n = 3.418 )</td>
<td>( \alpha &lt; 0.05 \text{ cm}^{-1} )</td>
<td></td>
</tr>
<tr>
<td>ice</td>
<td>( n = 1.793 )</td>
<td>( \alpha = 8.6 \text{ cm}^{-1} )</td>
<td>387</td>
</tr>
<tr>
<td>0.19-Ω-cm N-GaAs</td>
<td>( n \approx 2.97 )</td>
<td>( \alpha = 320 \text{ cm}^{-1} )</td>
<td>388</td>
</tr>
<tr>
<td>0.36-Ω-cm P-GaAs</td>
<td>( n \approx 3.44 )</td>
<td>( \alpha = 270 \text{ cm}^{-1} )</td>
<td></td>
</tr>
</tbody>
</table>
Early free-space T-ray systems were used for spectroscopy of common homogeneous solids. Arjavalingham et al. measured the complex dielectric constants, and their polarization and angular dependence, of fused silica, sapphire and plexiglass slabs up to 130 GHz with ps optical pump pulses.\textsuperscript{132,341,389-391} Wire grid polarizers were used to control the THz polarization. Grischkowsky et al. extended these measurements to 2 THz, studying common solids, as summarized in Table 3, and modeled the conductivity of doped Si with an extension of the Drude dielectric model to include energy-dependent carrier relaxations.\textsuperscript{392}

T-ray spectroscopy is an attractive non-contact, non-destructive and rapid technique that has been applied to semiconductor wafer characterization in a variety of geometries. Transmission measurements, detailed above, are limited to samples with low absorption. Doped n-GaAs, doped n-Si and bulk GaAs wafers have been characterized in reflection, with an Al mirror providing the reference pulse, and modeled using Drude theory to determine carrier density and mobility.\textsuperscript{384,393} The dielectric properties of 600 to 15-\mum films of highly doped semiconductors have been measured in reflection, where the thickness is measured by observing frequency of destructive interference in the THz spectrum.\textsuperscript{345} THz ellipsometry has been demonstrated on doped Si wafers to estimate \( \varepsilon \),\textsuperscript{347} and the Hall effect has been used to estimate the full conductivity tensor of n-GaAs, implemented with a 1.3-T magnetic field and dual detectors for the two emitted polarizations.\textsuperscript{298,383}

The optical pulse driven nature of T-rays lends itself to the study of photoexcited carrier dynamics in semiconductors; optical-pump and THz-probe experiments have the advantage of ultrafast resolution coupled with THz bandwidth.\textsuperscript{394} The THz reflectivity of photoexcited GaAs can be measured by T-ray reflection.\textsuperscript{395} The time-dependent conductivity of GaAs-AlGaAs quantum wells,\textsuperscript{396} bulk GaAs and epitaxial LT-GaAs have been studied under 800-nm and 400-nm light, and modeled using a modified Drude model.\textsuperscript{397,398} Two-color EOS, with time-delayed collinear pulses transmitted through a GaAs sample, has been used to study T-ray generation from surface field dynamics.\textsuperscript{245} Using ultrafast optical excitation pulses, T-rays have been used to observe photocarrier generation and the subsequent screening processes in semiconductors.\textsuperscript{282,399,400} Optical-pump and THz-probe experiments have also enabled a form of near-field imaging with a dynamic aperture, as detailed in Sec. 5, and the creation of transient mirrors for T-ray pulse slicing.\textsuperscript{319}

Characterization of insulating and conducting materials in the GHz-THz range is invaluable for the future development of high speed circuits. As potential substrates for superconducting circuits, MgO has been measured with a low THz transmission loss, unlike YSZ (yttrium stabilized zirconia) and LaAlO\textsubscript{3}.\textsuperscript{401} T-rays can be used to observe static and dynamic characteristics of superconducting films themselves.\textsuperscript{402,403} Goniometric measurements have been made of FLARE, TiO\textsubscript{2} and PZT thin films\textsuperscript{346} and parylene-N films have been characterized with differential spectroscopy.\textsuperscript{404} Non-contact characterization, and modeling with localization-modified Drude theory, is particularly valuable for conducting polymers, such as polypyrrole and poly-3-methylthiophene,\textsuperscript{405,406} and single-walled carbon nanotube films.\textsuperscript{407} Organic thin polymer films were first characterized by T-rays in 1992.\textsuperscript{408} Microwave ceramics for telecommunications have been analyzed in transmission\textsuperscript{409} and corrosion layers beneath opaque paints analyzed in reflection.\textsuperscript{410}
4.3.4. Biological materials

With the development of T-ray spectroscopy, increasingly complex materials are being characterized, specifically biological and medical samples. Many of these materials are particularly sensitive to sample preparation and environmental conditions.

THz-TDS has been previously used to monitor broadband THz optical properties of biomolecules in binding, denaturation, temperature and humidity studies. Biomolecules, such as proteins and nucleic acids, have broad THz features arising from a multitude of dense rotational, vibrational, inter-domain and hydrogen bond energy transitions. Molecular dynamics simulations of simple biomolecules can be used to understand the measured THz spectra.

Using planar waveguides and a THz resonator, Bolívar et al. have demonstrated a highly sensitive device for probing the binding state of DNA. This thin film micro-stripline approach could be extended to two-dimensional gene chips for high speed DNA analysis. The THz waveguide and resonator structure is sensitive to single-base defects in an 1.1 femtomol volume of 0.52 g/L DNA-in-water.

Biomolecular films are extremely sensitive to conditions of humidity, temperature, pH and preparation. At room temperature, the THz spectra of protein structures are heavily populated, resulting in a continuous spectral response with few signature resonances. The THz spectrum is difficult to measure because dried biomolecular films have very small THz responses. These problems can be overcome by careful sample preparation and the highly sensitive technique of DTDS, detailed in Sec. 4.2.5. Another solution to environmental control of biomolecules is to study them as suspensions in a non-polar organic solvent, where activity and hydration studies can be carried out with high accuracy.

An important consideration in inhomogeneous structures is scattering. Scattering of single-cycle T-rays has been studied with samples constructed from teflon balls, with expressions derived for the mean free THz path length in the model medium.

Larger, more complex structures have been studied phenomenologically, with an emphasis on contrast rather than full spectroscopic information. This is typically the case for sensor applications, or in imaging of tissues (see Sec. 5). One example is the image of a desiccating leaf, which has been studied spectroscopically with CW techniques in the 100 to 500-GHz range. THz biosensing, using contrast derived from a combination of sample properties, including thickness, absorption, scattering and phase delay, has been used to detect the binding of lipids to proteins with a sensitivity of 1 ng/mm². The push to clinical applications of T-rays (see Sec. 5) has led to studies of potential damage caused by THz. Two reviews of T-ray studies in biomedicine have been written by Chamberlain et al.

4.4. Radar and ranging sensing

T-rays can be used for purposes other than spectroscopy. As free-space pulses of radiation, T-rays have been used for scaled-down versions of radar applications, such as distance and thickness measurements. T-rays can be used to study the radar profiles of scaled-down objects, such as 1/200th-sized model planes and tanks, or to characterize micromachined components. Ranging measurements occur in...
Fig 22. The number of pages in a book can be counted by measuring the phase delay of T-ray pulses in transmission. The resolution shown here is better than 1 in 400.

Fig 23. In a similar application to Fig. 22, the time delay of T-ray pulses in transmission have been used to count stacks of paper currency. Fig. 23(a) shows an expanded part of the time-domain waveform. Fig. 23(b) shows both the time delay and pulse spreading as T-ray pulses are transmitted through thicker stacks of paper currency.

reflection, either using a small incidence angle or a transceiver.\textsuperscript{197,323}

The interpretation of reflected pulses necessitates the modeling of single-cycle electro-magnetic pulse interactions with dielectric objects.\textsuperscript{430,431} Measuring T-ray pulse delay through a sample enables the THz refractive index profile of a flame to be estimated\textsuperscript{298} and the number of pages in a book to be counted, as demonstrated in Fig. 22. This work has been applied to studying the THz time delay caused by dollar bills, whereby cash can be counted by its thickness, seen in Fig. 23.

T-ray scattering from objects has been extended to imaging research, including quasi-optical imaging,\textsuperscript{432} synthetic aperture imaging\textsuperscript{433} and T-ray propagation around a cylinder.\textsuperscript{434,435} These topics are further discussed in Sec. 5.
4.5. *Terahertz-induced activity*

T-rays, despite their low average power, can be used for coherent manipulation of states in atoms, for spectroscopy and for quantum information processing. Rydberg wave packets have been created, probed and ionized using ultrashort THz pulses.\(^436-439\) T-ray pulses have similarly been used to excite Rabi oscillations in donor impurities in GaAs.\(^319\) Theoretical modeling suggests that narrowband THz could be used for tuning the resonance of semiconductor microcavities, and pulsed THz may be used to control coherent mode oscillations.\(^440\)

5. Imaging

![Fig 24. T-ray transmission through ink on currency. This figure indicates how a T-ray image can detect contrast in dry, non-polar samples. This image was taken with a raster-scanned imaging system. Most interestingly, the T-ray image remains unchanged when the plastic Australian $50 note is placed inside a paper envelope.\(^441\)](image)

5.1. *Introduction*

T-rays have been extended to two-dimensional imaging using a variety of scanning and CCD techniques, which provide parallel processing for speed and image processing for visualization. The first T-ray imaging systems involved raster scanning a sample to build up a 2D image, as discussed in Sec. 5.2. Using EO detection, this has been extended capturing the THz image of the whole sample with a CCD. Using a 2D array of data speeds up the acquisition of information at the expense of losing signal power at each pixel. Simple contrast images in the THz domain can be developed using absorption information, delay information or a combination of the two. A T-ray image is shown in Fig. 24 of two digits on a plastic $50 note, showing the different THz transmission of the inks. Dry substances, such as paper and plastic, transmit T-rays with less than 1% attenuation. More interesting than a simple contrast image is a two-dimensional array of waveforms. Spectral analysis at each pixel can then provide molecular information about the whole sample. With a metric of FIR molecular classification, it would be possible to map samples in terms of their specific molecular composition, potentially valuable in chemistry, biology and medicine.
5.2. Scanning and synthetic aperture imaging

The first T-ray images were obtained by simply scanning a sample across the T-ray beam, and presenting neighboring data points as pixels. The images were constructed of transmitted T-ray waveforms, with a number of contrast mechanisms available: total transmitted amplitude, peak delay at each pixel, and transmitted amplitude or delay of a given Fourier component. The technical advances that led to T-ray imaging were primarily methods of increasing the T-ray signal strength, coupled with faster acquisition speeds. The first demonstration of T-ray imaging had a 10–20 pixel-per-second waveform acquisition rate with an SNR of 100:1. This system relied on a 50-μm-gap SI-GaAs PCA emitter, a 5-μm-gap PCA detector, a 7.5-mm, 20-Hz scanning delay line for the gated detector, and a dedicated DSP chip to sample, process and display the data.420

T-ray imaging based on detector scanning has been simplified by the availability of fiber-coupled PCA detectors, where the probe time delay and alignment remain fixed by the fiber.9 A scanned PCA detector has been used to characterize the emission patterns from PCAs,167-169 and image the spatial reshaping of T-rays as they tunnel through a narrow air gap.442

Reflection imaging requires reconstruction algorithms to estimate the shape and structure of the reflecting object. The reflected data can be acquired along a straight line, perpendicular to the emitter-sample normal, and the object shape can be reconstructed using Kirchhoff migration,443 or for data acquired over a hemisphere, the object profile can be calculated by numerical back-propagation.444 Enhanced depth resolution is achieved by scanning the sample in an interferometer, as described in Sec. 4.2.6.356,445

Synthetic aperture processing is used to reconstruct 3D T-ray images and images with enhanced resolution.446

5.3. PCA array imaging

It is preferable in an imaging system to have a 2D array of detectors, so the sample and detector remain stationary. Imaging array detectors are available for CW terahertz detection, from applications in mm and sub-mm astronomy.447 Arrays of gated PCAs have been proposed but not yet developed.441,448

5.4. EOS CCD imaging

An effective method of 2D T-ray detection is to use an expanded probe beam in EOS, and detect the probe beam’s polarization rotation with a polarizer and a CCD, as shown in Fig. 25. The advantage of using EOS is the increased bandwidth over PCAs and the simplicity of array imaging. The central trade-off remains between speed and SNR; for 2D imaging, the T-ray beam is expanded to encompass the entire object, which spreads out the available T-ray power over the number of pixels. The main advantage is that neither sample nor detector need be moved to acquire the entire image.

An advance of EOS imaging is the ability to use chirped optical probe pulses, as described above in Sec. 3.4.2. The advantage of chirped probe measurements is that the entire T-ray waveform can be sampled with only one pulse, as schematically described in Fig. 10. Full waveforms can be acquired from 1D images with a CCD.291
Fig 25. 2D EOS differs from point EOS in that both the probe beam and the T-ray beam are expanded onto the EO detector, and the transverse intensity modulation of the probe is detected by a CCD camera. The data is loaded directly to a computer for presenting a visual image. 3D data sets can be built up and analyzed in XY slices, or with time on one axis (YT, XT).

Fig 26. Examples of single-shot 1D imaging of dipole and quadrupole T-ray fields. In this data, the emitted field of dipole and quadrupole PCA emitters have been imaged using EOS. The time evolution of the T-ray pulse is shown simultaneously across the X-dimension of the emitter.

Two example waveforms are shown in Fig. 26, where a cross-sectional array of time-domain waveforms have been imaged simultaneously. This technique is close to the theoretical speed limit of T-ray imaging, although a reference measurement needs to be made to interpret the chirped waveform.

5.5. Tomography

When detected in reflection, the pulsed nature of T-rays enables analysis of a sample’s internal structure, using tomographic reconstruction. Tomography is an extension of T-ray ranging studies to more complex internal structures and 1D or 2D imaging. Tomography enables the visualization of internal structure as different...
Fig 27. Schematic layout of a T-ray CT experiment. Data is acquired with a single-shot imaging system using chirped probe pulses. The sample, in this case a hollow plastic ball, is scanned X-Y and rotated about its axis. It takes approximately 1 hour to scan 100x100 images at 18 projection angles, which is sufficient to reconstruct a 3D sectionable profile of the object using computed tomography algorithms. A T-ray CT profile is shown in the bottom left, with a quarter cut away to reveal the internal structure. The spatial resolution of THz CT is given by the angular rotation multiplied by the radius of rotation at the edge of target. For the plastic ball, the radius is approximately 2 cm, therefore the spatial resolution on the surface of the ball is 3.5 mm.

interfaces reflect pulses with varying delay and intensity. For example, a slice of a floppy disc reveals layers of air, plastic and metal. Information on the complex dielectric profile of a multilayered object can be estimated using Fresnel equations (Sec. 4.2) and an iterative algorithm. Using the frequency-dependent focal length of a Fresnel lens (Fig. 12(a)) enables images at different depths in a sample to be obtained using the different Fourier components of the broadband T-ray pulse.

T-ray computed tomography (CT) and diffraction tomography are techniques for reconstructing an estimate of the internal structure of an object based on numerous transmission measurements. Fig. 27 shows a schematic T-ray CT spectroscopy system for imaging a small, in this case a ping-pong ball.

5.6. Near-field imaging

The lateral resolution of far-field T-ray images is limited by the wavelength of the radiation, \( \lambda \), to approximately \( 0.61 \lambda / (n \sin \theta) \), where the refractive index of the focusing medium \( n \) is typically unity for air and \( \theta \) is the half angle of the focal point. T-ray pulses have wavelengths spanning from 3 mm to 100 \( \mu \)m (0.1–3 THz), limiting the average resolution at the T-ray peak to approximately 500 \( \mu \)m. By selecting only high-frequency components after numerical Fourier analysis, the resolution can be improved to approximately 100 \( \mu \)m. To study broadband T-ray pulses interacting with sub-wavelength areas, however, near-field techniques are required. Near-field techniques rely on an aperture, diameter \( a \), placed in the optical near-field of the sample to be studied, with separation distance \( L < a \), so the size
Fig 28. Collection mode near-field EO imaging is performed by holding an EO sensor in the THz near-field of a sample, and measuring the T-ray electric field with a probe pulse reflected from the sample side of the crystal. This technique has a reduced signal because the internal reflection is small, but it provides resolution sufficient for the image of the letters 'THz'; the line width of the word is 0.5 mm and the image was taken in less than 1 s.

Fig 29. These images of the T-ray beam wavefront were taken with collection mode near-field EOS. The measurements were made in real-time and time evolution of the wavefront can be directly observed. The circular waves are due to the curved nature of the wavefront.

of the interaction spot is defined by \( a \) and not the wavelength of the radiation. A review of the development of FIR nearfield microscopy was prepared in 2002 by Rosner and van der Weide.

An early evaluation of using tapered waveguides as apertures for THz light was carried out using CW THz from a gas-vapor laser, and the first reports of near-field T-ray imaging used an elliptical aperture in the end of a tapered metal tip, \( a \approx 50 \mu m \) by \( 80 \mu m \). T-rays were focused to a diffraction-limited spot with a parabolic mirror, the tapered tip was placed at the focal point and a sample was placed in the near-field of the aperture, that is, within a length less than \( a \). A resolution of approximately 50 \( \mu m \) was achieved. Two typical aperturing effects were observed: the transmitted THz electric field dropped by approximately 130 times, and THz frequencies below 0.5 THz were strongly attenuated. A similar
Fig 30. 1D image of T-ray propagation through a focus. This image shows an image of a T-ray pulse passing through the focus of a lens. The focal point and Gouy phase shift are clearly visible. This image was sampled using 2D EOS and a 2-f imaging configuration with a polyethylene lens.\cite{171}

Fig 31. Near-field dynamic aperture schematic. As described in Sec. 5.6, a sub-wavelength-sized spot of THz can be blocked by the gated generation of photocarriers in a semiconductor. GaAs is normally transparent to T-rays, but the photocarriers act as a mirror screen.\cite{456} The schematic is the same as a typical T-ray spectrometer based on OR and EOS, except a gated beam for the sample is incorporated. Such a layout allows other optical-pump THz-probe experiments on ultrafast processes.\cite{400} The $\lambda/2$ and $\lambda/4$ waveplates are used to optimize the pump polarization and balance the photodiodes. The lens, $L$, is used to focus the optical beam to a sub-100 $\mu$m spot on the GaAs wafer.
Fig 32. This figure demonstrates the improved spatial resolution achieved by near-field imaging with a dynamic aperture. The sample is a metal circuit deposited on a GaAs wafer. This technique has demonstrated resolution of better than 50 μm.\textsuperscript{457}

Fig 33. T-ray contrast in biological tissue, showing diffraction-limited resolution and cell structure visible in the THz band.\textsuperscript{458}
resolution was achieved without an aperture, by holding the sample in the near-field of a PCA – in this case a virtual aperture existed due to the sub-wavelength dimensions of the spot where T-rays are generated.\textsuperscript{464}

Images can easily be acquired with the detector placed in the near-field of the sample, or in collection mode, using a large area EO detector.\textsuperscript{171} A schematic of near-field collection mode EO imaging is shown in Fig. 28. This provides a method to directly measure the 3D T-ray beam profile by moving the detector crystal along the beam path, and generates images of T-ray wavefronts or passing through a focus, shown in Figs. 29 and 30.

A resolution of 7 \( \mu \text{m} \) was achieved in 2001 using a collection-mode PCA detector.\textsuperscript{465,466} The detector had a 10-\( \mu \text{m} \) diameter aperture fabricated directly onto a 4-\( \mu \text{m} \)-thick PCA, so T-rays passed through the aperture and were detected by the PCA all in the near-field. The operation of the aperture is enhanced by having a high-refractive-index GaAs tip, \( n_{\text{GaAs}} = 3.6 \), protruding through it. The tip refracts the T-rays, decreasing the wavelength of the radiation as it passes through the aperture by a factor \( n_{\text{GaAs}} \).\textsuperscript{467} Two important considerations in T-ray near-field imaging are temporal reshaping of the pulse and the finite thickness of the aperture itself.\textsuperscript{310,330,468} Sub-wavelength T-ray spatial resolution can also be achieved using near-field metal tips,\textsuperscript{469} based on work in the infrared.\textsuperscript{470,471}

As mentioned in Sec. 4.3.3 above, dynamic T-ray experiments can be performed by generating transient mirrors in semiconductors, using photocarriers generated by an optical pulse. If the optical pulse is tightly focused, a dynamic inverse aperture can be created to block a sub-\( \mu \text{m} \) cross-section of the T-ray beam.\textsuperscript{457} A schematic of the dynamic aperture system is shown in Fig. 31. For a semiconductor placed at the focal point of a T-ray beam, and samples placed in the near-field of the aperture (\( L < a \)), a 40-\( \mu \text{m} \) resolution has been achieved with the dynamic aperture.\textsuperscript{456}

Wynne \textit{et al.} mounted a sample directly onto a EO crystal, so T-rays were generated in the near-field of the sample.\textsuperscript{472} A resolution of 200 \( \mu \text{m} \) was demonstrated, limited by two-photon absorption due to high intensity at the focal point.\textsuperscript{192} THz generation via OR in sub-THz-wavelength volumes is optimal when the optical spot size is comparable with the wavelength of the generated T-rays.\textsuperscript{473,474}

### 5.7. Speed

The trade-off of speed versus SNR is central to all T-ray spectrometers, particularly in imaging where large data sets have to be acquired in a realistic time, for example, while a person waits in proposed clinical applications. The speed of image acquisition can be estimated from the time taken to acquire a sufficiently-long T-ray waveform at each pixel at a certain SNR. For example, the first scanning PCA imaging system demonstrated 25-ps-long scans acquired in 5 ms with an SNR of 100:1.\textsuperscript{420} In 2001, high-power PCA T-ray systems reported 35-ps-long scans acquired in 20 ms with an SNR \( \approx 3000:1 \).\textsuperscript{143} The SNR increases proportional to \( \sqrt{t} \), where \( t \) is the time spent averaging data samples. For an SNR of 100:1, 2D EO imaging can be used to sample 200-pt-long 20-ps T-ray waveforms over a 288\times384-pixel array in approximately 200 ms. Signal processing can be used to improve the SNR of T-ray images\textsuperscript{298} including wavelets.\textsuperscript{475,476}
5.8. Dielectric imaging

Samples that have been studied in the search for applications of T-ray imaging include those where spectroscopic information is of interest, such as gas flames, and more complex samples, where contrast cannot be clearly linked to a single dielectric constant. Dielectric studies have been performed to image the carrier concentration and mobility in silicon wafers,\textsuperscript{477} the Hall effect (Sec. 4.3.3), biological tissues,\textsuperscript{458,478} currency watermarks\textsuperscript{479} and thin ceramic oxide with tomography.\textsuperscript{480} An example of T-ray contrast in biological tissue is shown in Fig. 33.

5.9. Classification algorithms

![Fig 34. Finite impulse response filter classification. This plot shows the effective classification of T-ray waveforms using coefficients b1 and b2 derived by fitting the data to one of two finite impulse response filters. The filters were developed with known training data. The fast algorithm clearly differentiates between bone and meat in this chart, depending only on the two coefficients.\textsuperscript{294} The use of classification in imaging is shown in Fig. 35.]

One of the advantages of T-ray imaging over incoherent imaging techniques, such as those using X-rays or visible light, is the coherent nature of the measurements. A full T-ray waveform can be acquired for each pixel and used to interpret the image spectroscopically. For 2D images, the intensity of each pixel can be linked to a number of spectroscopic parameters, including absorption or phase delay of a specific spectral component, peak absorption or peak delay. Images of flames have been false color coded to indicate the magnitude of T-ray pulse delay in transmission.\textsuperscript{298} The magnitude of the pulse peak has been used to code images of currency.\textsuperscript{441,481} More complex metrics can be developed to represent a combination of factors, specifically chosen to emphasize differences in spectra of different materials. This becomes a classification task, and the metrics can be algorithms trained from known data to be either one sample or another. The results of the application of a simple classification algorithm are shown in Fig. 34. Classification is a digital signal processing task, often using linear filters to generate the parameters of importance. T-ray images colored by classification metrics have been used for imaging the Hall effect in
Fig 35. These three images demonstrate the utility of a simple classification algorithm. The visible image on the left shows the sample biological imaged with T-rays: meat and bone. Using the classifier algorithm, the background, meat and bone can be automatically and clearly differentiated, as shown in the center image. Using the classifier provides far clearer data than just using transmission amplitude information. The image on the right is a plot of transmitted THz amplitude, and shows poor discrimination between the sample meat and bone. Classification is a method of utilizing the entire sample waveform to differentiate between just two classes.

6. Conclusion

T-ray sensing and imaging has developed and spread widely since the early 1990s. Initially present primarily in physics and electrical and electronic engineering research groups, T-rays are now used by physicists, engineers, chemists and biologists. The THz bandwidth can play a role in industry, homeland security and medicine.

There are currently over a hundred groups world-wide using T-rays and commercial systems are on the market. There are, however, still numerous hurdles to be overcome before T-ray spectrometers compete with other techniques in spectroscopy, biosensing, medical diagnostics and industrial imaging. These challenges include size, cost, output power, SNR, bandwidth, depth penetration, water sensitivity, spatial resolution, speed of data acquisition and the lack of a THz-frequency knowledge base. As indicated in this chapter, these challenges are being addressed and future development will probably find T-rays implanted in vital applications.

Acknowledgements

This work was funded in part by the US National Science Foundation, the US Army Research Office and the Australian Research Council. S. P. Mickan would like to acknowledge the support of the Fulbright Commission and Clough Engineering, Australia, and his supervisors Prof. Derek Abbott (the Centre for Biomedical Engineering, and the Department of Electrical and Electronic Engineering) and Prof. Jesper Munch (the Department of Physics, Applied Physics and Astronomy) of The University of Adelaide, Australia.
References


T-Ray Sensing and Imaging


323


464. I. Brener, S. Hunsche, Y. Cai, M. C. Nuss, J. Wynn, J. Lopata and L. Pfeiffer, Time resolved near field imaging and diffraction with sub-wavelength far-infrared dipole


Multistatic Reflection Imaging with Terahertz Pulses

Timothy D. Dorney
Department of Electrical and Computer Engineering, MS 366, Rice University, Houston, TX 77251-1892, USA

William W. Symes
Department of Computational and Applied Mathematics, MS 134, Rice University, Houston, TX 77251-1892, USA

Daniel M. Mittleman
Department of Electrical and Computer Engineering, MS 366, Rice University, Houston, TX 77251-1892, USA

Phone: (713) 348-5452 Fax: (713) 348-5686 E-mail: daniel@rice.edu

Recent advances in the technique of terahertz time-domain spectroscopy have led to the development of the first fiber-coupled room-temperature broadband terahertz sources and detectors. The fiber coupling permits the repositioning of the emitter and receiver antennas without loss of temporal calibration or alignment, thus enabling multistatic imaging. We describe a new imaging method which exploits this new capability. This technique emulates the data collection and image processing procedures developed for geophysical prospecting. We use a migration procedure to solve the inverse problem; this permits us to reconstruct the location, shape, and refractive index of targets. We show examples for both metallic and dielectric model targets, and we perform velocity analysis on dielectric targets to estimate the refractive indices of imaged components. These results broaden the capabilities of terahertz imaging systems, and also demonstrate the viability of the THz system as a test bed for the exploration of new seismic processing methods.

Keywords: Terahertz imaging, time-domain spectroscopy, inverse problem, tomography

1. Introduction

The development of terahertz 'T-ray' imaging in the mid-1990's generated a great deal of interest, as this technology can be used for a wide range of problems involving remote monitoring and inspection. The unique capabilities of the terahertz time-domain spectrometer have also inspired numerous explorations of new imaging techniques for terahertz radiation, including time-of-flight reflection imaging and ranging, emission imaging, near-field techniques for improved spatial resolution, interferometric imaging for improved depth resolution, chirped pulse imaging for single-shot acquisition, and near real-time display-mode imaging. More recently, the availability of fiber-coupled THz transmitters and receivers has created new opportunities for data acquisition and image formation. For example, Ruffin and co-workers describe a technique in which a THz wave interacts with a two-dimensional target, and is subsequently measured at many different positions. The data is used to reconstruct the field distribution at the position of the target through back-propagation of the Huygens-Fresnel diffraction kernel, a familiar result from the theory of holography. A more explicit application of computed tomography techniques has also been reported recently. We recently described a similar image formation procedure, also using fiber-coupled THz antennas. This latter approach, known as Kirchhoff migration,
borrows not from the optics community but instead from the field of seismic prospecting. Here, we describe in detail this novel method for generating images using terahertz pulses. Because of the similarities between sub-picosecond THz pulses and the seismic impulses used for geophysical studies, we are able to employ a mature set of algorithms for image formation. These algorithms are based on much of the same underlying physics as the more familiar electromagnetic propagation, in the sense that both are direct consequences of the description of wave propagation using Green's functions. In our case, some simplifying assumptions can be made because the propagating wave is in the form of a broadband transient. In essence, the migration approach places more significance on the travel time than on the amplitude of the measured wave. This approximation neglects important aspects of the collected data, so it necessarily reduces the quality of the generated image. It leads, however, to an imaging algorithm that is very simple to implement and is extremely robust against losses due to scattering or absorption of the propagating wave. For this reason, it is well suited for image formation in situations where the THz wave must propagate through a lossy or disordered medium either before or after interacting with the target. A situation of this sort would, of course, be very challenging to handle using conventional approaches.

In addition to broadening the scope of THz imaging capabilities, this work also offers new possibilities for the development of novel seismic processing algorithms. Generally, such advances in signal processing techniques are evaluated using simulated data, for which the effects of scattering or noise can be difficult to emulate. Another option in the testing of new imaging algorithms is to use real data from field studies. In such cases, however, the 'correct' answer is often not known with high precision, so comparisons between the new imaging algorithms and conventional methods are not always revealing. A third option is the use of scale models as test beds, an approach that can provide a valuable new perspective. Physical models have found extensive use in the geophysics community, both for the validation of new imaging algorithms and for the study of complex structures. The majority of these have used acoustic techniques, and there have been few considerations of geophysical test beds based on electromagnetic propagation. The results presented here demonstrate the feasibility of using few-cycle terahertz electromagnetic impulses as models for seismic image formation. Of course, there are a number of important differences between acoustic and electromagnetic waves, the most significant involving the issue of longitudinally polarized waves. Nevertheless, there are many commonalities as well, including such phenomena as refraction, diffraction and Snell's law. Also, unlike seismic waves, the wavefront of the free-space THz radiation can be easily manipulated using lenses or other bulk optical components. Finally, the use of electromagnetic radiation eliminates the issue commonly encountered in the use of acoustic transducers, involving the strong sensitivity to the contact conditions due to the strong impedance mismatch between the transducer and the surrounding medium. As a result, one might expect to learn a great deal from table-top model systems, such as described here.

In this work, we discuss the method of Kirchhoff migration as applied to terahertz imaging. In Section 2, we describe the fiber-coupled terahertz spectrometer used to perform these measurements. In Section 3, we review the basic principles of the migration approach and our experimental arrangement. In Section 4, we discuss issues of resolution, as well as the optimization of the data acquisition. This is most readily understood in the context of reflective (metallic) targets, for which we show several illustrative examples. Section 5 deals with the imaging of dielectric, rather than metallic,
targets. As with many inverse problems, there is a commonly encountered ambiguity in images for which the wave velocity is not known a priori. The seismic community has developed a number of sophisticated metrics for determining the optimal velocity maps in these cases. We discuss the application of the most appropriate of these metrics to the THz imaging problem.

2. THz Time-domain Spectrometer

The imaging methods described here require that the terahertz field be measured at a large number of spatial locations, in order to have enough data to perform a complete reconstruction of the desired image. Despite the obvious advantages inherent to obtaining these larger data sets, multistatic measurements of this type have not been widely reported. This is in large part due to the difficulties involved in designing a compact detector of THz radiation, which can be easily repositioned. For example, in terahertz time-domain spectroscopy (THz-TDS), a photoconductive antenna is triggered with a femtosecond optical pulse, creating a transient temporal gate for the sampling of an incident THz field. This requires a temporal synchronization between the optical trigger pulse and the sub-picosecond THz pulse, as well as the precise alignment of the optical beam onto the active area of the antenna. Typically, the temporal synchronization must be maintained to better than a few tens of femtoseconds, and the optical alignment must be accurate to within at least a few microns. Since the vast majority of THz-TDS implementations use a free-space coupling of the optical beam to the antenna, rapid repositioning of the detector is generally impractical. The above arguments also apply to detection methods based on free-space electro-optic sampling, in which the alignment of the co-propagating THz and visible beams is also critical.

The easiest method for maintaining both the temporal synchronization (i.e., the delay) and also the alignment of the optical beam (i.e., the relative sensitivity of the detector) is to couple the optical beam to the detector antenna by fiber optics. Froberg et al. first reported a fiber-coupling scheme in 1992, but this technique was not widely adopted. However, with the recent development of a commercial THz spectrometer employing fiber-coupled antennas, this tool is now generally available. This new capability permits easy repositioning of the detector without loss of either optical alignment or absolute temporal delay. Because it is not necessary to re-align the optical beam onto the antenna each time it is moved, many measurements can be made in a relatively short time, minimizing the effects of system drift. An additional important advantage afforded by the fiber-coupling is that one may rotate the detector as one translates it along a line, so that it is always pointing towards the target. This guarantees that the receiver antenna is always oriented normally to the propagation direction of the emitted wave. As a result, the angular sensitivity of the receiver antenna need not be a factor in these measurements.

Figure 1 shows a schematic of the setup used to perform these measurements. The experimental apparatus is derived from a commercially available THz-TDS system from Picometrix, Inc. This system is entirely fiber optic coupled and employs permanently aligned THz transmitter and receiver antenna modules. Many of the other components are similar to those in a conventional THz-TDS system. Laser pulses of approximately 100 femtoseconds duration at a repetition rate of 80 MHz are injected into single-mode optical fiber, after pre-compensation for group velocity dispersion in the fiber. The fiber is split into two arms, one for the transmitter and one for the receiver.
Fig. 1. We acquire data using the Picometrix fiber-coupled terahertz system that provides a constant temporal delay as the transmitter (Tx) and/or receiver (Rx) is translated. The system includes dispersion compensation to offset the effects of the optical fiber and a separate enclosure for fiber splitting, delay rail, and data acquisition. To simulate simultaneous multistage data acquisition, we translate and rotate the receiver along a line. A photograph of a packaged antenna module is shown.

The pulses in the receiver arm exit the fiber, transit through a computer-controlled optical delay, and are coupled back into fiber. The delay consists of a long rail that provides up to 1 nanosecond of delay, as well as a rapid scanning delay line that permits measurement of THz waveforms within a 40 picosecond window at a rate of 20 Hz. The optical power in the fiber is kept below 10 mW to avoid nonlinear pulse broadening. Even so, a signal-to-noise of 1000 per sweep of the scanning delay line is not unusual.

3. Two Dimensional Migration

Seismic data collection is generally arranged to acquire data at many points on a planar surface. An acoustic pulse is generated at the earth's surface, and propagates into the sub-surface strata in the form of an expanding spherical wave. This wave generates a series of reflected waves, which are recorded by an array of receivers on the surface. The task in migration is to transform this data set, knowing only the position of the transmitter and receivers and the travel times of the reflected pulse, into a useful image of the subsurface. In other words, given that the horizontal surface is the x-axis and depth is the z-axis, we wish to transform data from the (x, t) domain into the (x, z) domain. We review the theoretical and experimental foundations of migration. Schneider provides a useful guide from both the acoustic and optical perspectives that we closely follow in our discussion.
The theoretical foundation for the migration transform assumes a point source, generating a spherical wave that propagates into a homogeneous media. This wave is reflected back to a boundary at the recording surface $z = 0$. From first principles, the scalar wave equation governs the propagation of the wave. Given that we can observe the wave along the surface, the solution for the surface integral of the wave equation is desired. Several Green's functions have been suggested that solve this problem. The analysis is based on Huygen's principle with the solution commonly called the Kirchhoff integral.

![Diagram](image)

Fig. 2. (a) Schematic of a common shot experimental arrangement emulated by the THz system. Multiple symmetrically placed receivers are arranged to collect a series of reflected waveforms from a point scatterer. The travel time increases hyperbolically with the receiver offset from the transmitter. (b) Kirchhoff migration reconstructs the location and shape of a reflector by calculating the appropriate hyperbola, and summing the values of the recorded traces along that hyperbola, for an array of guessed locations. Incorrect locations generate a small summation since their hyperbolae do not pass through many reflected pulses. Two such incorrect guesses, and their corresponding hyperbolae, are shown.

Although we have not restated the mathematics of the transform, a more easily understood explanation can be formulated from the geometry of the experimental arrangement. In figure 2(a), a single transmitter with an array of symmetrically placed receivers is located at the recording surface. The emitted spherical pulse propagates downward, until a portion of it interacts with a point diffractor, generating a reflected wave. Given that the point scatterer is located at $(x_0, z_0)$, the transmitter is at $(0, 0)$, and a receiver is at $(x, 0)$, the travel time in a homogeneous media based on the Pythagorean theorem is:
\[ D(x) = v_0 \tau = \sqrt{x_0^2 + z_0^2} + \sqrt{(x-x_0)^2 + z_0^2}, \]  

(1)

where \( v_0 \) is the velocity in the medium, \( \tau \) is the two-way travel time, and \( D(x) \) is the total distance from transmitter to target to receiver.

In figure 2(b), a collection of waveforms from various receiver positions is shown so that the temporal axes run vertically. Based on Eq. (1), we are not surprised to find that the peaks of these waveforms follow a hyperbolic shape. Kirchhoff migration uses this fact to transform the collection of waveforms back to the location of the point diffractor. In practice, we use a grid of possible reflector locations. For each point, the expected hyperbola is calculated, and the amplitude from each waveform at the proper time offset is summed. The value of the Kirchhoff summation is then placed at the grid point. Correctly guessed points yield a large summation value since they traverse through the peaks of multiple waveforms. Incorrectly guessed points result in smaller values due to the low amplitude of the traversed waveforms and destructive interference from different waveforms in the summation.

Two incorrectly guessed points are shown as circles in figure 2(b). The hyperbolas associated with each of these points are centered above the circles. Both points have a small valued summation due to partial destructive interference. Some of the temporal waveforms have negative amplitude values at the points where they intersect the hyperbola, while others have positive amplitude values at the intersection points. The right-hand point is closer to the surface (\( z = 0 \)); therefore, its associated hyperbola has more curvature than the one on the left. These examples illustrate how even incorrect locations can generate non-zero amplitudes in the migration procedure and induce image artifacts. The characteristics of these artifacts are seen in several of the figures below. As the number of waveforms included in the summation increases, the amplitude of these artifacts decreases and the effects become less severe.

From the mathematical outline by Schneider, we note three major differences between theory and practice. First, the Kirchhoff surface integral is a continuous function, whereas practical implementation requires discrete summation to approximate the integral due to a finite number of receivers. Discrete summation introduces the need to prevent aliasing effects, which we discuss in Section 4. Second, Kirchhoff migration, strictly speaking, takes into account the reduction in amplitude of the wave as the transmitter-to-receiver distance increases. In our work, the transmitter-to-receiver distance is small enough to neglect this effect, given the negligible absorption of the intervening medium. The process of Kirchhoff migration without corrections for the relative change in amplitude between received waveforms is known as ‘diffraction summation’ as opposed to Kirchhoff migration, but these terms are often used interchangeably within the seismic community. We follow this trend and generally refer to our approach as Kirchhoff migration. Third, our waveforms are discrete time recordings. We must therefore employ interpolation when the intersection with a hyperbolic curve falls in between two time steps.

Kirchhoff migration provides a relatively simple means to reconstruct an image of a point diffractor from reflected waveforms. In general, we can also image extended objects since any surface may be represented as a collection of point scatterers. The ability to image an object, even if it contains surfaces tilted with respect to the incoming beam, emphasizes one unique feature of this imaging technique. With conventional THz reflection measurements, even those using a bistatic configuration, a surface with a tilt at a large angle does not reflect any radiation back to the receiver. The radar community
has understood the advantages of multistatic over bistatic imaging for many years.\(^3\)\(^5\)

Another unique aspect of Kirchhoff migration is the ability to estimate the parameters of a dielectric target such as thickness and refractive index. To properly create a depth image from temporal signals, we must also know the velocity of the wave in each stratum. We apply one such velocity analysis technique in Section 5. In fact, it should be possible to estimate velocity parameters for a complex two-dimensional dielectric target since Kirchhoff migration handles lateral velocity changes better than competing methods.\(^2\)\(^4\) There are fundamental limits, however, to the resolution of the reconstruction in both the horizontal (x-axis) and vertical (z-axis) directions. We examine these limits and discuss the requirements of the experimental arrangement in the Section 4.

Figure 1 shows a representation of the experimental configuration. The emitted radiation reflects off of a target, and is sensed by a receiver. Although we are collecting these waveforms serially, by repositioning a single receiver at many different locations, this process simulates simultaneous multistatic data acquisition. As noted above, since the receiver is fiber-coupled, repositioning does not alter the relative temporal delay of the laser arrival time or the alignment of the femtosecond optical beam onto the receiver. We translate the receiver along a line and rotate the antenna to compensate for its angular sensitivity.\(^1\)\(^7\) Great care is taken to ensure that the axis of this rotation passes through the antenna, so that the rotation does not introduce any temporal shifts in the measured waveforms. This configuration mimics the typical situation in geophysical imaging, where the transmitter is a device for generating a seismic impulse (e.g., dynamite) and the receivers are an array of geophones.\(^2\)\(^4\)

4. Resolution of reconstructed images

As discussed in the previous section, Kirchhoff migration provides a means for reconstruction of a point scatter from temporal data in a homogenous medium. This method also can reconstruct an arbitrary shape, since any surface may be regarded as a collection of point scatters. Limitations exist as to how well the reconstructed image is resolved. We discuss the theoretical limits and demonstrate the effects with both experimental and simulated data.

Horizontal resolution defines the smallest feature that can be resolved along the x-dimension (parallel to the array of receivers). It is generally defined in terms of the first Fresnel zone. Generally, the first Fresnel zone is defined by the size of an opening in an infinite plate such that only positive values of an incident spherical wave are able to penetrate the opening.\(^1\)\(^9\),\(^2\)\(^3\) For a broadband source, the first Fresnel zone can be defined by the size of an aperture which maximizes the transmitted energy.\(^3\)\(^6\),\(^3\)\(^7\) The size of the first Fresnel zone is given approximately by:

\[
\Delta x = \frac{v_0}{2} \sqrt{\frac{\tau}{f_{\text{mean}}}},
\]

where \(f_{\text{mean}}\) is the mean frequency of the THz source, and \(v_0\) and \(\tau\) are as defined in Eq. (1) above.\(^2\)\(^4\)

A feature that is smaller than the size of the first Fresnel zone cannot be resolved. As the feature size increases beyond the size of the first Fresnel zone, the image reconstruction provides a better representation. This viewpoint is equivalent to
asking how much the resolution improves as the hole in our infinite plate expands. As the size of the hole increases, more of the spherical wave is allowed to traverse the opening. Subsequent Fresnel zones alternate in sign compared to the first Fresnel zone. As additional Fresnel zones are available, the resolution improves, but at the same time, the effect diminishes as more zones are added.\(^{38}\)

Vertical resolution is defined as the smallest feature that can be resolved along the z-dimension (the initial beam propagation direction). This depends on the coherence length of the probing wave. A shift in position of a temporal waveform by an amount that is small compared to the duration of the THz pulse will not have a significant effect on the summation. Consequently, the migration summation is not sensitive to temporal shifts of this magnitude. The generated images exhibit blurred edges, due to the finite coherence length of the radiation. The vertical resolution is given by:

\[
\Delta z = \frac{v_0}{4\Delta f},
\]

where \(\Delta f\) is the bandwidth of the radiation.

We demonstrate the limits of both horizontal and vertical resolution with an experiment using long metal cylinders of various diameters. These cylinders may be taken to be ideal reflectors. Each cylinder is placed approximately 90 mm away from a fixed THz transmitter. As in figure 2(a), a series of 152 reflected waveforms are collected on either side of the transmitter in 1 mm steps. This step size is somewhat smaller than the effective aperture of the receiver antenna. The smallest transmitter-to-receiver offset is 38 mm, limited by the size of the antenna housings. Images are formed from the resulting set of waveforms using the migration procedure outlined above. These results are shown in figure 3, for five cylindrical targets with various diameters. The vertical axes show the distance from the transmitter, while the dashed circles show the actual positions and cross-sections of the targets. For all five images, we used a grid spacing (pixel size) of 50 \(\mu\)m for the image reconstruction. The migration results place the targets at the proper position within our ability to independently determine them.

We note that the two smallest cylinders are accurately placed; however, they are not well resolved. The diameters of these cylinders are very close to the horizontal resolution. For our measurements, the size of the first Fresnel zone is \(\Delta x = 2.9\) mm. For the 4.7 mm diameter cylinder, we begin to resolve features in the image which correspond with the location of the reflecting surface. The images of the two largest cylinders both have a clearly defined surface section. Only a portion of the surface is imaged due to the finite range of receivers along the x-axis. The limited number of receivers hinders complete cancellation of diffraction sums from regions without a reflector. Consequently, a number of image artifacts arising from aliasing effects are observed, as anticipated above. These artifacts blend into the surface reconstruction at its outer limits, and appear to simply continue the surface along a tangent line on either side of the cylinder. The angles of these artifacts are determined both by the distance from the transmitter to the object (which determines the curvature of the hyperbolic reflection curve) and by the shape of the reflecting object. The reconstructed curves of the cylindrical surfaces also exhibit a finite thickness. The finite coherence length of the radiation, which limits the vertical resolution, causes blurring that is 3-4 pixels wide.\(^{23}\)

The surface blurring is consistent with the calculated vertical resolution of \(\Delta z \approx 0.19\) mm.
Fig. 3. Kirchhoff migration images from five metal cylinder targets with diameters of (a) 2.4 mm, (b) 3.2 mm, (c) 4.7 mm, (d) 6.2 mm, and (e) 12.7 mm. The dashed curves represent the outlines of the targets for comparison with the migration images. In (a) and (b), the objects are correctly located, but their surface curvature cannot be resolved since their diameters are close to the resolution. In (c), the location of the object and a very faint surface curvature is resolved. In (d) and (e), the cylindrical surfaces are reconstructed over a limited region due to the finite range of receiver offsets.

The experimental arrangement may also have a large effect on the reconstructed image. Obviously, the resolution limits are determined by the bandwidth of the THz pulse, but increased artifacts and lack of detail in complex shapes may result if the experimental arrangement is not properly designed. Two important concerns are the receiver's range, or span across the surface, and the spacing between adjacent receiver locations.

Receiver span determines the angle of a target that can be observed. Simple geometry provides sufficient insight, and the reconstructed cylinder images in figure 3 provide a useful illustration. With a receiver offset from 38 mm to 189 mm on either side of the transmitter, the observed section of the cylinder's curvature is limited. Increasing the receiver's span allows more of the target to be reconstructed but at the expense of additional acquisition and processing time.

Given the desire to acquire data over a broader range of receiver positions, it might seem obvious to simply increase the step size between receiver positions.
Diffraction sums, however, rely not only on coherent summation to define where an object exists but also rely on destructive summation of incorrectly guessed hyperbola. As the receiver spacing increases, the ability to create destructive interference using Kirchhoff migration diminishes. In fact, a receiver spacing that creates a shift in the temporal waveforms greater than half of the temporal duration of the pulse effectively destroys the ability to properly resolve the image. Equally, the receivers must be arranged such that such a shift does not occur between any pair of adjacent receivers. Complex targets may have a large effect on the temporal shift of the acquired waveforms.

We define the upper bound of the stepping distance in terms of the signal bandwidth $\Delta f$:

$$
\Delta D_n = \left| D(x_n) - D(x_{n+1}) \right| \leq \frac{v_0}{2 \cdot \Delta f},
$$

where $D(x_n)$ is the transmitter-to-target-to-receiver distance of adjacent receivers and $v_0$ is the velocity of the wave in the homogeneous medium. For simple structures, this provides a means to determine the receiver spacing. For example, assume that a flat reflector, parallel to the transmitter/receiver line, is located 90 mm from the surface. A

![Fig. 4. Reconstruction of a planar reflector using a single transmitter and a group of receivers on one side of the transmitter with different receiver spacings demonstrate the effect of aliasing. In (a) through (d), the receivers are spaced every 1 mm, 2 mm, 3 mm, and 4 mm, respectively, within an offset range of 38 mm to 157 mm.](image-url)
receiver position step size of 1 mm (e.g., from 38 mm to 39 mm) results in a difference \( \Delta D_n \) of \( \sim 0.21 \) mm, which is less than the upper bound of \( \sim 0.3 \) mm for a 0.5 THz bandwidth. A stepping distance of 1 mm is appropriate, and it is used for all experimental data in this paper.

To demonstrate the effects of aliasing due to improper receiver placement, we use simulated data. A flat, reflective plate is located at \( z = 90 \) mm with a single transmitter and 120 receivers positioned on one side of the transmitter at 1 mm spacings. The minimum transmitter-to-receiver offset is 38 mm. Four images were reconstructed using the data, each with a progressively reduced data set. In figure 4(a), we use all 120 waveforms in the migration. Artifacts are apparent, but are relatively small compared to the reconstruction for the reflecting plate. For figure 4(b), alternate waveforms were discarded (i.e., receivers 1, 3, 5,..., 119 were used) before the Kirchhoff migration; while figure 4(c) used data from receivers 1, 4, 7,..., 118; and figure 4(d) used data from receivers 1, 5, 9,..., 117. We observe a dramatic effect as the aliasing artifacts get progressively worse with each increase in receiver stepping distance.

5. Velocity estimation with semblance

In Section 4, we demonstrated the ability of the THz system to image curved surfaces. Imaging steep angles relative to the transmitter is possible with a multistatic receiver arrangement. This experiment is a useful illustration, but contains only a single homogenous medium (air). A more interesting problem involves imaging dielectric targets. Dielectrics allow both reflection and transmission of the electromagnetic field and gives rise to regions of differing velocities within the reconstructed image. To properly create an image, we require a method to estimate the different velocities. In an effort to emulate seismic data acquisition, we study layered media, a common geologic structure. To illustrate the important steps in image formation, we start with simulated terahertz data. These simulations demonstrate the challenges inherent in this image reconstruction problem, which are formidable even without noise or other experimental considerations. Once we have covered the salient points, we then turn to experimental results, using a composite target with a very small impedance mismatch. This data illustrates the sensitivity of the THz system, and the applicability of the velocity estimation procedures outlined below.

For the simulations, a simple planar target is a useful starting point. A single homogenous layer is oriented parallel to the transmitter/receiver line. Its top surface is at \( z = 90 \) mm, and its bottom surface is at \( z = 100 \) mm. The regions both above and below this planar target are empty. The refractive index of the simulated material is set to 1.525, which is roughly equivalent to that of high density polyethylene (HDPE). The material is assumed to be nondispersive and non-absorbing. Figure 5 shows the data acquisition arrangement, in which a group of one transmitter and 20 receivers is moved across the layer. The initial transmitter-to-receiver offset is 38 mm, with the receiver spacing set at 2 mm. Data from these 20 receivers is collected every 2 mm at 10 different locations. We note that seismic data collected at sea uses a similar technique, where the acoustic source is an airgun followed by a towed streamer of acoustic receivers.

We wish to use the same Kirchhoff (diffraction) summation described above to create an image from the entire data set. Previously, we used only common shot data, or in other words, data acquired by a set of receivers from a single transmitter location. We now have several advantageous data arrangements, such as common midpoint and
Fig. 5. Data acquisition of a synthetic planar, homogenous dielectric material using a moving group of one transmitter and 20 receivers. This arrangement allows common offset, common midpoint, and common shot gathers. We use a refractive index of 1.525, similar to that of high density polyethylene (HDPE).

common offset, that can be extracted from multistatic data sets of this type. Common midpoint data allows multiple views of the same spatial location. In effect, we use these various views to reduce noise (e.g., through averaging) or infer structure by noting variations in waveforms not accounted for by the change in transmitter-to-receiver distance. Common offset is useful in determining angle and/or velocity changes as the transmitter-to-receiver offset is fixed. Time of flight variations indicate a change in depth or tilt of the reflector, or velocity of the intervening media.

One important difference between the application of Kirchhoff migration to electromagnetic and acoustic scattering is that, in the seismic experiment, the absolute propagation time is known while the wave velocity \( v_0 \) is generally not known. With our data, the wave velocity of the first layer (air) is known precisely, but the time between the emission and detection of the THz pulse is not known since the experimental measurements only provide relative delays. In both cases, however, the unknown quantity can be obtained from Green's equation:

\[
T_x^2 = T_0^2 + \frac{x^2}{v_0^2},
\]

where \( T_x \) is the absolute transmitter-to-receiver travel time for a receiver at \((x, 0)\) and \( T_0 \) is the travel time for a (hypothetical) receiver at \(x = 0\) (the location of the transmitter).\(^{39}\) Using a flat reflector oriented parallel to the receiver line, a plot of \( T_x^2 \) versus \( x^2 \) produces a straight line. If the absolute travel time is known, then the slope of the line determines the wave velocity. In our case, we vary the temporal offset of all waveforms until the slope equals \( 1/v_{air}^2 \); then the intercept gives the absolute value of \( T_0 \), and therefore the transmitter-to-target distance \( (z_0 = v_{air} T_0 / 2) \). Figure 6 shows the arrival time of the
Fig. 6. The absolute propagation time for the THz system is determined by plotting the squared receiver offset for a flat plate while adjusting the squared two-way travel time until the slope equals $1/v_{air}$. Green's technique is used to determine the wave velocity $v_w$ when the absolute time is known. We know the wave velocity precisely but do not have an absolute time reference.

primary reflection from common shot data after the appropriate time shift of the waveforms. We used the absolute time $T_0$ to calibrate the time axis in the above cylinder experiments.

After adjusting the temporal waveforms to an absolute time scale, we perform Kirchhoff migration using a 50 µm pixel size for the flat plate drawn in figure 5. The results are shown in figure 7; as expected, both the front and back surfaces of the plate are visible. We note, however, that the distance between the surfaces of the plate is not equal to the thickness of the sample. This is a result of the assumption, implicit in the

Fig. 7. Kirchhoff (diffraction) summation of the synthetic data collected in Fig. 5 using a 50 µm grid clearly shows the two surfaces of the flat plate. The top surface is accurately placed; however, the z-distance separation between the two surfaces is larger than the simulated thickness of 10 mm. Velocity analysis is required and is performed over the region indicated by the dashed box. Figure 8 displays the results of this velocity analysis.
migration procedure, that the wave velocity is constant everywhere within the imaged region. The velocity of light within the material is ~33% slower than in air. Consequently, the surfaces appear to be farther apart than they actually are. This type of image artifact was recognized even in the earliest reports of THz reflection imaging. We now examine several means to estimate velocity, with the goal of correcting this artifact.

Several simple but effective methods to determine velocities in geophysical imaging are constant-velocity gathers (CVG) and constant-velocity stacks (CVS). With CVG, a uniform velocity is assumed for all data (e.g., common midpoint), and the temporal waveforms are drawn side-by-side, or gathered. Each waveform's time base is adjusted to properly align the data. The adjustment, or moveout, is determined by:

\[ \Delta T = T_x - T_0 = \sqrt{T_0^2 + \frac{x^2}{v_g^2}} - T_0, \]

where \( T_0 \) is the travel time for a (hypothetical) receiver at \( x = 0 \) (the location of the transmitter) as in Eq. (5), \( v_g \) is the guessed constant velocity, and \( T_x \) is the absolute transmitter-to-receiver travel time for a receiver at \( (x, 0) \). The moveout value \( \Delta T \) is used to shift the time axis of each reflected waveform. A series of different gathers, each with a slightly different guessed velocity, are plotted side-by-side on a time versus estimated velocity graph. Visual trends indicate good velocity estimates. An under-corrected gather (where the guessed velocity \( v_g \) is too fast) creates downward tilting curves from flat reflectors (parallel to the line of receivers). Over-corrected gathers produce upward tilting curves from flat reflectors. By selecting regions at differing times that have 'flat' features, the correct velocities at each time can be determined.

Constant velocity stacks use a similar approach to CVG, but data is first migrated (summed) across hyperbola as in Eq. (1). By plotting the stacked (migrated) data, each with a slightly different constant velocity, we look for 'flat' regions that indicate coherent summation (i.e., correct velocities). Typically, CVG and CVS are used concurrently with the CVG data taken from one of the data sets used in the CVS. Both of these procedures rely to a great extent on a visual inspection of the data; however, these methods are appropriate for seismic reconstruction since acoustic velocity generally increases with depth. In fact, a relatively new technique called differential semblance optimization goes so far as to assume a smooth velocity function and moves the discontinuities created by impedance mismatches into a reflectivity function.

Both CVG and CVS methods assume a root-mean-square (rms) velocity. In other words, if a layered target has velocities \( v_1, v_2, v_3, \ldots, v_n \), and a one way travel time of \( t_1, t_2, t_3, \ldots, t_n \), for each layer, respectively, the rms velocity is defined as:

\[ v_{rms}^2 = \frac{\sum_{k=1}^{n} \frac{v_k^2 t_k}{t_1 + t_2 + t_3 + \cdots + t_n}}{\sum_{k=1}^{n} t_k}. \]

Given the sharp velocity discontinuities created as light moves from air to HDPE in our sample target, an average velocity does not provide adequate information. These types of sharp discontinuities are also exhibited in the very challenging Marmousi benchmark seismic velocity model.

This still leaves us in need of a method to analyze our simple target. One
approach is to break up the velocity function into sections and continue the analysis downward using an iterative approach, in which the velocity in each layer is determined using the value from the previous layer.\textsuperscript{23} We apply this approach by using Kirchhoff migration in tandem with a velocity field. The velocity field determines the one way travel time from a depth point ($z > 0$) to a point on the surface ($z = 0$). The model includes velocities defined over the entire $(x, z)$ space of interest and any deviations due to Snell's law at interfaces. The cumulative depth-point-to-transmitter and depth-point-to-receiver travel times determine the path of the hyperbola used for summation of the data set values. The velocity model is downward modified as each new interface is encountered.

![Graph showing velocity and coherence](image)

**Figure 8.** (top) Comparison of eight different velocity fields applied to the Kirchhoff migration of the second interface in Fig. 7 shows a dramatic change in the placement of the layer. We note a generally decreasing spread of the reconstructed interface as the guessed velocity decreases; however, there is no simple way to distinguish among these guesses. The background has been amplified to display the artifacts. (bottom) We apply the semblance coherency metric to each column of the migration indicated in Fig. 7 from 95 to 110 mm. The maximum semblance values for each column increase with decreasing velocity, but show a curve due to the artifacts on both the left and right sides of the reconstruction. The maximum semblance value occurs at $2.0 \times 10^8 \text{ m/s}$, close to the actual value of $1.97 \times 10^8 \text{ m/s}$.

Figure 8 (top) shows the Kirchhoff migration results for the second layer of the simulated target (the layer in which the index is 1.525) at guessed velocities from $2.8 \times 10^8$ to $1.8 \times 10^8 \text{ m/s}$, corresponding to guessed refractive indices from 1.07 to 1.67, respectively. The first layer location and velocity are determined by the results in figure 7, since the velocity in the uppermost layer is known. The migration images show a dramatic change in the location of the second interface as the guessed velocity in the
The background artifacts have been amplified in figure 8 (top) compared to figure 7 to emphasize the differences created by the reconstruction; however, there is no good way to distinguish by inspection between these various results. This ambiguity, arising from the fact that neither the thickness nor the wave velocity are known, is commonly encountered in many different manifestations of the inverse problem. A quantitative measure of the quality of the interface in the migration image is required to distinguish among the various guesses displayed in figure 8.

One aspect of these images that can be analyzed to provide this metric is the extent to which the wave fronts align during the migration procedure. Several coherency metrics exist such as velocity spectra, crosscorrelation, and semblance. A velocity spectrum relies on \( v_{\text{rms}} \) which, as noted earlier, will not suffice for this example. Semblance has been shown to differentiate smaller velocity changes than crosscorrelation; therefore, we use semblance as the coherency metric.

The semblance principle can be understood by noting the effect a change in velocity has on the location of a reflector, as in figure 8 (top) for example. We create multiple reconstructed images using different subsets of the data and the same guessed velocity. Since these multiple images are all of the same region of the target, they should all appear identical. This will only be true if the guessed velocity is the correct one. In essence, we create different views of the same target and note the variations from view to view, arising from incorrect velocity estimates. We define an image \( I(x, z) \) generated from Kirchhoff summation as:

\[
I(x, z) = \sum_T \left( \sum_R f(R, t(R, T, x, z)) \right),
\]

where the hyperbolic intersection occurs across many waveforms \( f \) for a given receiver location \( R \), at the time index \( t \). The intersection is determined by the location of the transmitter \( T \), the location of receiver \( R \), and the subsurface point of interest \((x, z)\). Given that many images are formed from different subsets of the data using the same guessed velocity, we have multiple images \( I_m(x, z) \). The semblance metric is defined as:

\[
S = \frac{\left( \sum_M I_M(x, z) \right)^2}{M \sum_M I_M^2(x, z)},
\]

where all images are pointwise summed then squared in the numerator, and pointwise squared then summed in the denominator. The division by \( M \) simply normalizes the metric, so \( 0 \leq S \leq 1 \).

For the simulated data acquired from figure 5, we generated Kirchhoff migration images using common offset data on a 50 \( \mu \text{m} \) square pixel grid. Common offset data reduces the amount of artifacts at the left and right edges of the reconstruction since it spans a wider range compared to common shot data. The reconstructed region from \( z = 95 \text{ mm} \) to 110 mm encompasses the second interface in all velocity estimates in figure 8 (top). From the twenty images created using data from 10 transmitter/receiver locations for each velocity, we generated a semblance image. The plots in figure 8 (bottom) show the maximum semblance value in each column of the semblance images.

The semblance results shown in figure 8 (bottom) provides a quantitative metric for selecting the appropriate guessed velocity for the region between the two interfaces.
Ideally, the semblance would provide a similar response across the flat secondary interface; however, artifacts, evident in figure 8 (top), create a reduced semblance on both the left and right sides of the reconstruction, and likewise, in the semblance results. These results, however, do indicate that an incorrectly guessed velocity positions the reconstructed interface at different locations depending on the receiver offset from the transmitter. Semblance records the lack of coherency as a lower normalized output-to-input energy ratio. The guessed velocity of $2.0 \times 10^8$ m/s has the maximum semblance, while the actual velocity in this simulated target is $1.97 \times 10^8$ m/s. Next, we extend these results to more complex targets.

The previous simulated example was very simple, and could have been handled by several less computationally expensive methods. Specifically, Dix provides a means to solve for an arbitrary number of velocity intervals. Using this analysis method, any errors on earlier or later interfaces does not effect subsequent velocity calculation. This seems to be an ideal approach; however, application of this method to highly complex targets quickly becomes untenable as the analyst must apply simple geometric shapes to represent complex reflectors.

Although continuous downward modifications will be needed to update the velocity field with respect to angled reflectors, multiple layers, synclines (concave up features), and anticlines (concave down features or humps), we continue to use Kirchhoff migration and semblance. Kirchhoff migration (diffraction sums) performs well even with lateral velocity $v(x)$ shifts and semblance is a more robust coherency measure.

To illustrate these points, we study a composite dielectric target with substantially smaller refractive index discontinuities than those in the simulated example described above, using the terahertz spectrometer. The experimental target is designed to mimic layered strata common in the geophysics community. The only a priori information used in the analysis is the refractive index of the initial layer, which is air, and the transmitter and receiver positions. This approach follows the work of Neidell and Tanner.

Figure 9(a) shows a schematic of the composite target. Only general dimensions are displayed since several irregularities exist as a result of the milling process. The target is composed of a top layer of teflon. Two additional plates, one of HDPE and a second teflon plate, are compressed onto the top plate. The difference in velocities between teflon and HDPE is only $\sim 6\%$ compared to the $30\%$ difference between teflon and air. Furthermore, the thickness of the bottom teflon plate is machined to 1.65 mm while the HDPE is 1.55 mm. With these thicknesses, the bottom two plates induce the same optical delays at normal incidence. The goals of this experiment are to correctly locate the four interfaces, and also to identify the composition of each layer by accurately determining its refractive index, using the semblance analysis.

To analyze this composite target, one transmitter and 19 receivers were translated over the target. The initial receiver offset was approximately 34 mm, and the receiver spacing was 1 mm. The group of transmitters and receivers acquired data at 25 positions with a 1 mm spacing, for a total of 475 acquired terahertz waveforms. Figure 9(b) shows one of the common offset gathers with the corrected temporal offset. The first flat region of the target was used to calculate the absolute temporal offset with the method described above for Eq. (5). The interfaces are a bit hard to discern due to the small amplitude of the reflections from the nearly impedance-matched interfaces.
Figure 9(c) shows the Kirchhoff migration, assuming a constant velocity throughout the imaged region. As above, the pixel size is 50 μm. From this migration, the four interfaces are clearly displayed. The noise is dramatically reduced compared to figure 9(b). The four marks superimposed on the figure show the actual locations of the interfaces. The first interface is correctly located because the index of the intervening medium (air) is known; however, the remaining interfaces are not correctly located because of the non-uniform velocity field. As in figure 8, we require velocity analysis to properly locate these buried interfaces.

Fig. 9. (a) Construction of an experimental, composite target includes a top layer of teflon followed by high density polyethylene (HDPE) and another layer of teflon. The second and third layers are machined to have the same optical delay at normal incidence. This stack creates a smaller change in velocity between layers compared to the simulated air/HDPE interface in Fig. 5. (b) Common offset data after filtering shows a large amount of noise. The interfaces are hard to discern; however, constant velocity migration results in (c) dramatically reduce the noise and reconstruct the image. The indicators at left denote the expected locations of each interface. These demonstrate the need for velocity analysis to properly locate the interfaces.
We apply the semblance analysis not to the entire image, but instead to a small region within the image. To determine the velocity below the first interface, we use a small analysis window that encompasses the interface of interest. In this example, the window size is chosen to be 3 mm high and about 1.5 mm wide. Each interface is analyzed in turn, starting with the layer closest to the top (known) interface. We guess a constant velocity below the previous interface, and compute the semblance at each pixel within this window. The semblance, averaged over all pixels in the window, is shown in

Fig. 10. The normalized semblance averages for the second, third, and fourth interfaces are shown in (a) through (c), respectively. The semblance values were calculated for each 50 μm wide column over an approximately 3 mm high window encompassing each interface. The average semblance for each velocity bin was taken over 30 columns. Unlike the results in Fig. 8, noise and multiple oscillations at the interfaces create a challenge for the semblance metric. The final Kirchhoff migration using the velocities taken from (a) through (c) is displayed in (d). The indicators, unlike Fig. 9(c), match the locations of the reconstructed interfaces.
figure 10(a) as a function of the guessed velocity. The maximum average semblance determines the correct velocity in the first dielectric layer. This process is then repeated at the third interface (bottom of the second dielectric layer), using the (now known) velocity of the first layer to determine the velocity of the second layer. The process is iterated once more to determine the velocity in the third layer. The average semblance values for each layer are shown in figure 10.

Unlike the simulated results in figure 8 (bottom), noise and multiple reflections at the interfaces create a challenge for the semblance metric. Nevertheless, the velocities in each of the unknown regions are determined with surprising accuracy. The refractive indices of the upper and lower layers (both teflon) are found to be 1.46, while the index of the middle layer (HDPE) is 1.54. These values compare favorably with the literature values of 1.43 and 1.52 for these two materials. From the widths of the semblance peaks in figures 10(a) and (b), we estimate an uncertainty in the determined velocities of roughly $\pm 0.5 \times 10^7$ m/s, corresponding to an uncertainty in refractive index of about $\pm 0.04$ for these materials. It is clear from the arrows in Figs 10(a)-(c) that this margin is sufficient to distinguish between the two very similar materials used in this composite target. Using the determined values for the velocities in each region, we can compute a revised image by Kirchhoff migration. This image is displayed in figure 10(d). We note that, unlike the uncorrected image in figure 9, the layer spacing in figure 10(d) is quite accurate. The positions of each of the buried interfaces (i.e., the thicknesses of the layers) are determined to within ~10%. The accuracy is limited by the coherence length of the THz pulse, as described above.

6. Conclusions

We have described a new image acquisition procedure using a fiber-coupled THz-TDS system. Multistatic THz-TDS data acquisition allows more complex targets to be analyzed than traditional approaches. This arrangement emulates the techniques of seismic imaging and borrows data processing methods from this mature field. The resulting images provide detailed information about the targets, using a very simple and easily implemented algorithmic approach. The image reconstruction techniques provide quantitative information about both the thickness and refractive index of the components of a composite target. It can be used to analyze buried interfaces through an iterative approach in which the velocity in any particular layer is determined using the velocity of the preceding layer(s). Because of the large data redundancy, as well as the fact that migration relies primarily on pulse travel times rather than amplitude information, this imaging method is more robust against noise than many other multistatic imaging techniques.

To demonstrate these procedures, we simulated seismic data collection configurations on an optical bench using terahertz time-domain spectroscopy, and use migration signal processing techniques for image formation. These basic techniques are well known in the geophysics community, and have many analogies to imaging methods developed for electromagnetic waves. The development of new imaging algorithms for geophysical prospecting is an active area of research. Although we do not attempt to expand the current geophysical imaging techniques, it is clear that the use of real world data creates situations not generally modeled in computer simulations. These results demonstrate the viability of the THz system as a test bed for the exploration of new seismic processing methods involving more complex model systems. These first steps
open the door for the THz system to probe more challenging media and morphologies, and the investigation of new or improved reconstruction algorithms. Examples of challenges for which the THz system could be useful include strongly scattering media, or stratified samples containing dispersive layers. Both of these situations are encountered in geophysical prospecting, and both pose significant difficulties in image reconstruction.

Acknowledgements

We acknowledge partial support for this work by the Environmental Protection Agency and the National Science Foundation. The assistance of J. Van Rudd and Matt Warmuth at Picometrix, Inc., is gratefully acknowledged.

References

698 T. D. Dorney, W. W. Symes & D. M. Mittleman


32 W. A. Schneider, "Integral formulation for migration in two and three dimensions," Geophys., 43, 49-76 (1978).


