Piloting Epitaxy with Ellipsometry as an In-Situ Sensor Technology

by

Sean C. Warnick

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

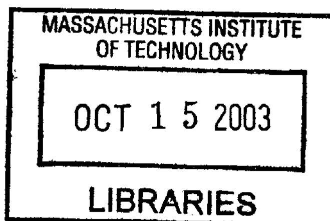
September 2003

© Massachusetts Institute of Technology 2003. All rights reserved.

Author .....
Department of Electrical Engineering and Computer Science
July 25, 2003

Certified by ....
Munther A. Dahleh
Professor
Thesis Supervisor

Accepted by
r C. Smith
Chairman, Department Committee on Graduate Students


MASSACHUSETTS INSTITUTE OF TECHNOLOGY
OCT 1 5 2003
LIBRARIES

BARKER

Piloting Epitaxy with Ellipsometry as an In-Situ Sensor Technology

by

Sean C. Warnick

Submitted to the Department of Electrical Engineering and Computer Science
on July 25, 2003, in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Abstract

Epitaxial processes are deposition processes that produce crystalline films with nano-scale precision. Many compound semiconductor devices rely on epitaxy to produce high-quality crystalline films with a specified compositional profile as a function of film thickness. Although various epitaxial processes, such as molecular beam epitaxy or metal-organic chaemical vapor deposition, have long been known to be capable of producing such films with mono-layer precision, doing so often requires significant calibration and many trial-and-error attempts before success is realized.

Ellipsometry is an established optical characterization method that uses the polarization state of reflected light to determine the alloy composition of a grown film. As a non-invasive sensor technology, ellipsometry has more recently been deployed as an in-situ sensor to characterize films as they grow.

This thesis takes a comprehensive view of the issues and impact associated with the use of ellipsometry as an in-situ sensor to control epitaxy in real time. A dynamic model of the physics associated with ellipsometry and growing films is developed from first principles, and this model is used to highlight some of the capabilities and limitations of ellipsometry as a measurement device for feedback control. A control problem is then formulated that substitutes the actual design objective with one more amenable to feedback design, and standard linear tools are used for feedback design. Simulations show that these designs look promising, even for the reliable growth of quaternary graded structures.

Thesis Supervisor: Munther A. Dahleh
Title: Professor

Acknowledgments

To my friends, I thank you for your stimulating conversation, critical ear, and willingness to share your insights toward my work. To my family, I thank you for your inspiration, rollicking laughter (even at my worst jokes!), and unfailing faith in me.

This thesis is dedicated to the memory of my brother, Ward W. Warnick.

Contents

1 Introduction **13**

1.1 Overview 13

1.2 Motivation 14

1.2.1 Optical Coatings Industry 15

1.2.2 Non-optical Thin Film Technologies Industry 16

1.2.3 Silicon-based Semiconductor Device Industry 16

1.2.4 Compound Semiconductor Device Industry 17

1.3 Background 18

1.3.1 Light Emitting Diodes (LEDs) 18

1.3.2 Transistors 19

1.3.3 Photodetectors 21

1.3.4 Non-Optical Sensors 23

1.3.5 Solar Cells 24

1.3.6 Optical Filters 25

1.3.7 Lasers 26

1.3.8 Vertical Cavity Surface Emitting Lasers 28

2 Ellipsometry as a Sensor Technology **33**

2.1 Background Physics 34

2.1.1 Electromagnetic Fields and Maxwell's Equations 34

2.1.2 Constitutive Relations 35

2.1.3 Uniform Plane Waves 36

2.1.4 Boundary Conditions and Inhomogeneous Media 39

2.1.5 Optical Response of Plane-Stratified, Stationary Films 41
2.2 Modeling Ellipsometry 48
2.2.1 Measuring Polarization 48
2.2.2 Non-Stationary Growing Films 55
2.2.3 Frozen Time Assumptions 60
2.3 Analysis 61
2.3.1 Equilibria and Fresnel's Formulas 62
2.3.2 Local Behavior of the System 63
2.3.3 Zero Dynamics and Brewster's Angle 63

3 Control of Epitaxy using Ellipsometric Feedback 67

3.1 Epitaxial Processes 68
3.1.1 Growth Dynamics 68
3.1.2 Actuation Devices 69
3.1.3 Production Processes 72
3.1.4 Process Model 76
3.2 Growth Objective 77
3.2.1 Band-gap Engineering 77
3.2.2 Measuring Film Quality 81
3.2.3 Performance Criterion 84
3.3 Tracking Graded Quaternaries 85
3.3.1 Problem Formulation 85
3.3.2 Linearization 92
3.3.3 Simulation Results 96

4 Piloting Epitaxy 99

4.1 Implementation 99
4.1.1 Experimental Set-up 100
4.1.2 Pre-Experiment Testing 101
4.2 Future Work 104

List of Figures

1-1 Fabrication processes in InP HBT and HEMT MMIC production. Critical nodes are shaded for transistor performance, rounded for passive component performance. 21

1-2 Typical reverse biased p-i-n photodiode with corresponding I-V characteristic. 22

1-3 Avalanche photodiode design, and a schematic of the resulting fabricated device. 23

1-4 Schematic of a single junction solar cell (left), and the cross-section of a multi-junction cell developed at NREL (right). 25

1-5 Typical optical data link (left) and a schematic of a channel dropping filter (right). 26

1-6 Semiconductor lasers use population inversion to induce conditions for spontaneous emission. 27

1-7 Edge emitting (left) and vertical cavity surface emitting lasers (right). 29

2-1 A linearly polarized wave oscillates without twisting around its axis of propagation. 39

2-2 Geometry of the physical system defining the optical response of a stationary stratified film with $\phi = 0$. 42

2-3 Rotating analyzer ellipsometer experimental set-up 51

2-4 The rotating analyzer configuration yields various detector signals based on the polarization state of the reflected light. 52

2-5 The surface reaction layer of a growing film is modeled as an equivalent planar-homogeneous film, with index graded smoothly over thickness. 56
2-6 The structure of $\epsilon(z,t)$ for an idealized growing film. 59
3-1 Vertical warm-wall MOCVD apparatus for AlGaAs, illustrating the geometry of sources, reaction chamber, and in-situ ellipsometer during growth. 73
3-2 Elemental sources of a HSMBE system introduce line-of-sight constraints on reactor geometry and ellipsometer configuration (figure courtesy of Gale Petrich). 75
3-3 Classification of the physical structure of solids. 78
3-4 Columns II through VI of the periodic table, highlighting the most common materials used in semiconductor devices (left). Diamond and Zincblende lattice structures characterized by four-nearest-neighbor bonding (right). 78
3-5 The Bond (left) and Band models (right) serve to visualize different features of the lattice. 79
3-6 Lattice constant versus bandgap for a variety of material systems. . . 81
3-7 Desired composition profile for a multiple quantum well laser design. 83
3-8 Synchronization as a performance objective allows one to choose different growth-time profiles d(t) to make difficult maneuvers easier to track. 85
3-9 Different compositions can yield the same dielectric constants due to a "fold" in the map from composition to bulk permittivity. Maps for different wavelengths fold in different places, however, indicating that each composition maps to a unique set of permittivities collected over multiple wavelengths. 89
3-10 Alternate approaches for generating tracking error, where P represents the growth process, E represents the sensor dynamics, and K is the feedback algorithm responsible for delivering the desired performance. 90

3-11 Simulated closed-loop HSMBE growth of a graded InGaAsP structure. 97
3-12 Simulated closed-loop HSMBE growth of a graded structure with multiple quantum wells. 98

4-1 Simulated response with delay (dashed) and with delay and hold (solid).103

Chapter 1

Introduction

This thesis takes a comprehensive view of the issues and impact associated with the use of ellipsometry as an in-situ sensor to control epitaxy in real time. A dynamic model of the physics associated with ellipsometry and growing films is developed from first principles, and this model is used to highlight some of the capabilities and limitations of ellipsometry as a measurement device for feedback control. A control problem is then formulated that substitutes the actual design objective with one more amenable to feedback design, and standard linear tools are used for feedback design.

1.1 Overview

Historically, the use of in-situ monitoring began in 1983 with the introduction of Reflection High Energy Electron Diffraction (RHEED) for Molecular Beam Epitaxy. This sensor was limited to use with processes that operated in a vacuum environment, but had become quite standard within years.

Ellipsometry, an established ex-situ sensor, was introduced for in-situ real-time control of epitaxy in 1990 with the control of AlGaAs by Molecular Beam Epitaxy [15]. In 1992 the same group had successfully used feedback to grow graded AlGaAs structures with the closed-loop growth of parabolic quantum wells [16]. In 1993 and 1994, another group used ellipsometry for the control of HgCdTe using Metal-organic Chemical Vapor Deposition [32, 57]. Feasibility studies such as these continued for

various material systems and from various groups [36, 33, 20], and modeling efforts of the growth processes were expanded [29]. Understanding of ellipsometry as an in-situ characterization tool became a central focus [14, 13, 65, 18, 39].

In 1993 we began work on this problem, resulting in the closed loop growth of AlGaAs by MOCVD two years later [67, 68]. As opposed to earlier works that focused on reporting the best growth performance for a particular system, our emphasis was in understanding how feedback design should be approached for MOCVD, largely independent of the material system or particular sensor technology.

Increasingly, groups began to approach the control of epitaxy by controlling different process parameters individually. Temperature controls on effusion cells, for example, using pyrometer measurements were standard, but tighter control was sought through the use of flux sensors. One group reported flux control of MOCVD in 1996 [27], relating the increased control of source inputs to the desired impact on composition and thickenss. In 1997 a modeling study of the effusion cell and the resulting closed-loop flux control for MBE was conducted [64]. Substrate temperature was another parameter that demanded tighter control [58] . These efforts to use feedback one loop at a time have helped decrease the uncertainty exhibited by different process parameters.

In spite of these successes, truly closed-loop growth, where the growth process itself is in feedback, remains an anomaly instead of being the standard. This thesis uncovers possible reasons, revealing how a relatively simple open-loop process becomes tremendously complex through the sensor dynamics necessary for feedback; designing ad-hoc controllers can be difficult and easily destabilize such a system. Nevertheless, this study suggests that in spite of these dangers, standard linear design techniques can yield reliable performance, even for complicated graded quaternary structures.

1.2 Motivation

Deposition processes are necessary in the manufacture of a variety of products from a variety of industries. Although these processes can be quite varied depending on

the industry or product being served, simplified models of material accumulation and mixing are often general enough to describe many of these processes.

This section provides broad motivation for this research by briefly introducing four dominant deposition industries. Many of the technologies used by these industries may employ the same models and performance criteria as those discussed in this research, thus many of the issues encountered in this study should often be viewed as rather generic to deposition processes in general. Keeping this context in mind, the hope is that lessons learned from piloting epitaxy may be readily applied to the feedback control of many of these applications as well.

1.2.1 Optical Coatings Industry

The oldest of these industries is the optical coatings industry, dating to last century with the preparation of reflective and anti-reflective coatings, and developed by the work of Fraunhofer, Rayleigh, Fabry, Perot, Bunsen, and Faraday, among others. The challenges of this industry revolve around the design, deposition, and characterization of optical coatings [38] that modify the optical properties, e.g. reflectance as a function of wavelength, of a substrate surface. Typical designs include multi-layer coatings of particular materials with precise thicknesses, but the selection of materials and realization of such coatings remains a challenge even today [38]. Addressing these challenges, however, have lead to a rich variety of applications such as anti-reflection coatings (used for lenses of optical instruments, display windows, etc.); multipurpose optical filters; thermal reflectors and cold mirrors used in projectors; dichroic mirrors, consisting of band-pass filters deposited on the faces of prismatic beam splitters to divide light into red, green and blue channels in color television cameras; and highly reflecting dielectric mirrors used in gas lasers and Fabry-Perot interferometers. Although a variety of processes are used for the deposition of such coatings, much of current research uses Metal-organic Chemical Vapor Deposition (MOCVD) or Molecular Beam Epitaxy (MBE) to explore the optical properties of exotic materials such as organic or polymer films, used in organic lasers [11, 63] and display technology [34].

1.2.2 Non-optical Thin Film Technologies Industry

The technology of the optical coatings industry, however, lead to the creation of a similar market for thin film technologies emphasizing non-optical characteristics of surface coatings. These characteristics may include abrasion resistance, hardness, chemical or corrosion resistance, environmental isolation (for hermetic seals, packaging, etc.), low friction or wear, wettability (retaining a lubricating boundary layer), adhesion, biostability (reducing the biological rejection of foreign objects), electrical insulation, and infection resistance (preventing the attachment or migration of biological proteins or bacteria) [23, 38]. Such technologies suggest a variety of applications, including artificial joints and orthopedic devices, catheters, mechanical properties of optical films, stealth coatings on aircraft, viscosity reduction in marine vehicles, or coatings on contact lenses cases or other sterile environments to prevent protein or bacterial contamination. Again, as with the optical coating industry, a variety of deposition processes are used in production. MOCVD, MBE, or their derivatives are used in current research and production of new materials for emerging markets such as diamond films, used as a heat sink in power and micro electronics [45], magnetic films, and super-conducting ceramic films.

1.2.3 Silicon-based Semiconductor Device Industry

While thin-film industries may be the oldest to use deposition processes for production, the silicon-based semiconductor industry is by far the largest market. The industry is dominated by micro-electronic devices such as integrated circuits and dynamic random access memory (DRAM) chips, although other devices such as silicon-based solar cells also use Chemical Vapor Deposition (CVD) techniques in their fabrication. An important distinction of this market from the others discussed so far is that the products here are complex devices, not just stacks of films. Thus, fabrication of these devices is considerably more complicated, and deposition is only one of many processes necessary for production. Another key feature of this industry is its unprecedented economic "jetstream" of production. Fueled primarily by the convenience of using

the native oxide SiO_2 in processing, silicon device technology has experienced rapid advances leading to even greater economies of scale (with the production of smaller and smaller devices). This economic drive has allowed silicon-based devices to win markets from scientifically superior, but more expensive, alternatives [10], and has fueled research into process optimization, including the feedback control of a few of the simpler fabrication processes. The chemistry behind many of these processes, such as plasma etching or CVD, is very complex, leading to difficult *in-situ* measurement and control problems [60, 21, 22]. As a result, management has been reluctant to invest in the development of full automation, particularly since software failure currently plays a significant role in reducing throughput. Thus, run-by-run techniques still dominate industrial practice and command the lion-share of industry-supported research dollars [59].

1.2.4 Compound Semiconductor Device Industry

The youngest deposition industry is the compound-semiconductor device industry, effectively starting around 1978 with the demonstration of GaInAs/InP lasers compatible with optical fiber, and with the development of MOCVD and MBE reactors that removed "the heterostructure technology bottleneck" [53]. This segment represents about 5% of the total semiconductor market, and it is driven by three primary applications. First, the tremendous growth of the mobile communications market, including cell phones and satellite systems, has driven demand for GaAs/AlGaAs based devices such as analogue and digital integrated circuits and high performance transistors. Second, the optical communications industry, driven by the economics of cheap glass fiber, induces a demand for GaInAsP based lasers and detectors. These lasers have the unique capability of lasing at $\lambda = 1.3\mu m$, which is the minimum dispersion wavelength (hence achieving maximum bandwidth), or $\lambda = 1.55\mu m$, which is the wavelength for minimum loss (hence cheapest for telecom) for silica fiber. Finally, the compact disk and optical memory market, in consumer electronics and computer applications, has demanded short wavelength laser diodes for compact data storage. Currently 780nm AlGaAs lasers drive the CD market, but the emerging markets of

635-650nm AlGaInP lasers are important for optical memory using the DVD standard. Even shorter wavelengths, leading to even more compact data storage, are achievable with the new GaN based lasers [48, 50].

1.3 Background

Although the broad deposition market provides some general motivation for this research, the specific incentive for this work is fueled by the compound semiconductor device industry in general, and devices built from III-V heterostructures in particular. These devices use epitaxy as a deposition technology to generate crystalline alloy films that mix various species from columns III and V of the periodic table. Often other fabrication technologies are necessary to produce a given device, but all of these devices demand the ability to grow high quality crystalline films with a specified alloy mixture or material composition.

This section provides background for this research by introducing some of the most common compound semiconductor devices produced by the processes considered here. Of these devices, vertical cavity surface emitting lasers (VCSELs) are explored in more depth to highlight many of the issues relevant to the production of a specific device.

1.3.1 Light Emitting Diodes (LEDs)

At less than 36 cents each [5], LEDs are some of the most common compound semiconductor devices, arising in applications such as automotive displays (e.g. brakelights, running lights, signals, dashboard displays), traffic signals, variable message signs for traffic management, and large area displays for advertising (such as the 170 $\times$ 56 foot full-color screen used for stage effects by the band U2 in various tours [7, 49]). Blue light emitters can be used for spectroscopic detection to analyze blood, monitor oxygen levels, or sniff out biological and toxic agents [43]. Infrared LEDs are used in TV and other remote controls, and have a rapidly expanding market in cordless data transfer in office environments [47]. With the demonstration of high-quality blue LEDs in 1993, attention has focused on mixing red, green, and blue sources to develop

a rugged, low-power white solid state light source [24]. Such a product would not only be invaluable in specialty applications such as safety flashlights, interior lighting in military vehicles, underwater or mining operations, etc., but a potential energy savings of $35 billion/year could be realized by replacing the ordinary household 100 W incandescent bulb [43]. The first volume production of white-light LEDs began in late 1996 [9].

Although the markets for LEDs are substantially different then for lasers, for example, the devices are physically quite similar. LEDs involve spontaneous emission, while lasers invoke the stimulated emission of light. The similarity between LEDs and lasers leads to similar measures of performance, such as power-output (brightness), reliability (lifetime in stressful environments), and efficiency (power-output to power consumption). However, the priority on these measures can be quite different. For LEDs, reliability, including device lifetime, is the key factor. One reason for this is the outdoor applications LEDs support. For example, the mean time between failures for LEDs is typically 1.5 million hours or more, depending on operating conditions, where today's incandescent bulbs burnout after approximately 8000 hours [48]. Another reason reliability is important is due to the stress of LED fabrication. The LED market is extremely cost sensitive, relying on high volume production for economic viability. Thus, fabrication processes such as epoxy encapsulation (for packaging), high speed die attach, and wire bonding are motivated by throughput economics and can be extremely stressful to the device. Thus, reliability is crucial for the devices to survive economical fabrication [24]. Brightness and efficiency are also important, but reliability is often what drives the economics behind a decision to use LEDs in a product instead of alternative light sources.

1.3.2 Transistors

The semiconductor revolution has largely been the story of the development of the silicon transistor. In particular, the incredible speed at which fabrication technology evolved to enable increasingly complex integrated devices is unprecedented. Compound semiconductor alloys can also be exploited to realize transistors and integrated

circuit devices, but the lack of silicon dioxide as a native oxide for these materials has kept the fabrication process from evolving economically as it has for silicon. As a result, compound semiconductor devices leverage unique properties of the alloys to carve niche markets in high-performance industries.

The current market motivation for compound semiconductor transistors and integrated devices is the wireless communications industry, dominated by cell phones and pagers but rapidly expanding to cordless phones, direct broadcast satellite (DBS) TV, wireless local area networks (WLAN), global positioning systems (GPS), and mobile computing applications [42, 48]. The devices used in these applications include silicon bipolar junction transistors (BJTs) and CMOS devices, and compound semiconductor (especially GaAs) MESFETS, HBTs, and HEMTs. Epitaxy is not required for MESFET production, but the market trend is strongly toward HEMTs, and especially HBTs, which require sophisticated epitaxy growth technology [3, 42].

One of the main contests in the wireless communications industry is not between silicon and compound semiconductor devices, as might be expected, but between monolithic microwave integrated circuits (MMICs) vs. discrete devices. Larger companies, with design teams necessary to optimize product designs using discrete components, favor discrete solutions, while smaller companies typically opt for MMIC designs [42, 2]. This issue is important because the distribution of production costs is different for each type of device, thus modulating the impact of high quality epitaxy on the value of the final component. Figure 1-1, reproduced from [52], shows the critical processing steps for HEMT and HBT InP based devices.

Research directions in compound semiconductor electronics are varied and active. One direction is the development of novel devices such as resonant tunnel diodes for memory cells [73], or in improved fabrication processes such as planar processing, an alternative to conventional mesa-etched structures [12, 8]. Another direction is in the integration of different devices on the same chip, such as HBT/HEMT or HBT laser structures [6, 1], or in novel materials such as GaN [46]. One of the most important parts of DARPA's Microwave and Analog Front End Technology (MAFET) Program is the Design Environment Consortium, with goals of developing computer aided

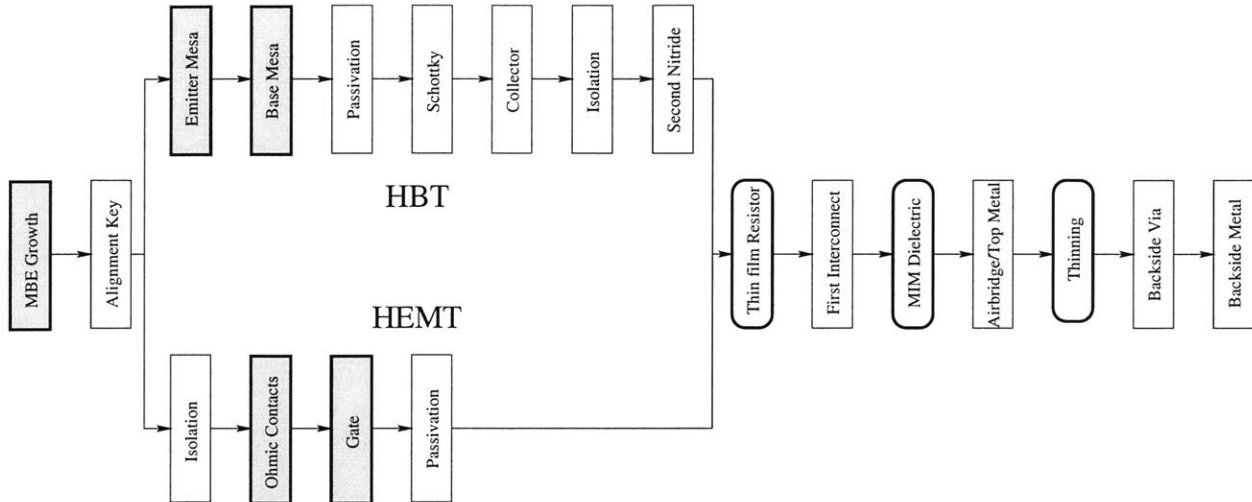


Figure 1-1: Fabrication processes in InP HBT and HEMT MMIC production. Critical nodes are shaded for transistor performance, rounded for passive component performance.

design tool for the virtual prototyping of MMIC and microwave modules [3].

1.3.3 Photodetectors

Photodetectors are devices that convert optical signals into electrical signals [41]. The most basic such device is perhaps the p-i-n photodiode. A diode is a device that allows current to pass in one direction, but restricts the amount of current that can flow when it is "reverse biased." A p-i-n photodiode is a diode with with an undoped "intrinsic" region (see Figure 1-2). The diode is reversed biased so as to deplete the entire intrinsic region of carriers (thereby effectively eliminating current flow through the device). As the surface of the diode is exposed to light, however, hole-electron pairs are generated and whisked away by the strong electric field present in the depletion region. This photo-generated current leads to I-V curves as shown in Figure 1-2.

Not all photons incident at the surface, however, are absorbed. Some, for example, will be reflected at the surface. Two important factors increasing efficiency are a large depletion width and using materials with large absorption coefficient. Increasing the depletion width, though, can significantly increase the carrier transit time, effectively decreasing the bandwidth and speed of the device. So there is a trade-off between

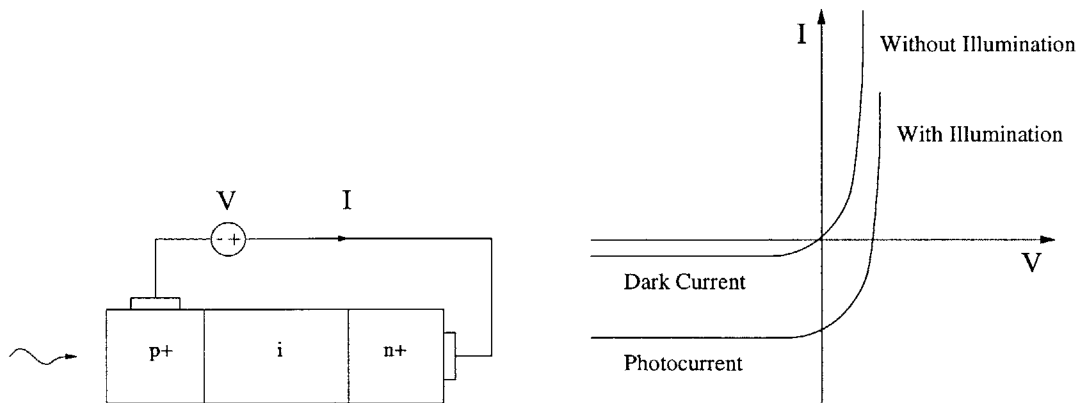


Figure 1-2: Typical reverse biased p-i-n photodiode with corresponding I-V characteristic.

efficiency and bandwidth for a given material system.

In p-i-n photodiodes, each absorbed photon results in the generation of a single hole-electron pair, i.e. the multiplication factor is unity. The result is that, for most applications, the output current must be amplified to usable values. This amplification can introduce significant noise into the signal. Alternative detector structures attempt to reduce this noise by using built-in gain mechanisms to increase the output current, making cascade amplification unnecessary.

An avalanche photodiode is such a device. These devices use an intrinsic instability to increase the multiplication factor. The idea is that a carrier swept through the depletion region will collide with a lattice atom, creating another hole-electron pair. These carriers continue the process, resulting in a carrier "avalanche." Figure 1-3 shows a standard APD design, and a schematic of the fabricated device.

Such complex structures place stringent demands on epitaxy. Moreover, the demands on other processing steps, such as diffusion or implantation, can be elevated with high-quality epitaxy by introducing diffusion barrier layers into the design. Other important fabrication steps include etching, photo-lithography, and electron-beam evaporation.

The market for photodetectors is dominated by silicon devices that do not require epitaxy. However, specialty applications, especially high speed optical networks, has driven an epitaxy-fueled compound semiconductor niche. This segment of the market,

however, is now well established and continues to grow.

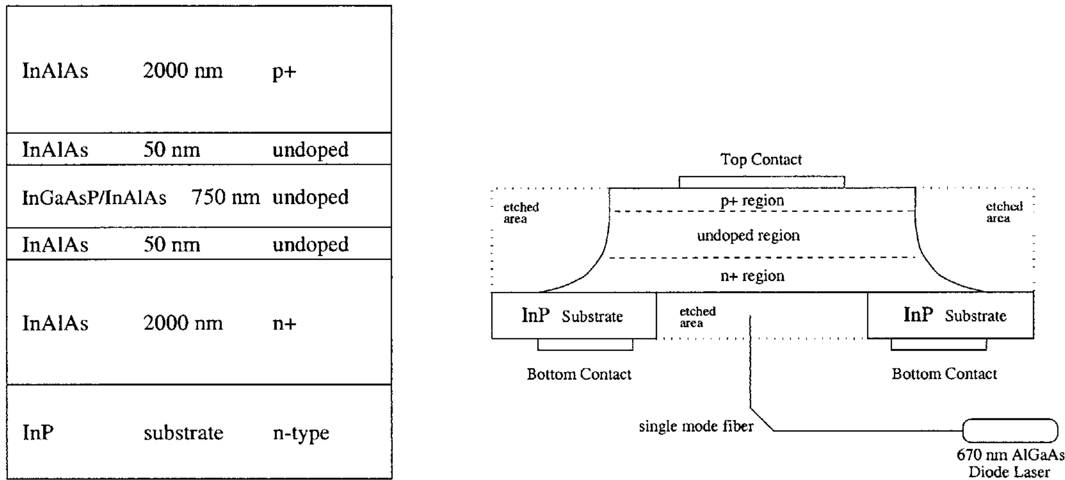


Figure 1-3: Avalanche photodiode design, and a schematic of the resulting fabricated device.

1.3.4 Non-Optical Sensors

Although not as large an industry as other devices, non-optical sensors are an important class of semiconductor devices because they often are enabling technologies for new or emerging applications. Many of these devices are still primarily manufactured from silicon, but there is a small but steady trend toward developing the fabrication technology to make compound semiconductor alternatives economically viable.

One important such device is the Hall-effect sensor, or Hall element. These devices are used primarily as high-precision magnetic sensors for controlled rotation of motors in VCRs, PCs, CD-ROMs, and other electronic products [72]. The idea behind these devices is that semiconductors exhibit a transverse voltage due to the application of a longitudinal current and a perpendicular magnetic field. This phenomenon, called the Hall effect, can be exploited as a magnetic sensor by passing a (constant) current through a small semiconductor "chip". The resulting voltage seen by the Hall element is then proportional to the magnetic field orthogonal to the "chip" surface. Current devices use very low tech processing and do not require epitaxy. The next generation of devices, however, are building on multi-wafer MBE growth

to deliver higher performance products that incorporate novel structures, such as quantum wells, that require precision growth [4].

Another growing market is in the area of high-temperature electronics, including transistors and sensors (even some Hall-effect sensors). [51]. Although silicon is, and is projected to remain, the dominant technology for most of these devices for at least ten years, this market is natural for the wider bandgap compound semiconductors due to their greater tolerance for thermal excitation. Development of these devices is currently a spin-off application of more immediately promising markets such as radiation hardening, power handling, and blue-light emission. 40% of the current market consists of data acquisition electronics monitoring the environment around drilling heads for the oil industry. The most rapidly expanding market for these devices, though, is in reduced-noise rugged sensors for the aerospace and automotive industries. Other applications include industrial process control and nuclear power.

1.3.5 Solar Cells

Solar cells [44, 54] are an important class of detectors that convert sunlight to electricity, providing a clean, inexhaustible energy resource. The simplest structure is single junction solar cell, which consists of a simple large-area pn junction. When the junction is exposed to light, it will efficiently separate photo-excited hole-electron pairs, generating electrical current when the cell is connected to a load.

More advanced designs are emerging, though, which stack multiple junctions using different bandgap materials (Figure 1-4). The idea is that the top junction absorbs and converts the high energy part of the spectrum while transmitting lower energy photons to the underlying cells. Optically buries cells absorb and convert increasingly longer wavelength, lower energy photons, thus increasing the overall efficiency (power per unit area) of the design.

Currently the space-satellite market is driving the demand for solar cells since photovoltaics are their preeminent energy source. Compound semiconductor cells are especially in demand, even though they are more than five times more expensive than silicon cells, because their high-efficiency balances the cost of placing larger,

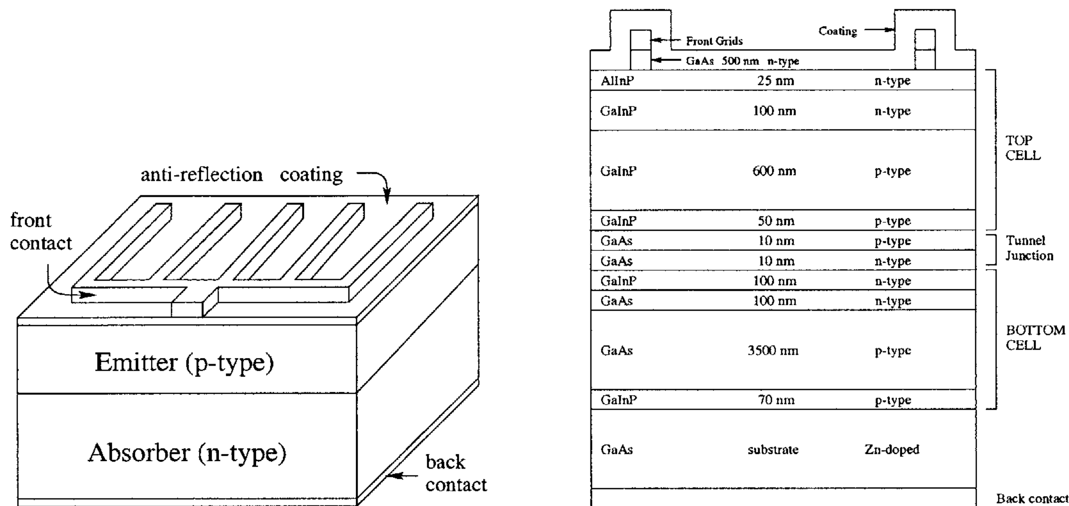


Figure 1-4: Schematic of a single junction solar cell (left), and the cross-section of a multi-junction cell developed at NREL (right).

heavier, silicon arrays in orbit. Another advantage is that compound semiconductor cells can be "radiation hardened," leading to higher reliability in high radiation environments. Other applications include water pumping, cathodic protection, communications, lighting and small appliances, and power sources for remote villages or electric vehicles.

1.3.6 Optical Filters

Optical filters [25] are devices that tap or modulate particular frequencies in optical signals. They are important in high-bandwidth optical networks because they allow a high-bandwidth optical signal to be separated into single channels. This facilitates wavelength division multiplexing (WDM) to accommodate varied data rates and expand user access to the full bandwidth of the system.

Two types of filters are the full spectral resolver and channel dropping filter. The full spectral resolver acts like a prism to separate channels by wavelength, and can be realized by a prism, grating, or phased-array. A channel-dropping filter is a device that selectively taps a single channel from the optical signal while leaving the others undisturbed. Figure 1-5 shows a typical optical data link and a schematic of a channel-dropping filter.

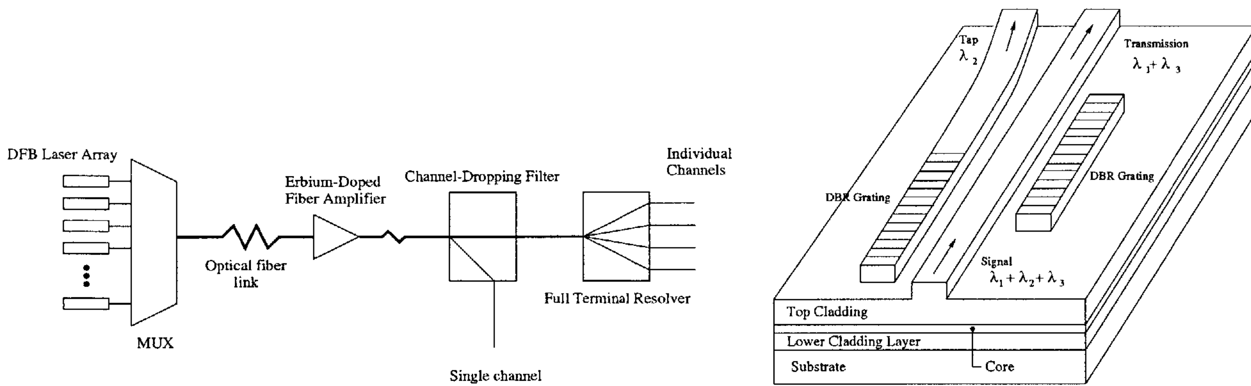


Figure 1-5: Typical optical data link (left) and a schematic of a channel dropping filter (right).

The physics behind a channel dropping filter are similar to those of a piano tuning fork. The idea is to place an optical resonators near the center waveguide. Just as the tuning fork resonates with a particular note, the optical filter resonates with a particular optical frequency. In this way, a single channel can be read or changed without affecting the others. These devices are difficult to fabricate, though, pushing lithographic and epitaxial limits for precision. Research toward developing them is thus ongoing.

1.3.7 Lasers

The possibility of stimulated emission of radiation was first suggested by Einstein. It was not until 1960, however, that laser action was demonstrated in optical wavelengths, with the development of the ruby laser by Maiman. In 1961 Bell Laboratories invented the He-Ne laser, and in 1962 the first semiconductor laser (or laser diode) was reported. Since then the market for semiconductor lasers has surpassed that of any other laser source, owing mostly to the optical communication and compact disk technology that they enable. Other applications are varied, though, ranging from mammography and photochemistry to smart pixels and optical interconnects.

To understand the action of lasers, consider a system where free electrons are confined to two states, as shown in figure 1-6. The number of electrons in the lower energy state is typically greater than that of the higher energy state. Thermal equilibrium

implies that transitions between the states are balanced, in the sense that the number of 'up' transitions equal the number of 'down' transitions. A variety of mechanisms can cause such transitions, but when radiative transfer is involved an 'up' transition is called photon absorption and a 'down' transition (in thermal equilibrium) is called *spontaneous* emission.

Suppose a photon with energy equal to the gap is shone on the material. Now either the photon combines with an lower-energy electron to produce an 'up' transition (absorption), or it stimulates the emission of another photon by inducing the transition of a higher-energy electron (*stimulated* emission). It turns out that these transitions are equally likely. However, since there are many more lower-energy electrons, shining a light on a two-state system results in a net absorption at the wavelength of the gap (thus defining the material's visible color).

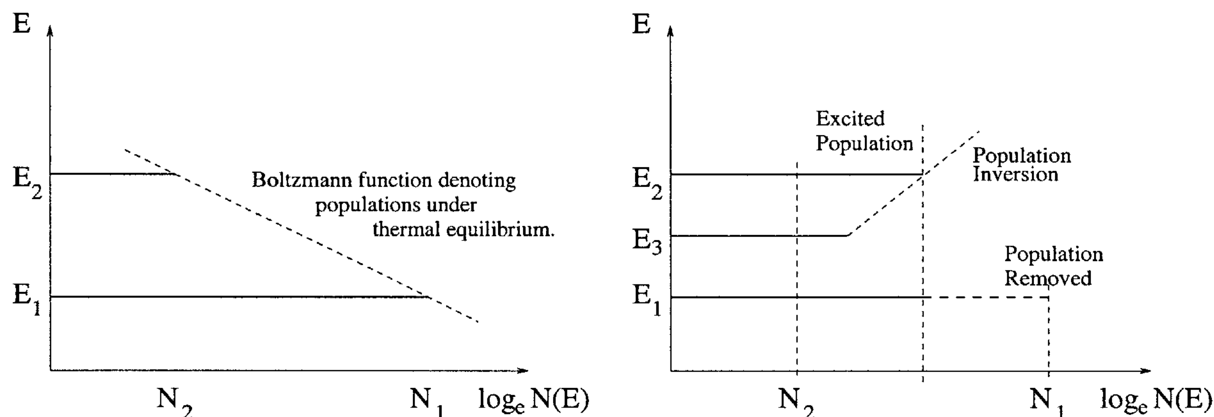


Figure 1-6: Semiconductor lasers use population inversion to induce conditions for spontaneous emission.

Semiconductor lasers typically use a forward biased p-n junction to create a "population inversion," i.e. conditions where there are more high-energy electrons than lower-energy electrons (relative to a third state–see Figure 1-6). This implies that there are lots of electrons in the conduction band and holes in the valence band waiting to recombine. Now, an incident photon of the bandgap energy will tend to induce stimulated emission (rather than absorption), resulting in a net gain of the number of photons that is proportional to the volume of active material (or material under the population inversion conditions).

These photons are confined to the active layer by surrounding it with reflective surfaces, creating an optical cavity that induces standing waves of the stimulated light. No reflective surface is perfect, though, and so there will be some losses unless enough current is applied (the "threshold current") to compensate for these losses and maintain the standing wave. Any additional current applied to the system will result in emission of stimulated light from the reflective surfaces.

A number of structures have been devised to induce lasing action in semiconductor devices. Two basic categories of lasers are edge emitters and surface emitters (see Figure 1-7). While edge emitters, due to their high power potential, are the primary source for telecommunication applications, surface emitters, and especially vertical cavity surface emitting lasers (VCSELs), are receiving increasing attention for a number of reasons. First, they can be conveniently packaged into two-dimensional arrays for applications such as interconnects, data transmission, printing, or as optical pumps for solid state lasers. Second, their circular, low divergence output beams allow efficient coupling to fiber. Third, small active volumes lead to low threshold currents and low power consumption. Fourth, simple monolithic manufacturing allow for economical, high volume/density production with on-wafer testing. This last reason is important from the perspective of this study because it highlights the way epitaxial complexity is used to replace fabrication complexity, by eliminating the need, for example, for precisely cleaved surfaces. This elimination of processing steps places epitaxy in a more central role to impact device performance, suggesting that the quality of such devices will be more sensitive the quality of the epitaxy process.

1.3.8 Vertical Cavity Surface Emitting Lasers

Vertical cavity surface emitting lasers (VCSELs) have optical cavities aligned with the (vertical) direction of epitaxial growth (see Figure 1-7). This simple change in cavity orientation produces radical differences in the beam characteristics, scalability, optoelectronic design, fabrication, and array configurability. In particular, this configuration has the advantage of selecting the active layer to be thin enough to support only a single optical mode, thereby enhancing the quality of light emitted.

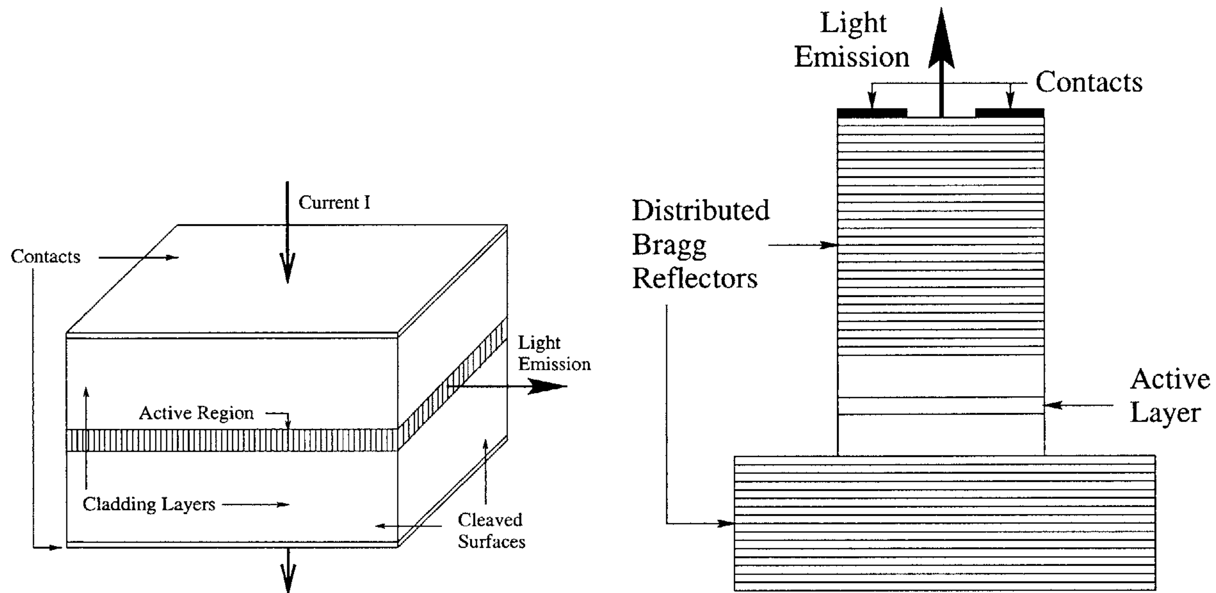


Figure 1-7: Edge emitting (left) and vertical cavity surface emitting lasers (right).

Also, since the active layer volume can be very small, low threshold current and power consumption are typical. The emitted light is typically circular, and thus can be more efficiently coupled to fiber. Finally, since the mirrors can be fabricated as part of the epitaxy processes, epitaxial complexity replaces fabrication complexity, in that cleaving or other complicated processing steps may not be necessary. This monolithic processing supports on-wafer testing, which can greatly enhance throughput and quality control measures in production. In short, VCSELs are the most manufacturing-friendly laser on the market provided their stringent demands on epitaxial growth can be adequately met.

How VCSELs Work

The basic VCSEL structure consists of a central optical cavity surrounded by two highly reflective distributed Bragg reflector (DBR) mirrors. These mirrors ensure vertical optical confinement, and they are usually composed of many periods of repeated layers of low and high refractive index materials. In addition, these mirrors must be transparent at the laser wavelength, and they should permit current injection with the least possible resistance. The electrical properties of the mirrors are usually achieved with appropriate doping profiles throughout the structure. Also the mirror

heterojunctions can be modified with compositional grading or short-period chirped superlattices to achieve the right optical and electrical properties.

The optical cavity is usually one wavelength thick to ensure single mode operation. It contains an active layer consisting of one or more quantum wells formed in a material corresponding to the right wavelength regime. There a number of ways to achieve lateral electrical and optical confinement. Figure 1-7 suggests an etched air-post geometry, but ion implantation and other methods can also be used for isolation. Smaller active volumes lead to lower threshold currents, which can be an important design criterion for these devices.

Design Criterion

Although the intended application usually determines the priority given to various design criterion, VCSELs share a number of measures quantifying the performance of a given structure. These include lasing wavelength, and pulsed and continuous L-I characteristics, pulsed and continuous I-V characteristics.

The first measure, lasing wavelength, is a function of the materials used in the active layer, and its resulting thickness-composition profile. Uniformity patterns over a wafer governs how this characteristic changes spatially, but these patterns can be detected using reflectance measurements before the wafer is segmented into individual devices. Wafer rotation during growth typically leads to radially symmetric wafer uniformity within 1% on modern MBE or MOCVD reactors. Generally a number of thickness/composition profiles in the active layer yield a specified lasing wavelength.

The L-I characteristic shows light output, in watts, as a function of input current. A typical curve shows little output below a threshold value of the current, and an approximately linear relationship thereafter until a maximum current and output power is reached. The power drops off rapidly as the current is increased above this maximum. Desirable features of this characteristic are a low threshold current, obtained from a small active volume, high efficiency (steep slope in the linear regime), resulting from good alignment between the lasing wavelength and the optimal optical characteristics of the (high quality) mirrors, and high maximum power.

Realizing a high maximum power requires good electrical characteristics as revealed by the I-V curve. In particular, the I-V characteristic shows how resistance of the bragg reflectors begins to limit current density as the bias voltage is increased. At some point, thermal runaway, or a rapid decrease in device resistance, is observed, indicating the maximum achievable current density. The maximum current, and thus power output, can then be predicted from the device size and its efficiency. Greater power outputs are achievable from larger devices (clearly a trade-off with low threshold current) with good electrical properties. These properties are determined by the composition and doping profiles in the mirrors.

Although these measures are the most dominant for device design, other features can also play an important role in measuring performance for certain applications. These include output lasing spectra at different pulse currents and far-field pattern (showing beam diameter and coherency).

Fabrication Processes

Much of what determines performance in a VCSEL relies on the ability to grow a specified compositional profile with accuracy. For example, efficiency of the device is tied to the optical properties of the mirrors, thus the ability to grow multiple periods of identical layer-pairs is critical. Likewise, lasing wavelength is determined by a precise profile in the active region. This is where feedback can possibly make a significant contribution.

Other processes are also important in a working device, however. For example, the doping profile of the mirrors is just as critical as the composition profile for good electrical properties. In mesa etched devices, the etching process is important to define the device's lateral dimensions and hence impacts the threshold current and power output. Ion-implantation plays a similar role for implantation-isolated devices. In some cases, epitaxial regrowth is desired to yield a planar device following mesa etch isolation. The regrowth dynamics can differ suitably from nominal growth, and hence demands consideration as an independent process. Finally, electrical contacts must be attached. Even with these other processing steps, though, epitaxy is the

performance-critical stage of processing for VCSELs, and VCSEL designs pose some of the most demanding requirements for epitaxial growth.

Applications

VCSELs have a different set of capabilities that tend to complement, rather than compete with, those of edge emitting lasers. Edge emitters will dominate high power applications, such as telecom, but VCSELs are the technology of choice for optical interconnects. These connections are used to transmit high bandwidth data over short distances (e.g. less than 1km) with immunity from electromagnetic interference. The ultimate goal of free-space optical interconnect research is a "smart pixel," or a single-chip array of optical input and output devices with controlling logic to guide a lens array that collimates and routes the light as it passes through the air. A similar structure has been proposed for integrated optical neural processors that use the artificial neural-network model as a paradigm for optical computing.

VCSELs' near monolithic processing also make them manufacturing-friendly companions for other devices on a single chip. The integration of VCSELs and microelectronics is a current research trend, and Sandia National Laboratories has recently demonstrated the integration of VCSELs with microelectromechanical (MEMs) technology. Other research directions include the development of blue-emitting VCSELs and polarization control of the light emitted from VCSELs. All of this research represents emerging markets for this technology, while optical interconnects continue to drive the current demand for these devices.

Chapter 2

Ellipsometry as a Sensor Technology

Ellipsometry is a well established optical technique for non-invasively characterizing a specified material. The technique consists of measuring the change in the polarization state of light that interacts with the material in a carefully designed experimental arrangement.

Although there are various experimental arrangements and hardware configurations used to perform these measurements, at the heart of each technique is the optical response of a material to an incident beam of light. This response is characterized by the way the material reflects, and transmits, a beam of light impinging at a specified angle of incidence.

This chapter analyzes ellipsometry as a sensor technology by first reviewing the relevant background physics and deriving the theoretical optical response of stationary stratified materials. A dynamic model of ellipsometry is then developed from assumptions that link the optical response of a growing film to that of a stationary stratified material. Finally, this model is analyzed to understand its behavior as a sensor technology in feedback applications.

2.1 Background Physics

2.1.1 Electromagnetic Fields and Maxwell's Equations

Our interest is in the behavior of four vector fields throughout a region of space in $\mathbb{R}^3$ [19, 31, 35]. The vector fields of interest are the electromagnetic fields:

$\boldsymbol{E}$=electric field strength	(volts per meter),
$\boldsymbol{D}$=electric flux density	(coulombs per square meter),
$\boldsymbol{H}$=magnetic field strength	(amperes per meter),
$\boldsymbol{B}$ = magnetic flux density	(webers per square meter) or (teslas).

These are functions of time and space, mapping $\mathbb{R}^4 \rightarrow \mathbb{R}^3$, and they are discussed here in SI units. Note that our general notation will use a slanted font to indicate real-valued quantities (versus an upright font for complex-valued quantities, e.g. E vs. E), and a heavy-weight font for vector-valued functions (versus normal-weight text for scalar-valued quantities, e.g. $\boldsymbol{E}$ vs. E). The exception will be Greek symbols, where normal weight indicates real-valued quantities, while heavy-weight denotes complex-valued quantities (e.g. ϵ vs. $\boldsymbol{\epsilon}$).

Maxwell's equations describe the behavior of these fields in the presence of electromagnetic excitation:

$$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t} \tag{2.1}$$

$$\nabla \times \boldsymbol{H} = \boldsymbol{J} + \frac{\partial \boldsymbol{D}}{\partial t} \tag{2.2}$$

$$\nabla \cdot \boldsymbol{B} = 0 \tag{2.3}$$

$$\nabla \cdot \boldsymbol{D} = \rho_v \tag{2.4}$$

where the nature of the electromagnetic source excitation is characterized by

$\boldsymbol{J}$ = electric current density	(amperes per square meter)
ρ_v = electric charge density	(coulombs per cubic meter)

as maps $\mathbb{R}^4 \rightarrow \mathbb{R}^3$ and $\mathbb{R}^4 \rightarrow \mathbb{R}$. These equations thus determine the behavior of $\boldsymbol{E}$, $\boldsymbol{D}$, $\boldsymbol{H}$, and $\boldsymbol{B}$, at every point in space and at all times, as driven by an admissible source

excitation ($\boldsymbol{J}$, ρ). Note that the sources $\boldsymbol{J}$ and ρ are not independent; Maxwell's equations also proscribe that an admissible source pair satisfy the conservation of charge relation

$$\nabla \cdot \boldsymbol{J} + \frac{\partial \rho_v}{\partial t} = 0.$$

Moreover, the lack of a magnetic charge variable in equation 2.3, corresponding to ρ_v in equation 2.4, can also be considered a constraint on an admissible source; admissible sources do not have free magnetic charge.

2.1.2 Constitutive Relations

Even with admissible excitation, however, Maxwell's equations are underdetermined. That is, given an admissible source that respects conservation of charge, there are still many combinations of $\boldsymbol{E}$, $\boldsymbol{D}$, $\boldsymbol{H}$, and $\boldsymbol{B}$ fields that will satisfy Maxwell's equations driven by the source. In particular, given an admissible source, equations 2.3 and 2.4 can be derived from equations 2.1 and 2.2. This indicates that Maxwell's equations provide six independent scalar relations on twelve independent scalar variables.

To allow Maxwell's equations to uniquely characterize these vector fields, we must provide additional information specifying which of the valid solutions is appropriate for a particular point in space and time. This information is provided as six additional independent constraints on the fields. Such constraints are called constitutive relations, and they provide a mechanism for classifying various media by their electromagnetic properties.

Let $\mathcal{C}$ be the 6×6 matrix operator relating the fields as

$$\begin{bmatrix} \boldsymbol{D} \\ \boldsymbol{B} \end{bmatrix} = \mathcal{C} \begin{bmatrix} \boldsymbol{E} \\ \boldsymbol{H} \end{bmatrix}.$$

We typically decompose $\mathcal{C}$ into the 3×3 operators $\bar{\epsilon}$ (permittivity tensor), $\bar{\xi}$, $\bar{\zeta}$, and

$\bar{\mu}$ (permeability tensor), where

$$\begin{bmatrix} \boldsymbol{D} \\ \boldsymbol{B} \end{bmatrix} = \left[\begin{array}{c|c} \bar{\epsilon} & \bar{\xi} \\ \hline \bar{\zeta} & \bar{\mu} \end{array}\right] \begin{bmatrix} \boldsymbol{E} \\ \boldsymbol{H} \end{bmatrix}. \tag{2.5}$$

The structure and functional dependencies of $\mathcal{C}$ provide a system to classify various media by their electromagnetic properties. Bi-anisotropic media are general media with non-zero parameter tensors, while anisotropic materials are characterized by a block-diagonal structure, with $\bar{\xi} = \bar{\zeta} = 0$. Isotropic materials are a special case of anisotropic media, where not only $\bar{\xi} = \bar{\zeta} = 0$, but also $\bar{\epsilon}$ and $\bar{\mu}$ have the special form

$$\bar{\epsilon} = \epsilon \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \bar{\mu} = \mu \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

with ϵ and μ being scalar-valued quantities.

In addition to being classified by the block-structure of $\mathcal{C}$, a material may be further sub-classified by the particular functional dependencies of its constitutive parameters. For example, we call a medium homogeneous if $\mathcal{C}$ is independent of spatial coordinates, stationary if it is independent of time, time-dispersive if it a function of time derivatives, spatially dispersive if a function of spatial derivatives, and linear if it operates linearly on the electromagnetic field vectors. The simplest isotropic medium is free space or vacuum, a homogeneous, stationary, non-dispersive, linear medium characterized by $\epsilon = \epsilon_o \equiv 8.85 \times 10^{-12}$F/m and $\mu = \mu_o \equiv 4\pi \times 10^{-7}$H/m.

Equipped with constitutive relations that model the electromagnetic properties of the physical system of interest, one can solve Maxwell's equations for the field vectors given an admissible source and/or appropriate boundary conditions.

2.1.3 Uniform Plane Waves

Equipped with appropriate constitutive relations, the problem of solving Maxwell's equations is well-posed when appropriate initial and boundary conditions are speci-

fied. Consider a simple isotropic homogeneous medium that fills the region $\mathcal{R} \subset \mathbb{R}^3$, and suppose no sources exist in $\mathcal{R}$. Maxwell's Equations then give us

$$\nabla \times \boldsymbol{E} = -\mu \frac{\partial \boldsymbol{H}}{\partial t}, \tag{2.6}$$

$$\nabla \times \boldsymbol{H} = \epsilon \frac{\partial \boldsymbol{E}}{\partial t}. \tag{2.7}$$

Taking the curl of (2.6), substituting (2.7), and making use of the vector identity $\nabla \times (\nabla \times \boldsymbol{E}) = \nabla (\nabla \cdot \boldsymbol{E}) - \nabla^2 \boldsymbol{E}$ yields the wave equation

$$\nabla^2 \boldsymbol{E} - \epsilon\mu \frac{\partial^2 \boldsymbol{E}}{\partial t^2} = 0. \tag{2.8}$$

A similar expression could be found for $\boldsymbol{H}$. Consider the following solution to the wave equation

$$\boldsymbol{E} = \text{real}\left\{E_o \exp^{j\omega t - jkz}\right\} \hat{\mathbf{x}} \tag{2.9}$$

where $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ refer to the standard basis vectors in $\mathbb{R}^3$. Substituting into (2.8), we find

$$-k^2 E_o \exp^{j\omega t - jkz} + \omega^2 \epsilon\mu E_o \exp^{j\omega t - jkz} = 0,$$

which is a solution throughout $\mathcal{R}$ provided $k^2 = \omega^2\epsilon\mu$. This is a particular wave propagating in the $+\hat{\mathbf{z}}$ direction; we could have easily chosen a wave propagating in a different direction or with very different properties and found that it is also a solution to the wave equation given above for this simple medium.

Although there are many kinds of wave phenomenon that are consistent with the wave equation derived from Maxwell's Equations, the solution that exists in our region of interest, $\mathcal{R}$, is determined by the particular boundary conditions we impose. The simplest of these wave phenomenon is the solution described above (2.9). This type of wave is called a uniform plane wave because the amplitude of the electric field at a point (x, y, z) at a time t is equal to the amplitude of the electric field at any other point $(\tilde{x}, \tilde{y}, \tilde{z})$ in a plane (in this case, the plane is given by $\tilde{z} = z$) at the same time. Other wave types include cylindrical or spherical waves. Often, instead of defining

the boundary conditions that give rise to the wave phenomenon most descriptive of the physical situation, one simply defines the wave that is deemed sufficiently descriptive to model the physical situation and specifies an initial condition for the electromagnetic field throughout the region $\mathcal{R}$ at time $t = 0$.

Focusing our attention on the uniform plane wave, we note that we have chosen a harmonic wave since the harmonic functions are the eigenfunctions of the linear wave operator. In general we could consider other plane-wave variations, but the analysis would proceed by decomposing the general variation into its Fourier basis and considering each harmonic component separately. Thus, we consider harmonic excitation without loss of generality. When all fields in the region vary harmonically, it is convenient to operate in the time-transform domain and use standard phasor notion to represent the field amplitudes. In such a case, we may consider complex-valued parameters in the constitutive relations, such as a complex permittivity $\boldsymbol{\epsilon}$, and simplify the representation of dispersive or conductive effects on the fields by the medium.

Nevertheless, the uniform plane wave described above is not in its most general form. For a wave traveling in the $+\hat{\mathbf{z}}$ direction, we would more generally have the field

$$\mathbf{E} = \begin{bmatrix} \mathbf{E}_x \exp^{-jkz} \\ \mathbf{E}_y \exp^{-jkz} \\ 0 \end{bmatrix} \tag{2.10}$$

where the vector components are understood to coincide with the (x, y, z) system, and the time dependence $\exp^{j\omega t}$ has been suppressed. This wave is seen to still be a uniform plane wave traveling in the $+\hat{\mathbf{z}}$ direction, nevertheless it differs from (2.9) in that the electric field vector is rotating in the x-y plane as a function of t and z. The simpler solution from (2.9) is seen to always point in a fixed direction (in this case, the $\hat{\mathbf{x}}$ direction) and hence is called linearly polarized. The general expression above, (2.10), is seen to trace out an ellipse in the x-y plane, depending on the relative phases of $\mathbf{E}_x$ and $\mathbf{E}_y$. This is the most general expression for a uniform harmonic plane wave, and it is important to note that it can always be decomposed into the

superposition of two linearly polarized waves. Substituting this expression into (2.1) reveals that

$$\mathbf{H} = \begin{bmatrix} \sqrt{\frac{\epsilon}{\mu}}\mathrm{E}_y \exp^{-jkz} \\ -\sqrt{\frac{\epsilon}{\mu}}\mathrm{E}_x \exp^{-jkz} \\ 0 \end{bmatrix}, \tag{2.11}$$

which shows how a generally polarized electric field is related to its corresponding magnetic field. We observe that the electric and magnetic fields are indeed mutually orthogonal to each other and to the direction of propagation (Figure 2-1).

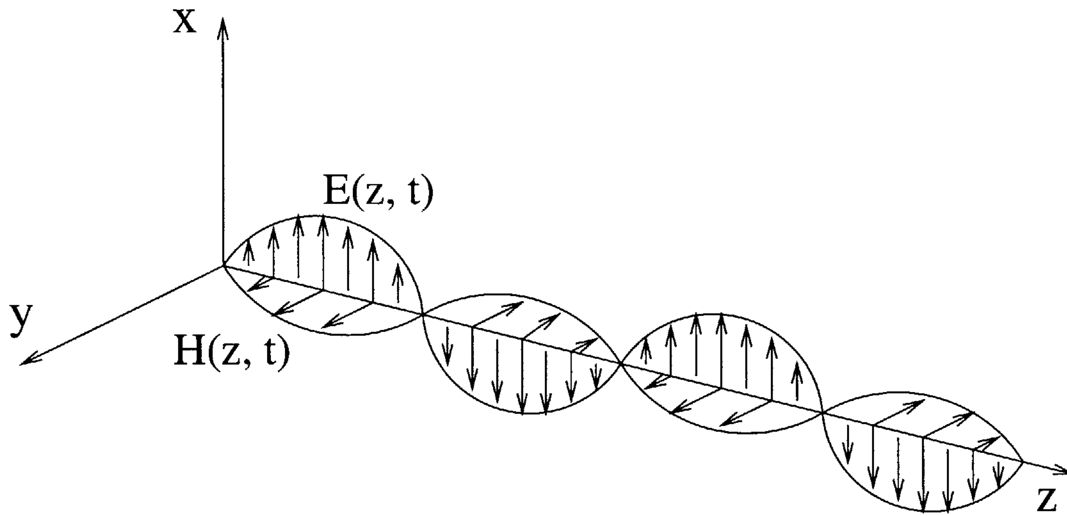


Figure 2-1: A linearly polarized wave oscillates without twisting around its axis of propagation.

2.1.4 Boundary Conditions and Inhomogeneous Media

While uniform plane waves may be descriptive of field excitation in homogeneous media, it is often the effect of inhomogeneous media on the fields that are of most interest. In particular, one may wonder how to "solve" Maxwell's Equations at points where the values of the constitutive parameters are discontinuous, such as at the interface between two semi-infinite homogeneous media.

This situation is handled by using the integral form of Maxwell's Equations to derive boundary conditions at the interface. Thus, the region of interest $\mathcal{R}$ is effectively split into two regions, with solutions related by the boundary condition at the

interface.

For stationary boundaries, it is straightforward to show that the following conditions must hold

$$\begin{aligned} \hat{n} \times (\mathbf{E}_1 - \mathbf{E}_2) &= 0 \\ \hat{n} \times (\mathbf{H}_1 - \mathbf{H}_2) &= \mathbf{J}_s \end{aligned} \tag{2.12}$$

where $\hat{n}$ is the direction normal to the boundary surface, J_s is the complex amplitude of the surface current (if any), and the subscripts 1 and 2 refer to the media on either side of the interface. These conditions simply require that the tangential components of the electric field be continuous through the interface, and that any mismatch in the tangential components of the magnetic field be offset by the presence of a surface current.

Imposing these conditions at a planar interface between two simple, semi-infinite media demonstrates that an incident wave impinging from the first medium generates a reflected wave in the first medium, and a transmitted wave in the second. This is easily seen from the imposition of the boundary condition. Let the first medium be characterized by (ϵ_1, μ_1), for the region $z < 0$, and the second medium be given by (ϵ_2, μ_2), for the region $z > 0$. At time t=0, let the electromagnetic field in Medium 1 be given by

$$\mathbf{E}_i = \begin{bmatrix} \mathrm{E}_0 \exp^{-jk_1 z} \\ 0 \\ 0 \end{bmatrix} \quad \mathbf{H}_i = \begin{bmatrix} 0 \\ \sqrt{\frac{\epsilon_1}{\mu_1}} \mathrm{E}_0 \exp^{-jk_1 z} \\ 0 \end{bmatrix}$$

where $k_1^2 = \omega^2 \epsilon_1 \mu_1$, and let the fields in Medium 2 be identically zero. Since the electric and magnetic fields are tangential to the interface, and assuming no surface current is present, the boundary conditions imply that a similar field exists in Medium 2 for all time $t > 0$, given by

$$\mathbf{E}_2 = \begin{bmatrix} \mathrm{E}_2 \exp^{-jk_2 z} \\ 0 \\ 0 \end{bmatrix} \quad \mathbf{H}_2 = \begin{bmatrix} 0 \\ \mathrm{H}_2 \exp^{-jk_2 z} \\ 0 \end{bmatrix}$$

where $\mathrm{E}_2 = \mathrm{E}_0$ and $\mathrm{H}_2 = \sqrt{\frac{\epsilon_1}{\mu_1}}\mathrm{E}_0$. Solving Maxwell's equations in Medium 2 fixes $k_2^2 = \omega^2 \epsilon_2 \mu_2$, but also creates a contradiction by requiring that $\mathrm{H}_2 = \sqrt{\frac{\epsilon_2}{\mu_2}}\mathrm{E}_0$. This inconsistency is resolved, though, by the presence of a reflected wave in Medium 1. Now the boundary conditions imply

$$\mathrm{E}_0 + \mathrm{E}_R = \mathrm{E}_2$$

$$\sqrt{\frac{\epsilon_1}{\mu_1}}\mathrm{E}_0 - \sqrt{\frac{\epsilon_1}{\mu_1}}\mathrm{E}_R = \sqrt{\frac{\epsilon_2}{\mu_2}}\mathrm{E}_2$$

and, solving for E_R and E_2, we find $\mathrm{E}_R = \mathrm{r}\mathrm{E}_0$ and $\mathrm{E}_2 = \mathrm{t}_{\mathrm{trans}}\mathrm{E}_0$, where

$$\mathrm{r} = \frac{\sqrt{\frac{\epsilon_1}{\mu_1}} - \sqrt{\frac{\epsilon_2}{\mu_2}}}{\sqrt{\frac{\epsilon_1}{\mu_1}} + \sqrt{\frac{\epsilon_2}{\mu_2}}}$$

$$\mathrm{t}_{\mathrm{trans}} = \frac{2\sqrt{\frac{\epsilon_1}{\mu_1}}}{\sqrt{\frac{\epsilon_1}{\mu_1}} + \sqrt{\frac{\epsilon_2}{\mu_2}}}.$$

are defined as the reflectance and transmittance at the interface. In general we expect that when an inhomogeneous medium is excited by an incident wave, a reflected and transmitted wave will result, characterized by a reflectance or transmittance function, respectively.

2.1.5 Optical Response of Plane-Stratified, Stationary Films

To derive the field effects of more complicated inhomogeneities, we consider a complex medium embedded between simple, homogeneous media [30, 19, 37, 66, 26]. Let the complex medium be defined in the region $0 < z < d$, with d fixed, and let the region $z > d$ be the *ambient*, characterized by (ϵ_a, μ_o), and the region $z < 0$ be the *substrate*, characterized by (ϵ_s, μ_o), where μ_o is the permeability of free space. We will restrict the film to be a linear stationary medium with inhomogeneity only in the $\hat{\mathbf{z}}$ direction.

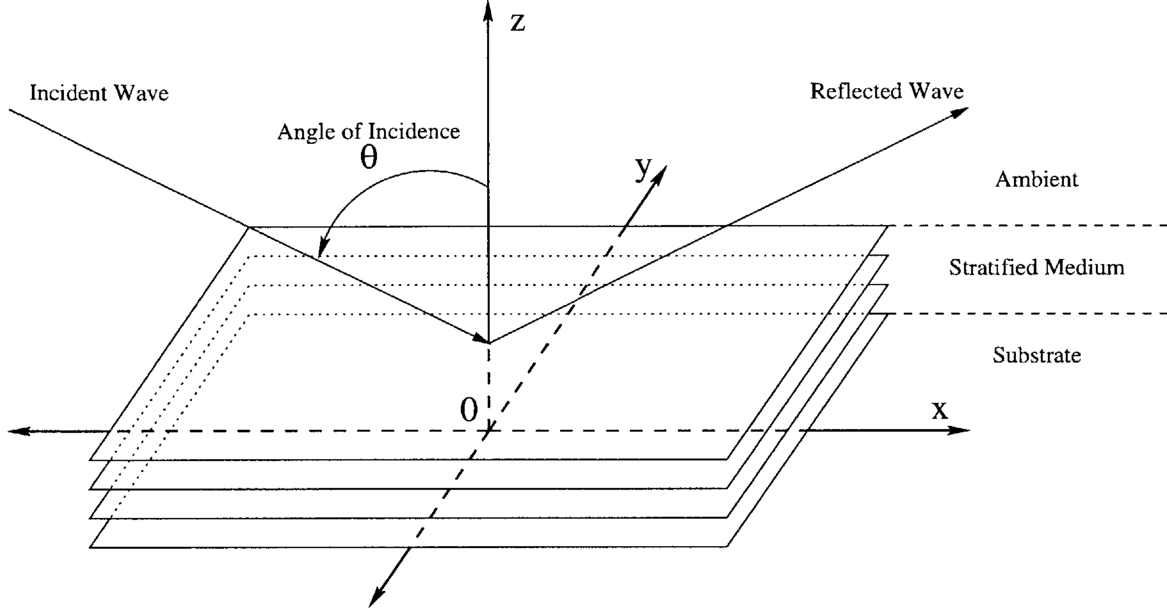


Figure 2-2: Geometry of the physical system defining the optical response of a stationary stratified film with $\phi = 0$.

Otherwise, let the constitutive parameters be arbitrary, given by

$$\begin{bmatrix} \mathbf{D} \\ \mathbf{B} \end{bmatrix} = \left[\begin{array}{c|c} \bar{\epsilon}(z) & \bar{\xi}(z) \\ \hline \bar{\zeta}(z) & \bar{\mu}(z) \end{array}\right] \begin{bmatrix} \mathbf{E} \\ \mathbf{H} \end{bmatrix}. \tag{2.13}$$

Initial conditions are established by a uniform plane wave with time dependence $\exp^{j\omega t}$ impinging obliquely from the ambient at an angle θ with the surface normal, and an angle ϕ with the $\hat{\mathbf{x}}$ axis; all other fields throughout the system are zero before $t = 0$ (Figure 2-2). Given the planar structure of the (linear) film, and the harmonic nature of the excitation, we expect the x, y, and t dependence of all fields throughout the system for $t > 0$ to be given by $\exp^{[j(\omega t - k_x x - k_y y)]}$, where

$$k_x = k_o \sin\theta \cos\phi \quad k_y = k_o \sin\theta \sin\phi$$

and where $k_o = \omega\sqrt{\epsilon_a \mu_o}$ is the wave number of the incident field.

Maxwell's equations governing the system are thus given by

$$\begin{bmatrix} \nabla\times & 0 \\ 0 & \nabla\times \end{bmatrix} \begin{bmatrix} \mathbf{E} \\ \mathbf{H} \end{bmatrix} = j\omega \begin{bmatrix} -\bar{\boldsymbol{\zeta}} & -\bar{\boldsymbol{\mu}} \\ \bar{\boldsymbol{\epsilon}} & \bar{\boldsymbol{\xi}} \end{bmatrix} \begin{bmatrix} \mathbf{E} \\ \mathbf{H} \end{bmatrix} \tag{2.14}$$

where $\nabla\times$ is the curl operator given by

$$\nabla\times = \begin{bmatrix} 0 & -\frac{\partial}{\partial z} & \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} & 0 & -\frac{\partial}{\partial x} \\ -\frac{\partial}{\partial y} & \frac{\partial}{\partial x} & 0 \end{bmatrix} = \begin{bmatrix} 0 & -\frac{\partial}{\partial z} & -j\omega q_y \\ \frac{\partial}{\partial z} & 0 & j\omega q_x \\ j\omega q_y & -j\omega q_x & 0 \end{bmatrix},$$

where $q_x = \sqrt{\epsilon_a \mu_o} \sin\theta \cos\phi$ and $q_y = \sqrt{\epsilon_a \mu_o} \sin\theta \sin\phi$. One can always express the $\hat{\mathbf{z}}$ components of $\mathbf{E}$ and $\mathbf{H}$ in terms of the tangential fields as

$$\begin{bmatrix} \mathrm{E_z} \\ \mathrm{H_z} \end{bmatrix} = P \begin{bmatrix} \mathbf{E}_{xy} \\ \mathbf{H}_{xy} \end{bmatrix} \tag{2.15}$$

where $\mathbf{E}_{xy} = [\ \mathrm{E}_x \quad \mathrm{E}_y\ ]^T$, $\mathbf{H}_{xy} = [\ \mathrm{H}_x \quad \mathrm{H}_y\ ]^T$, and where

$$P = \begin{bmatrix} -\zeta_{33} & -\mu_{33} \\ \epsilon_{33} & \xi_{33} \end{bmatrix}^{-1} \begin{bmatrix} q_y + \zeta_{31} & -q_x + \zeta_{32} & \mu_{31} & \mu_{32} \\ -\epsilon_{31} & -\epsilon_{32} & q_y - \xi_{31} & -q_x - \xi_{32} \end{bmatrix},$$

provided the inverse exists. Substituting 2.15 into 2.14 yields

$$\frac{\partial}{\partial z} \begin{bmatrix} \mathbf{E}_{xy} \\ \mathbf{H}_{xy} \end{bmatrix} = D \begin{bmatrix} \mathbf{E}_{xy} \\ \mathbf{H}_{xy} \end{bmatrix}, \tag{2.16}$$

where D is given by

$$D = j\omega \begin{bmatrix} -\zeta_{21} & -\zeta_{22} & -\mu_{21} & -\mu_{22} \\ \zeta_{11} & \zeta_{12} & \mu_{11} & \mu_{12} \\ \epsilon_{21} & \epsilon_{22} & \xi_{21} & \xi_{22} \\ -\epsilon_{11} & -\epsilon_{12} & -\xi_{11} & -\xi_{12} \end{bmatrix} + j\omega \left(\begin{bmatrix} -\zeta_{23} & -\mu_{23} \\ \zeta_{13} & \mu_{13} \\ \epsilon_{23} & \xi_{23} \\ -\epsilon_{13} & -\xi_{13} \end{bmatrix} + \begin{bmatrix} -q_x & 0 \\ -q_y & 0 \\ 0 & -q_x \\ 0 & -q_y \end{bmatrix} \right) P.$$

We expect the inhomogeneity of the system to induce a reflected wave in the system everywhere $z > 0$, but it is important to note that $\mathbf{E}_{xy}$ and $\mathbf{H}_{xy}$ represent the total electric or magnetic fields, incident and reflected, at each point in space; the forward and backward propagating modes are undistinguished. We can change basis, however, and split the electromagnetic field differently to explicitly distinguish between modes propagating in the $+\hat{\mathbf{z}}$ or $-\hat{\mathbf{z}}$ directions. In general such a transformation will depend on z, but we will fix a transformation that corresponds to a valid wave-splitting in the ambient but does not carry any special physical interpretation elsewhere in the system.

We construct such a transformation by considering the fields in the ambient, with the goal of relating the total electromagnetic fields ($\mathbf{E}$, $\mathbf{H}$) to the forward and backward components of the electric field, $\mathbf{E}^+$ and $\mathbf{E}^-$. The total electric field $\mathbf{E}$ is clearly the sum of the incident field $\mathbf{E}^-$, traveling in a direction with negative $\hat{\mathbf{z}}$ component, and the reflected field $\mathbf{E}^+$, traveling in a direction with a positive $\hat{\mathbf{z}}$ component. Likewise, the total magnetic field is the sum of its forward and backward components, $\mathbf{H} = \mathbf{H}^- + \mathbf{H}^+$. So the connection is in relating the $+$ and $-$ components of the magnetic field to the corresponding components of the electric field. Recalling our previous computation leading to (2.11), we note that in a coordinate system (x', y', z'), where z' is the direction of propagation, x' is orthogonal to the plane of incidence and y' to be parallel to the plane of incidence, we have

$$\begin{bmatrix} \mathrm{E}_{\mathrm{x}'} \\ \mathrm{E}_{\mathrm{y}'} \\ 0 \end{bmatrix} = \sqrt{\frac{\mu_o}{\epsilon_a}} \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathrm{H}_{\mathrm{x}'} \\ \mathrm{H}_{\mathrm{y}'} \\ 0 \end{bmatrix}. \tag{2.17}$$

Rotating back to the (x, y, z) frame we then find

$$\begin{bmatrix} \mathrm{H}^+_{\mathrm{x}} \\ \mathrm{H}^+_{\mathrm{y}} \\ \mathrm{H}^+_{\mathrm{z}} \end{bmatrix} = \begin{bmatrix} 0 & -\cos\theta & \sin\theta \\ 1 & 0 & 0 \\ 0 & \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \mathrm{H}^+_{\mathrm{x}'} \\ \mathrm{H}^+_{\mathrm{y}'} \\ 0 \end{bmatrix} = \begin{bmatrix} -\mathrm{H}^+_{\mathrm{y}'}\cos\theta \\ \mathrm{H}^+_{\mathrm{x}'} \\ \mathrm{H}^+_{\mathrm{y}'}\sin\theta \end{bmatrix}, \tag{2.18}$$

$$\begin{bmatrix} \mathrm{H}_{\mathrm{x}}^{-} \\ \mathrm{H}_{\mathrm{y}}^{-} \\ \mathrm{H}_{\mathrm{z}}^{-} \end{bmatrix} = \begin{bmatrix} 0 & \cos\theta & \sin\theta \\ 1 & 0 & 0 \\ 0 & \sin\theta & -\cos\theta \end{bmatrix} \begin{bmatrix} \mathrm{H}_{\mathrm{x'}}^{-} \\ \mathrm{H}_{\mathrm{y'}}^{-} \\ 0 \end{bmatrix} = \begin{bmatrix} \mathrm{H}_{\mathrm{y'}}^{-}\cos\theta \\ \mathrm{H}_{\mathrm{x'}}^{-} \\ \mathrm{H}_{\mathrm{y'}}^{+}\sin\theta \end{bmatrix}, \tag{2.19}$$

with similar expressions for $\mathbf{E}$. Changing bases in (2.17) then yields

$$\begin{bmatrix} \mathrm{E}_{\mathrm{x}}^{+} \\ \mathrm{E}_{\mathrm{y}}^{+} \\ \mathrm{E}_{\mathrm{z}}^{+} \end{bmatrix} = \sqrt{\frac{\mu_o}{\epsilon_a}} \begin{bmatrix} 0 & \cos\theta & 0 \\ -\cos\theta & 0 & \sin\theta \\ 0 & -\sin\theta & 0 \end{bmatrix} \begin{bmatrix} \mathrm{H}_{\mathrm{x}}^{+} \\ \mathrm{H}_{\mathrm{y}}^{+} \\ \mathrm{H}_{\mathrm{z}}^{+} \end{bmatrix}, \tag{2.20}$$

$$\begin{bmatrix} \mathrm{E}_{\mathrm{x}}^{-} \\ \mathrm{E}_{\mathrm{y}}^{-} \\ \mathrm{E}_{\mathrm{z}}^{-} \end{bmatrix} = \sqrt{\frac{\mu_o}{\epsilon_a}} \begin{bmatrix} 0 & -\cos\theta & 0 \\ \cos\theta & 0 & \sin\theta \\ 0 & -\sin\theta & 0 \end{bmatrix} \begin{bmatrix} \mathrm{H}_{\mathrm{x}}^{-} \\ \mathrm{H}_{\mathrm{y}}^{-} \\ \mathrm{H}_{\mathrm{z}}^{-} \end{bmatrix}. \tag{2.21}$$

From (2.18) and (2.19), we note that the $\hat{\mathbf{z}}$ component of the magnetic fields are related to the $\hat{\mathbf{x}}$ component as $\mathrm{H}_z^+ = -\mathrm{H}_x^+ \tan\theta$ and $\mathrm{H}_z^- = \mathrm{H}_x^- \tan\theta$. Substituting for the $\hat{\mathbf{z}}$ component in (2.20) and (2.21) then yields

$$\begin{bmatrix} \mathbf{E}_{xy}^{+} \\ \mathbf{E}_{xy}^{-} \end{bmatrix} = \begin{bmatrix} Z_a & 0 \\ 0 & -Z_a \end{bmatrix} \begin{bmatrix} \mathbf{H}_{xy}^{+} \\ \mathbf{H}_{xy}^{-} \end{bmatrix}, \quad Z_a = \sqrt{\tfrac{\mu_o}{\epsilon_a}} \begin{bmatrix} 0 & \cos\theta \\ -\frac{1}{\cos\theta} & 0 \end{bmatrix}. \tag{2.22}$$

This, combined with the fact that $\mathbf{E} = \mathbf{E}^- + \mathbf{E}^+$ and $\mathbf{H} = \mathbf{H}^- + \mathbf{H}^+$, then produce the desired transformation

$$\begin{bmatrix} \mathbf{E}_{xy}^{-} \\ \mathbf{E}_{xy}^{+} \end{bmatrix} = T_a \begin{bmatrix} \mathbf{E}_{xy} \\ \mathbf{H}_{xy} \end{bmatrix}, \quad T_a = \tfrac{1}{2} \begin{bmatrix} I & -Z_a \\ I & Z_a \end{bmatrix} \tag{2.23}$$

In the new basis, Maxwell's Equations (2.14) become

$$\frac{\partial}{\partial z} \begin{bmatrix} \mathbf{E}_{xy}^{-} \\ \mathbf{E}_{xy}^{+} \end{bmatrix} = T_a D(z) T_a^{-1} \begin{bmatrix} \mathbf{E}_{xy}^{-} \\ \mathbf{E}_{xy}^{+} \end{bmatrix} \equiv \begin{bmatrix} \boldsymbol{\alpha}(z) & \boldsymbol{\beta}(z) \\ \boldsymbol{\gamma}(z) & \boldsymbol{\nu}(z) \end{bmatrix} \begin{bmatrix} \mathbf{E}_{xy}^{-} \\ \mathbf{E}_{xy}^{+} \end{bmatrix}, \tag{2.24}$$

where α, β, γ, and ν are 2×2 matrices. This wave-splitting is useful because it relates the forward and backpropogating modes, from which the reflection coefficients are

defined, to the dynamics of Maxwell's equations.

Let the reflection coefficient matrix $\mathbf{r}(z)$ be given by

$$\mathbf{E}^{+}(z) = \mathbf{r}(z)\mathbf{E}^{-}(z) \equiv \begin{bmatrix} \mathrm{r}_{11}(z) & \mathrm{r}_{12}(z) \\ \mathrm{r}_{21}(z) & \mathrm{r}_{22}(z) \end{bmatrix} \mathbf{E}^{+}(z). \tag{2.25}$$

Note that since the tangential components of the fields must be continuous, $\mathbf{r}$ is continuous in z even though the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, or $\boldsymbol{\nu}$ may be discontinuous. Substituting (2.25) into (2.24), we observe that the reflection coefficients evolve in z, independent of the fields, according to the following matrix Riccati equation

$$\frac{\partial \mathbf{r}}{\partial z} = \boldsymbol{\gamma} + (\boldsymbol{\nu}\mathbf{r} - \mathbf{r}\boldsymbol{\alpha}) - \mathbf{r}\boldsymbol{\beta}\mathbf{r}. \tag{2.26}$$

Given a suitable boundary condition $\mathbf{r}(0)$, this equation can be integrated to compute the optical response of the stratified film given by $\mathbf{r}(d)$, where energy conservation implies that

$$|\mathrm{r}_{ij}| \leq 1, \quad i, j = 1, 2$$

everywhere. Note that the structure of this equation holds even for general bian-isotropic media; the essential assumptions were existence of solutions with particular functional dependencies. Given harmonic plane-wave excitation, planar stratification and stationarity of the film are sufficient conditions for such solutions to exist.

Nevertheless, most applications do not require such generality, and for the common case of isotropic, non-magnetic films with $\phi = 0$, the expressions simplify substantially. In this case we find D to be given by

$$D = j\omega \begin{bmatrix} 0 & 0 & 0 & \mu_o\left(\frac{v^2}{\epsilon(z)} - 1\right) \\ 0 & 0 & \mu_o & 0 \\ 0 & \epsilon(z) - v^2 & 0 & 0 \\ -\epsilon(z) & 0 & 0 & 0 \end{bmatrix}$$

where $v \equiv \sqrt{\epsilon_a}\sin\theta$. Define $w_1 \equiv \sqrt{\epsilon_a \mu_o}/\cos\theta$ and $w_2 \equiv \sqrt{\epsilon_a \mu_o}\cos\theta$, then the

matrices $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\nu}$ are given by

$$\boldsymbol{\alpha} = \frac{j\omega}{2}\begin{bmatrix} \frac{\mu_o}{w_1}\epsilon(z) - w_1\left(\frac{v^2}{\epsilon(z)} - 1\right) & 0 \\ 0 & \frac{\mu_o}{w_2}\left(\epsilon(z) - v^2\right) + w_2 \end{bmatrix}$$

$$\boldsymbol{\beta} = \frac{j\omega}{2}\begin{bmatrix} \frac{\mu_o}{w_1}\epsilon(z) + w_1\left(\frac{v^2}{\epsilon(z)} - 1\right) & 0 \\ 0 & \frac{\mu_o}{w_2}\left(\epsilon(z) - v^2\right) - w_2 \end{bmatrix}$$

$$\boldsymbol{\gamma} = \frac{j\omega}{2}\begin{bmatrix} -\frac{\mu_o}{w_1}\epsilon(z) - w_1\left(\frac{v^2}{\epsilon(z)} - 1\right) & 0 \\ 0 & -\frac{\mu_o}{w_2}\left(\epsilon(z) - v^2\right) + w_2 \end{bmatrix}$$

$$\boldsymbol{\nu} = \frac{j\omega}{2}\begin{bmatrix} -\frac{\mu_o}{w_1}\epsilon(z) + w_1\left(\frac{v^2}{\epsilon(z)} - 1\right) & 0 \\ 0 & -\frac{\mu_o}{w_2}\left(\epsilon(z) - v^2\right) - w_2 \end{bmatrix}$$

Notice that the diagonal structure of these matrices imply that only r_{11} and r_{22} evolve in (2.26). That is, there is no change in the cross-polarized reflection coefficients for isotropic media. Substituting in to (2.26) we find

$$\frac{\partial \mathrm{r}_{11}}{\partial z} = \boldsymbol{\gamma}_{11} + 2\boldsymbol{\nu}_{11}\mathrm{r}_{11} + \boldsymbol{\gamma}_{11}\mathrm{r}_{11}^2 \tag{2.27}$$

$$\frac{\partial \mathrm{r}_{22}}{\partial z} = \boldsymbol{\gamma}_{22} + 2\boldsymbol{\nu}_{22}\mathrm{r}_{22} + \boldsymbol{\gamma}_{22}\mathrm{r}_{22}^2 \tag{2.28}$$

which can be rearranged to yield

$$\frac{\partial \mathrm{r}_{11}}{\partial z} = f_1(\mathrm{r}_{11}) + g_1(\mathrm{r}_{11})\epsilon(z) + h_1(\mathrm{r}_{11})\frac{1}{\epsilon(z)} \tag{2.29}$$

$$\frac{\partial \mathrm{r}_{22}}{\partial z} = f_2(\mathrm{r}_{22}) + g_2(\mathrm{r}_{22})\epsilon(z) \tag{2.30}$$

where

$$f_1 = w_1(1 - \mathrm{r}_{11})^2 \tag{2.31}$$

$$g_1 = -\frac{\mu_o}{w_1}(1 + \mathrm{r}_{11})^2 \tag{2.32}$$

$$h_1 = -w_1 v^2(1 - \mathrm{r}_{11})^2 \tag{2.33}$$

$$f_2 = \frac{\mu_o v^2}{w_2}(1 + \mathrm{r}_{22})^2 + w_2(1 - \mathrm{r}_{22})^2 \tag{2.34}$$

$$g_2 = -\frac{\mu_o}{w_2}(1 + \mathrm{r}_{22})^2 \tag{2.35}$$

2.2 Modeling Ellipsometry

An ellipsometer is a device that measures the polarization state of light interacting with an unknown film, or sample. These field measurements are then the basis of the optical inverse problem, that of determining the constitutive parameters of the sample given the measurements.

For ex-situ problems, the physics governing the optical response of stationary, stratified media relates the unknown parameters to these fields. In these cases the inverse problem, "characterization", is of primary interest. As a sensor technology for use in feedback applications, however, the in-situ film is necessarily non-stationary; the primary interest is modeling the dynamic response of the sensor so it can be used, along with process and actuator models, to guide the feedback design.

This section builds on the physics of stationary stratified films to develop a model of ellipsometry useful in feedback applications. First, we discuss sensor hardware and relate the measured variables to the reflection coefficients of a sample. Next, we explore growth dynamics of crystalline materials, showing how they impact the constitutive relations of a growing film. Finally, we suggest assumptions linking the physics of stationary media to the particular non-stationarity associated with film growth and present an associated model of ellipsometric dynamics.

2.2.1 Measuring Polarization

There are three primary hardware configurations that yield an ellipsometric measurement, and each of these use similar components. We will discuss each component and configuration briefly before focusing on the type of sensor used in in-situ applications [17, 74].

Hardware Components

All ellipsometer configurations use similar components to perform the measurement. These include sources, polarizers, modulators, compensators, and detectors.

- Light sources can be either monochromatic or broad spectrum, depending on whether single-wavelength or spectroscopic ellipsometry is desired. Key features of the source are that it is bright enough at the desired wavelength(s), and that its intensity is constant over time.

- Polarizers are lenses that only transmit light in one polarization direction. Thus, unpolarized light becomes linearly polarized when passed through a polarizer. Sometimes a polarizer placed just in front of the detector is called an "analyzer" to distinguish it from a "front-end" polarizer near the source.

- Modulators are devices that alter the light intensity or polarization state over time.

- Compensators are optical retarders that introduce a phase delay of 90° between orthogonal polarizations. Thus, a compensator will convert linearly polarized light to circularly polarized light.

- Detectors are photo-voltaic devices that convert light intensity to a voltage signal. Detector arrays are used in spectroscopic ellipsometers so that each detector in the array can respond to a particular wavelength of the incident light.

Hardware Configurations

All ellipsometer arrangements start with a light source and end with a detector; differences in configuration result from the arrangement of intermediate components. Here we discuss the three primary types of ellipsometers: null, polarization modulation, and rotating element ellipsometers.

Null ellipsometry is performed by adjusting the orientation of the polarizer, compensator, and analyzer until the measured light is extinguished. The typical configuration is:

Source $\rightarrow$ Polarizer $\rightarrow$ Compensator $\rightarrow$ Sample $\rightarrow$ Analyzer $\rightarrow$ Detector

The idea is that the polarizer and compensator are adjusted to elliptically polarize the source light "orthogonally" to the polarization effect of the sample. Thus, the reflected light is linearly polarized and can be extinguished by adjusting the analyzer. The alignment of the polarizer and compensator then reveal the intrinsic polarization affect of the sample. This techniques is usually performed manually, is slow, and is difficult for spectroscopic measurements since the "null" at each wavelength would need to be determined separately. Nevertheless, it can be very accurate.

Polarization modulation ellipsometers use incremental changes in the properties of the source light to generate incremental perturbations in the detector current. This is typically accomplished with a time-dependent optical retarder, yielding the following configuration:

Source $\rightarrow$ Polarizer $\rightarrow$ Modulator $\rightarrow$ Sample $\rightarrow$ Analyzer $\rightarrow$ Detector

Using periodic perturbations, the detector current can then be Fourier analyzed to yield information about the underlying sample structure. Advantages include a rapid data acquisition rate, but difficulty with noise, calibration, and temperature sensitivity have been documented.

Rotating element ellipsometers are similar to polarization modulation ellipsometers in that the detector current is Fourier analyzed for information about the sample, rather than relying on component settings. Such configurations do not require a compensator or modulator, leading to the following configuration:

Source $\rightarrow$ Polarizer $\rightarrow$ Sample $\rightarrow$ Analyzer $\rightarrow$ Detector

where either the polarizer or analyzer is continuously rotating at a specified frequency. By eliminating the compensator or modulator, this configuration relies on high quality polarizers, which are widely available, for accuracy in measurement. Figure 2-3 shows the experimental set-up of a typical rotating analyzer ellipsometer.

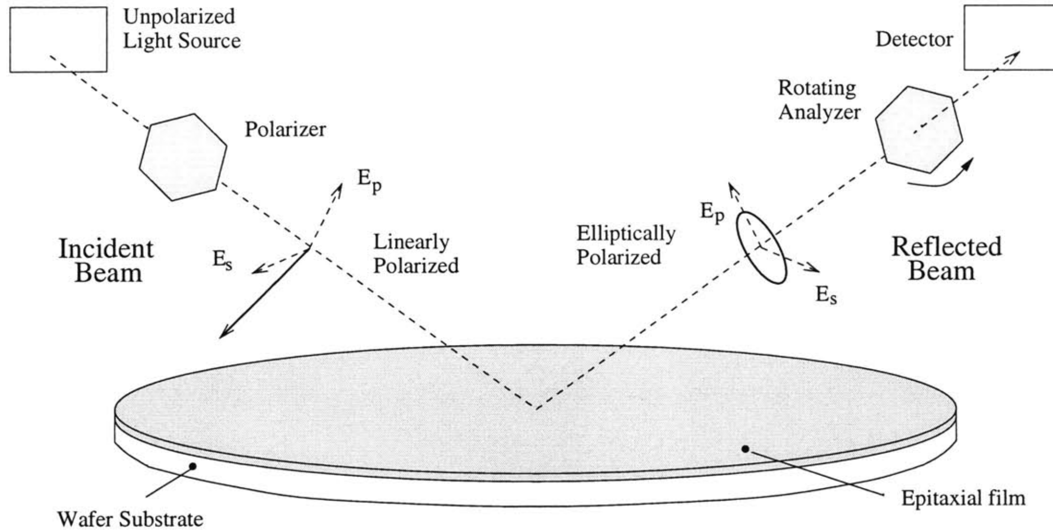


Figure 2-3: Rotating analyzer ellipsometer experimental set-up

Rotating Analyzer Ellipsometer

Figure 2-4 shows how the rotating analyzer ellipsometer (RAE) sweeps around the polarization ellipse to yield a measurement. The output signal is voltage as a function of time, and this voltage is proportional to the light intensity at the detector. Spectroscopic sensors collect multiple voltage measurements by using a grating at the front of the detector to defract the light, causing different wavelengths to strike different photo-detectors in an array.

To relate the fields and properties of the sample to this measured voltage, consider a linearly polarized beam given by

$$\begin{bmatrix} \mathbf{E}_{pi} \\ \mathbf{E}_{si} \end{bmatrix} = \begin{bmatrix} \cos(P) & -\sin(P) \\ \sin(P) & \cos(P) \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

where $\mathbf{E}_{pi}$ and $\mathbf{E}_{si}$ are the complex phasor coefficients describing the incident electric

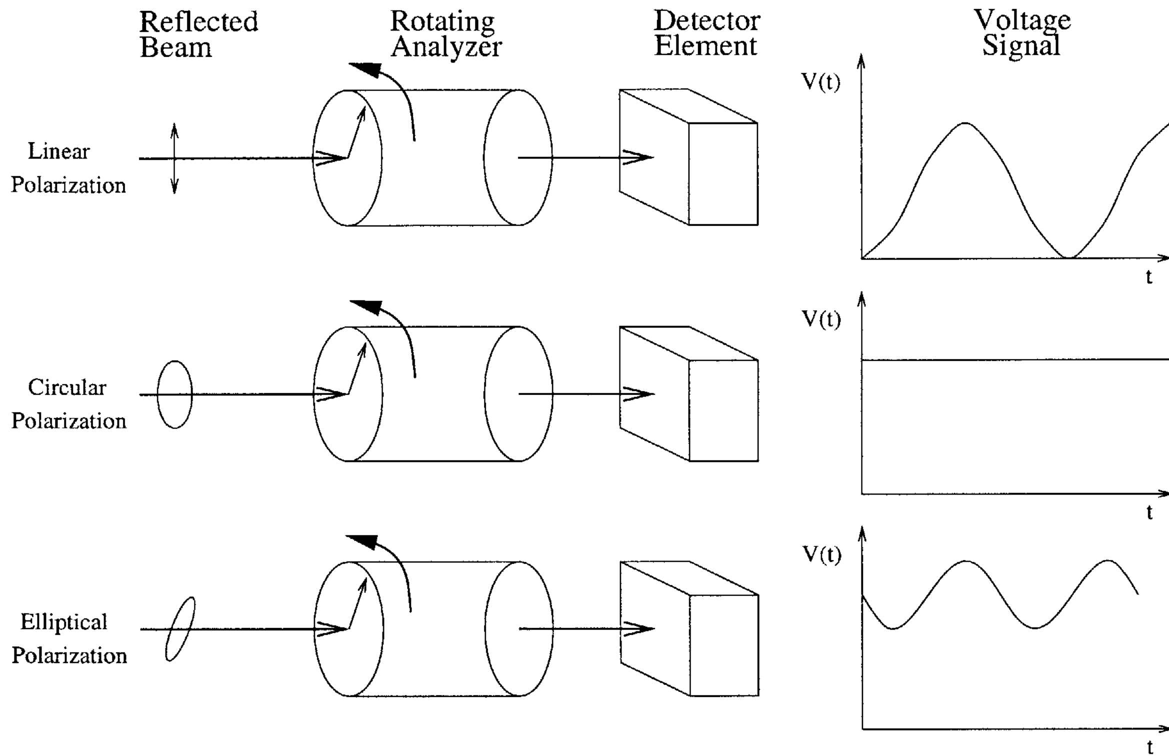


Figure 2-4: The rotating analyzer configuration yields various detector signals based on the polarization state of the reflected light.

field in the p and s-directions (see Figure 2-3), and P is the fixed polarizer orientation relative to the plane of incidence.

Now, reflection off an isotropic, smooth sample will alter the incident field as:

$$\begin{bmatrix} \mathbf{E}_{pr} \\ \mathbf{E}_{sr} \end{bmatrix} = \begin{bmatrix} \mathrm{r}_p(t) & 0 \\ 0 & \mathrm{r}_s(t) \end{bmatrix} \begin{bmatrix} \mathbf{E}_{pi} \\ \mathbf{E}_{si} \end{bmatrix}$$

where $\mathrm{r}_p = \mathrm{r}_{11}$ and $\mathrm{r}_s = \mathrm{r}_{22}$, and r_{11} and r_{22} are the reflection coefficients as given in (2.25). Recall that, in general, the reflection coefficients are complex valued quantities, and the explicit time dependence indicates that epitaxial growth may be in process.

The rotating analyzer will pass only the component of the reflected field aligned

with its polarization angle. Thus the field at the detector is given by

$$\begin{bmatrix} \mathbf{E}_{pd} \\ \mathbf{E}_{sd} \end{bmatrix} = \begin{bmatrix} \cos(\eta t) & -\sin(\eta t) \\ \sin(\eta t) & \cos(\eta t) \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \cos(\eta t) & \sin(\eta t) \\ -\sin(\eta t) & \cos(\eta t) \end{bmatrix} \begin{bmatrix} \mathbf{E}_{pr} \\ \mathbf{E}_{sr} \end{bmatrix}$$

where η is the frequency of analyzer rotation.

At the detector, the voltage $V(t)$ is proportional to the light intensity $V(t) \propto \mathbf{E}^H\mathbf{E}$ (where superscript H denotes the hermitian transpose). Putting this altogether we have

$$\begin{bmatrix} \mathbf{E}_{pd} \\ \mathbf{E}_{sd} \end{bmatrix} = \begin{bmatrix} \cos(\eta t)\left[\mathrm{r}_p(t)\cos(\eta t)\cos(P) + \mathrm{r}_s(t)\sin(\eta t)\sin(P)\right] \\ \sin(\eta t)\left[\mathrm{r}_p(t)\cos(\eta t)\cos(P) + \mathrm{r}_s(t)\sin(\eta t)\sin(P)\right] \end{bmatrix}$$

where, if we let $\mathcal{E} = [\mathrm{r}_p(t)\cos(\eta t)\cos(P) + \mathrm{r}_s(t)\sin(\eta t)\sin(P)]$, then we have

$$V(t) \propto \begin{bmatrix} \cos(\eta t)\mathcal{E}^H & \sin(\eta t)\mathcal{E}^H \end{bmatrix} \begin{bmatrix} \cos(\eta t)\mathcal{E} \\ \sin(\eta t)\mathcal{E} \end{bmatrix} = \mathcal{E}^H\mathcal{E}$$

$$\begin{aligned} \Rightarrow \mathcal{E}^H\mathcal{E} &= \mathrm{r}_p^H r_p \cos^2(P)\cos^2(\eta t) + \mathrm{r}_p^H r_s \cos(P)\sin(P)\cos(\eta t)\sin(\eta t) + \\ &\quad \mathrm{r}_s^H \mathrm{r}_p \cos(P)\sin(P)\cos(\eta t)\sin(\eta t) + \mathrm{r}_s^H \mathrm{r}_s \sin^2(P)\sin^2(\eta t) \end{aligned}$$

$$\begin{aligned} \Rightarrow \mathcal{E}^H\mathcal{E} &= \mathrm{r}_p^H \mathrm{r}_p \cos^2(P)\cos^2(\eta t) + \mathrm{r}_s^H \mathrm{r}_s \sin^2(P)\sin^2(\eta t) + \\ &\quad 2\mathrm{real}\left\{\mathrm{r}_p^H \mathrm{r}_s\right\}\cos(P)\sin(P)\cos(\eta t)\sin(\eta t) \end{aligned}$$

Using the identities $\sin(2\theta) = 2\sin(\theta)\cos(\theta)$, $\cos^2(\theta) = (1+\cos(2\theta))/2$, and $\sin^2(\theta) = (1-\cos(2\theta))/2$, we then have

$$\begin{aligned} \Rightarrow \mathcal{E}^H\mathcal{E} &= \mathrm{r}_p^H \mathrm{r}_p \cos^2(P)\cos^2(\eta t) + \mathrm{r}_s^H \mathrm{r}_s \sin^2(P)\sin^2(\eta t) + \\ &\quad \mathrm{real}\left\{\mathrm{r}_p^H \mathrm{r}_s\right\}\cos(P)\sin(P)\sin(2\eta t) \end{aligned}$$

$$\begin{aligned} \Rightarrow \mathcal{E}^H\mathcal{E} &= \mathrm{r}_p^H \mathrm{r}_p \cos^2(P)\left[\tfrac{1+\cos(2\eta t)}{2}\right] + \mathrm{r}_s^H \mathrm{r}_s \sin^2(P)\left[\tfrac{1-\cos(2\eta t)}{2}\right] + \\ &\quad \mathrm{real}\left\{\mathrm{r}_p^H \mathrm{r}_s\right\}\cos(P)\sin(P)\sin(2\eta t) \end{aligned}$$

$$\Rightarrow \mathcal{E}^H\mathcal{E} = \left[\mathrm{r}_p^H\mathrm{r}_p\cos^2(P) + \mathrm{r}_s^H\mathrm{r}_s\sin^2(P)\right] + \left[\mathrm{r}_p^H\mathrm{r}_p\cos^2(P) - \mathrm{r}_s^H\mathrm{r}_s\sin^2(P)\right]\cos(2\eta t) + \left[2\mathrm{real}\left\{\mathrm{r}_p^H\mathrm{r}_s\right\}\cos(P)\sin(P)\right]\sin(2\eta t).$$

Normalized, this signal has the form $1 + a(t)\cos(2\eta t) + b(t)\sin(2\eta t)$, where a and b are given by

$$a(t) = \frac{\left[\mathrm{r}_p^H\mathrm{r}_p\cos^2(P) - \mathrm{r}_s^H\mathrm{r}_s\sin^2(P)\right]}{\left[\mathrm{r}_p^H\mathrm{r}_p\cos^2(P) + \mathrm{r}_s^H\mathrm{r}_s\sin^2(P)\right]}$$

$$b(t) = \frac{\left[2\mathrm{real}\left\{\mathrm{r}_p^H\mathrm{r}_s\right\}\cos(P)\sin(P)\right]}{\left[\mathrm{r}_p^H\mathrm{r}_p\cos^2(P) + \mathrm{r}_s^H\mathrm{r}_s\sin^2(P)\right]}.$$

Measuring the voltage $V(t)$, then, allows one to estimate the coefficients $a(t)$ and $b(t)$, thereby eliminating η as an experimental parameter. Note that this estimation assumes that $a(t)$ and $b(t)$ vary slowly compared with η, which physically implies that the analyzer rotates quickly compared to film variation. These coefficients in turn can be related to the ellipsometric variables typically used to characterize the polarization state of a detected beam. Defining

$$\boldsymbol{\rho} \equiv \tan\Psi \exp^{j\Delta} \equiv \frac{\mathrm{r}_p}{\mathrm{r}_s},$$

where $\boldsymbol{\rho}$ is complex valued, while Ψ and Δ are real-valued, we note that these variables relate to the coefficients $a(t)$ and $b(t)$ as

$$a(t) = \frac{\left[\boldsymbol{\rho}^H(t)\boldsymbol{\rho}(t) - \tan^2(P)\right]}{\left[\boldsymbol{\rho}^H(t)\boldsymbol{\rho}(t) + \tan^2(P)\right]} = \frac{\left[\tan^2(\Psi(t)) - \tan^2(P)\right]}{\left[\tan^2(\Psi(t)) + \tan^2(P)\right]}$$

$$b(t) = \frac{\left[2\mathrm{real}\left\{\boldsymbol{\rho}(t)\right\}\tan(P)\right]}{\left[\boldsymbol{\rho}^H(t)\rho(t) + \tan^2(P)\right]} = \frac{\left[2\tan(\Psi(t))\cos(\Delta(t))\tan(P)\right]}{\left[\tan^2(\Psi(t)) + \tan^2(P)\right]}$$

and

$$\tan(\Psi(t)) = \sqrt{\frac{1 + a(t)}{1 - a(t)}}\,|\tan(P)|$$

$$\cos(\Delta(t)) = \frac{b(t)}{\sqrt{1 - a^2(t)}}\frac{\tan(P)}{|\tan(P)|}.$$

With these relationships, one can use knowledge of the polarizer angle P to deter-

mine $\Psi(t)$ and $\Delta(t)$ or, equivalently, $\boldsymbol{\rho}(t)$ from the detector voltage signal $V(t)$. We will thus hereafter consider the output of the ellipsometer to be the function $\boldsymbol{\rho}(t)$.

2.2.2 Non-Stationary Growing Films

The time dependence of $\boldsymbol{\rho}(t)$ is the result of time variation in $\mathrm{r}_p(t)$ and $\mathrm{r}_s(t)$, which did not arise in the optical response for stationary media. Nevertheless, the non-stationarity introduced by an actively growing film is also not as general as, say, the optical characterization of the atmosphere, with changing currents and moving cloud cover occurring simultaneously at different altitudes. Characterizing the particular time-dependence in a growing film's constitutive parameters begins with an understanding of the physical mechanisms governing deposition [61].

In general, deposition is a chemical accumulation process resulting in the incorporation of atoms onto the surface of a solid substrate. There are different techniques for inducing deposition, but ultimately each method creates thermodynamic conditions driving the intended chemical reaction, possibly limited by kinetic forces.

The process begins when an adatom or molecule approaches the surface. Within a few atomic distances of the surface, it will feel an attraction due to dangling bonds existing on the surface. This force propels the molecule toward the surface, where it may remain weakly bound if enough of its momentum dissipates into the surface. This weakly adsorbed state is called physical adsorption, and the molecule remains free to move about the surface.

Surface diffusion then governs the dynamics of the molecule as it moves randomly about the surface according to the available thermal energy. At high enough temperatures, the molecule may gain enough energy to desorb, otherwise it diffuses until it disassociates and each atom of the molecule forms chemical bonds with the surface atoms. These bonds involve the sharing of electrons in new molecular orbitals, and the atom will remain fixed at that lattice position under normal deposition conditions.

As atoms begin to accumulate on the surface, it becomes more and more likely that the diffusing molecules will encounter these fixed atoms already bound to the surface. One way this can happen is if the approaching molecule falls on top of an

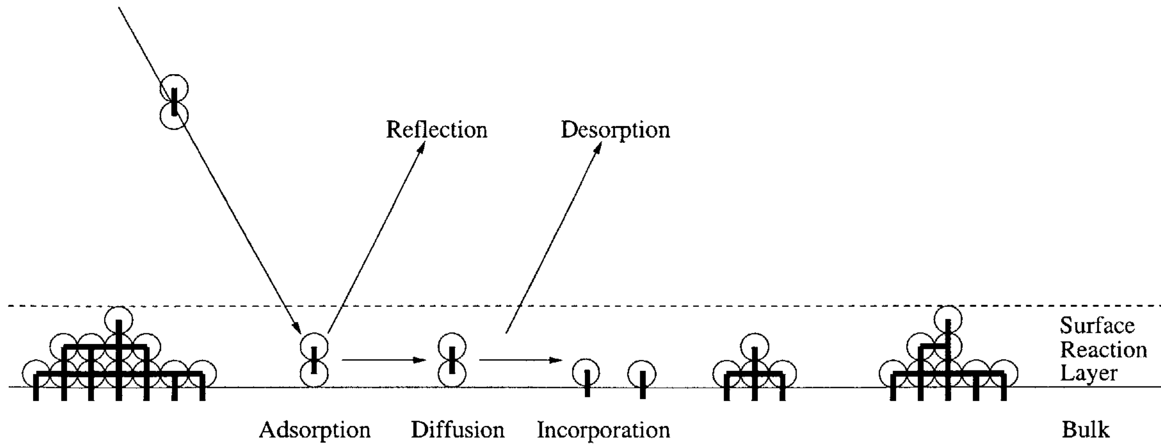


Figure 2-5: The surface reaction layer of a growing film is modeled as an equivalent planar-homogeneous film, with index graded smoothly over thickness.

existing bound atom. The likelihood of this happening is related to the proportion of the surface covered by bound atoms of the new atomic layer, but even then surface diffusion will have the tendency to transport this approaching molecule off the top of the new layer into lower energy positions. As a result, under the right thermodynamic conditions it is much more likely that an approaching molecule will bind next to a previously bound atom than on top of it. Islands of atoms thus form in a process called nucleation, and although new islands sometimes form before one atomic layer is completely coalesced into a smooth new layer, it is common to have no more than two or three atomic layers of islands growing before the bottom layer completely fuses together, although this depends strongly on the diffusion length [56].

This process suggests that in the surface reaction layer, between the last fully coalesced atomic layer and the height of the highest island cluster, the constitutive parameters are inhomogeneous in the plane orthogonal to the surface normal. The electromagnetic impact of this roughness on light of a given wavelength λ, however, is equivalent to that of a smooth, homogeneous medium provided the surface is uniformly rough, and the roughness is composed of many islands and valleys that are each small in diameter relative to λ. This equivalence is known as effective medium theory, and different models are available that relate the percentage of surface coverage of material with specified constitutive parameters to new, *effective* constitutive

parameters that describe a fictitious homogeneous film with equivalent optical effect.

Appealing to effective medium theory, our assumptions on the nucleation dynamics are that at each atomic level of the surface reaction layer the roughness of that level is uniform, with no large islands or large open regions without growth. Thus, initially the atomic level is covered by a uniform blanket of atoms that diffuse to create a uniform effective presence over some local region of the surface. As more of the surface is covered, a few small islands begin to form, spread out uniformly over the surface, with large numbers of diffusing atoms in between. This would suggest that the temperature is high enough to maintain larger diffusion lengths, yet low enough to prevent too much desorption. New island formation in the open spaces is preferred to large clustering at that level, since the probability of interaction between diffusing atoms is higher than that of an encounter with the small islands. As the density of island coverage increases, we assume the density of diffusing atoms between islands increases also, thereby maintaining the sense of uniform roughness by incorporation of many small inclusions in that atomic level of the reaction layer.

At higher (i.e. physically on top of) atomic levels of the reaction layer, we assume similar dynamics of uniform roughness are occurring, but with less coverage than that of lower atomic levels in the layer. In reality the need for long diffusion lengths for uniform coverage also suggests that adatoms from higher atomic levels will have a greater probability of finding holes in the next level down and expedite that level's coalescence rather than contribute to the formation of its own atomic level. Thus, the conditions necessary for uniform coverage and use of effective medium theory are also conditions that promote 2-D growth and a very thin surface reaction layer.

Even when these assumptions are not met, however, if the process ultimately produces relatively uniform films macroscopically, then the amount of time that the atomic levels of the surface reaction layer exhibit planar inhomogeneity because, for example, of large island formation, are relatively brief, on the order of the thickness of the surface reaction layer times the growth rate. Eventually, every atomic level of the reaction layer coalesces into a uniform film, admitting the use of effective medium theory and a clear definition of alloy composition as the relative average ratios of

elemental materials homogeneously mixed throughout the layer.

Nevertheless, although a specific atomic layer may have a well defined *composition* based on the planar mixing of elements uniformly throughout the layer, it is not clear how to meaningfully assign values of constitutive parameters on these scales, where classical approaches breakdown and quantum effects begin to dominate. Certainly it is not easy to grow a uniform monolayer of an alloy and measure these parameters, nor is it easy to abstract simple relations from the quantum effects. These challenges contribute to the difficulty of representing the constitutive parameters of a growing film along a temporal and spatial continuum.

Since the problem we are interested in is fundamentally not quantum mechanical, however, the essential feature of our representation is that it parsimoniously relates the observable optical effect of arbitrary multi-layer stacks of alloys, as they become thick enough to register with our sensor technology, to the process of accumulating thickness in film growth. To accomplish this, we first characterize composition of a growing film as a function of the continuums (z, t), and then we relate this composition function to the constitutive parameters.

The simplest way to embed composition in the continuum is to imagine fictitious values of composition that can be defined pointwise for any value z. Although this has no physical meaning on scales less than a monolayer, one may think of the true composition of the monolayer as the average of the fictitious values of composition defined at each height throughout the layer. We then can consider film thickness, characterizing the time-variation of the film, to be a continuous function, $d(t)$, such that $d(0) = 0$ and $\partial d/\partial t \geq 0$. Composition of a stratified (ternary) alloy, then, is a function $y(z,t) : \mathbb{R}^2 \rightarrow \{[0\ 1], y_s, y_a\}$ that has a fixed value y_s everywhere $z < 0$, and a fixed value y_a everywhere $z > d(t)$. On the region $0 < z < d(t)$, y takes on values on the unit interval $[0\ 1]$.

Note that although the film thickness is a continuous function of time, composition is not necessarily continuous in t or z. Also, the definition trivially extends for more complex alloys characterized by vector-valued compositions defined over the same region.

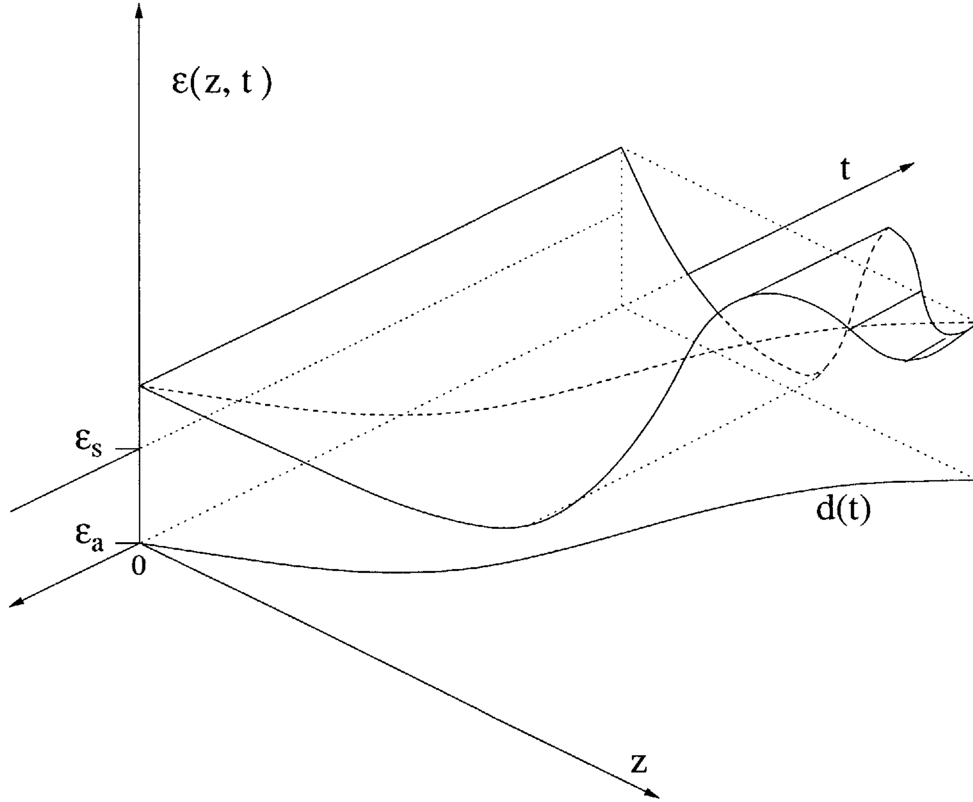


Figure 2-6: The structure of $\epsilon(z,t)$ for an idealized growing film.

Given this characterization of material composition for a growing film, we will assume the constitutive parameters of the film are then described by the pointwise assignment of the empirically derived bulk values of constitutive parameters associated with thick, homogeneous films of a specified composition, $H : y(z,t) \to \mathcal{C}(z,t)$. This characterization is known to be inaccurate for some materials, and it avoids dispersion effects caused by the differing penetration depths of light at various wavelengths. These effects can be modeled, however, by smoothing with a circularly symmetric low-pass filter where the aperture size varies with penetration depth δ_p. For an isotropic film, then, without smoothing, the permittivity is then a function of the form

$$\epsilon(z,t) = \begin{cases} \epsilon_s & z < 0 \\ H(y(z,t)) & 0 \leq z \leq d(t) \\ \epsilon_a & z > d(t) \end{cases} \tag{2.36}$$

In the absence of more accurate relations derived from first principles, this repre-

sentation of the optical constants for a growing film capture some important features of the growth process. First, we acknowledge the imperfections at small scales of embedding a quantized problem in the continuum by making the constitutive parameters continuous, so that even sharp discontinuities in composition yield blurred boundaries in terms of the optical functions. Second, the constitutive parameters at any fixed height z^* is the constant value of the ambient until it intersects the thickness function $d(t)$, at which time it rapidly transitions to another fixed value (the bulk value associated with the composition deposited at that height). Third, at any fixed time t^* the system is a substrate, film, ambient system with the film characterized by a fixed thickness $d(t^*)$ and stratified constitutive parameters given by $\mathcal{C}(z),\ 0 < z < d(t^*)$.

2.2.3 Frozen Time Assumptions

Given the structure of constitutive parameters exemplified in 2.36, we would then like to understand how to adapt our analysis of the evolution of r given by 2.29 and 2.30. Unfortunately, it is often difficult to solve Maxwell's equations for cases that are both non-stationary and stratified [30].

Nevertheless, we leverage the observation that at any fixed time t^*, the constitutive parameters are exactly of the form considered previously for a stationary stratified film. Thus, if the growth function $d(t)$ remained constant briefly, the results from 2.29 and 2.30 would apply. We might then consider a piecewise constant approximation to the growth function $d(t)$, where for each "frozen time" segment the stationary results apply. Patching together all of these frozen time solutions then yields the evolution of surface reflectivities $\mathrm{r}(t)$.

The key assumptions needed to make such reasoning apply are that the fields remain well behaved over the transitions between frozen time segments. We will assume this to be the case provided the jump in thickness between frozen time segments is not too large. Equivalently, this suggests that the growth rate $\partial d/\partial t$ is not too large.

Note that 2.29 and 2.30 describe how the amplitude reflectances of a stationary stratified film evolve over the height of the film, assuming that only the ambient is present above each position as it is considered. This analysis is thus very suggestive

of a growing film with thickness z. To model the evolution of r at the surface of a non-stationary film, we replace $\epsilon(z)$ by $\epsilon(z,t)$ and replace the independent variable z in the equations by the thickness function $d(t)$, describing the height of the film at time t. This yields

$$\frac{\partial \mathrm{r}_{11}}{\partial d(t)} = f_1(\mathrm{r}_{11}) + g_1(\mathrm{r}_{11})\epsilon(d(t),t) + h_1(\mathrm{r}_{11})\frac{1}{\epsilon(d(t),t)} \tag{2.37}$$

$$\frac{\partial \mathrm{r}_{22}}{\partial d(t)} = f_2(\mathrm{r}_{22}) + g_2(\mathrm{r}_{22})\epsilon(d(t),t) \tag{2.38}$$

Changing variables then yields a time evolution of the reflectivities which, combined with the ellipsometric output $\boldsymbol{\rho} = \mathrm{r}_{11}/\mathrm{r}_{22}$, produces a dynamic model of ellipsometry consistent with our frozen time assumptions [69]

$$\frac{\partial \mathrm{r}_{11}}{\partial t} = \left[f_1(\mathrm{r}_{11}) + g_1(\mathrm{r}_{11})\epsilon(t) + h_1(\mathrm{r}_{11})\frac{1}{\epsilon(t)}\right]\frac{\partial d}{\partial t} \tag{2.39}$$

$$\frac{\partial \mathrm{r}_{22}}{\partial t} = \left[f_2(\mathrm{r}_{22}) + g_2(\mathrm{r}_{22})\epsilon(t)\right]\frac{\partial d}{\partial t} \tag{2.40}$$

$$\boldsymbol{\rho}(t) = \frac{\mathrm{r}_{11}(t)}{\mathrm{r}_{22}(t)} \tag{2.41}$$

where here we have abused our notation and written $\epsilon(t)$ for $\epsilon(d(t),t)$, since it does not seem that the simpler notation would introduce any confusion.

2.3 Analysis

As a sensor technology to be used in feedback applications, we are interested in understanding the behavior of the dynamics described by 2.39, 2.40, and 2.41. Nevertheless, because of the nonlinear nature of these equations, it is not easy to come to such an understanding without some analysis. This section will explore the nature of the equilibria, stability, and the zero dynamics of the ellipsometer system.

2.3.1 Equilibria and Fresnel's Formulas

The dynamics of the ellipsometer model given by 2.39 and 2.40 reveal the conditions under which equilibrium is achieved. Suppose we are depositing a fixed material characterized by $\boldsymbol{\epsilon}$, so that $\boldsymbol{\epsilon}(t) \equiv \boldsymbol{\epsilon}$. As the film grows thicker, the system begins to physically approximate a simple interface between two semi-infinite homogeneous media, the ambient and the thick film.

We can compute the location of the system equilibrium under constant excitation $\boldsymbol{\epsilon}$ by equilibrating the system equations 2.39 and 2.40

$$0 = f_1(\mathrm{r}_{11}) + g_1(\mathrm{r}_{11})\epsilon_s + h_1(\mathrm{r}_{11})\frac{1}{\epsilon_s}$$

$$0 = f_2(\mathrm{r}_{22}) + g_2(\mathrm{r}_{22})\epsilon_s,$$

which can be rearranged to yield

$$0 = \boldsymbol{\gamma}_{11} + 2\boldsymbol{\nu}_{11}\mathrm{r}_{11} + \boldsymbol{\gamma}_{11}\mathrm{r}_{11}^2$$

$$0 = \boldsymbol{\gamma}_{22} + 2\boldsymbol{\nu}_{22}\mathrm{r}_{22} + \boldsymbol{\gamma}_{22}\mathrm{r}_{22}^2$$

where the parameters in the equations are as specified in 2.39, 2.40, 2.27 and 2.28. The quadratic nature of these equations reveal that for each polarization there are, in fact, two equilibria, given by

$$\mathrm{r}^{\{+,-\}} = \frac{-\boldsymbol{\nu} \pm \sqrt{\boldsymbol{\nu}^2 - \boldsymbol{\gamma}^2}}{\boldsymbol{\gamma}}.$$

Substituting for the values of the parameters $\boldsymbol{\nu}$ and $\boldsymbol{\gamma}$ yield the following characterization of the system equilibria

$$\mathrm{r}_{11}^{\{+,-\}} = \frac{\left(\sqrt{\epsilon_a}\cos\theta_\epsilon \pm \sqrt{\boldsymbol{\epsilon}}\cos\theta\right)^2}{\left(\sqrt{\epsilon_a}\cos\theta_\epsilon + \sqrt{\boldsymbol{\epsilon}}\cos\theta\right)\left(\sqrt{\epsilon_a}\cos\theta_\epsilon - \sqrt{\boldsymbol{\epsilon}}\cos\theta\right)}$$

$$\mathrm{r}_{22}^{\{+,-\}} = \frac{\left(\sqrt{\epsilon_a}\cos\theta \pm \sqrt{\boldsymbol{\epsilon}}\cos\theta_\epsilon\right)^2}{\left(\sqrt{\epsilon_a}\cos\theta + \sqrt{\boldsymbol{\epsilon}}\cos\theta_\epsilon\right)\left(\sqrt{\epsilon_a}\cos\theta - \sqrt{\boldsymbol{\epsilon}}\cos\theta_\epsilon\right)}$$

where θ_ϵ is defined from the relation $\epsilon_a \sin^2\theta = \boldsymbol{\epsilon} \sin^2\theta_\epsilon$.

Inspection of these equations reveal that the location of these equilibria are related by $\mathrm{r}^+ = 1/\mathrm{r}^-$. Clearly, then, only r^- will generally satisfy the conservation of energy requirement, $|\mathrm{r}| \leq 1$. These expressions for r^- are exactly the Fresnel formulas characterizing the reflection coefficients at the boundary of two homogeneous media. Thus Fresnel's formulas can be interpreted as the physically meaningful equilibrium of the optical system equations under constant input $\boldsymbol{\epsilon}$.

2.3.2 Local Behavior of the System

Characterizing the equilibria allows us to gain insight into the local behavior of the system by linearizing around an arbitrary equilibrium and studying the properties of the resulting linear system. In particular, we are interested in conditions for stability of the system near r^-.

Define the deviation of the system state variable from equilibrium as x, where $\mathrm{x} = \mathrm{r} - \mathrm{r}^-$. The dynamics $\dot{\mathrm{r}} = \gamma + 2\boldsymbol{\nu}\mathrm{r} + \gamma\mathrm{r}^2$ then yield

$$\frac{\partial \mathrm{x}}{\partial t} = 2\left[\boldsymbol{\nu} + \gamma\mathrm{r}^-\right]\frac{\partial d}{\partial t}\mathrm{x} + \gamma\frac{\partial d}{\partial t}\mathrm{x}^2$$

where, defining $\boldsymbol{\epsilon} = \epsilon_r + i\epsilon_i$, we have

$$\left[\boldsymbol{\nu} + \gamma\mathrm{r}^-\right] = -\sqrt{\boldsymbol{\nu}^2 - \gamma^2} = -\omega\sqrt{\mu_o}\sqrt{\left(\epsilon_a \sin^2\theta - \epsilon_r\right) - i\epsilon_i}.$$

Considering the square root of a complex quantity, this expression will have negative real part only if $\epsilon_i < 0$. Thus, we conclude that a sufficient condition for stability of the equilibrium r^- is that $\epsilon_i < 0$, which is expected for physically meaningful conducting materials. Note that the same condition implies instability of the equilibrium r^+.

2.3.3 Zero Dynamics and Brewster's Angle

Finally, we are interested in understanding the behavior of the zeros of the system, that is to say, the internal behavior of the system under conditions yielding identically

zero output. We note from the system output equation $\boldsymbol{\rho} = \mathrm{r}_{11}/\mathrm{r}_{22}$ that zeroing the output is achieved by an input $\boldsymbol{\epsilon}$ that maintains $\dot{\mathrm{r}}_{11} \equiv 0$, and an initial condition $\mathrm{r}_{11} = 0, \mathrm{r}_{22} \neq 0$.

Computing the $\boldsymbol{\epsilon}$ that zeros the output, we find that two choices of $\boldsymbol{\epsilon}$, denoted ϵ^o, do the job.

$$\begin{aligned}
\dot{\mathrm{r}}_{11} \equiv 0 \quad &\Rightarrow \quad \gamma_{11} + 2\boldsymbol{\nu}_{11}\mathrm{r}_{11} + \gamma_{11}\mathrm{r}_{11}^2 \\
&\Rightarrow \quad \gamma_{11} = 0 \\
&\Rightarrow \quad \tfrac{j\omega}{2}\left[-\tfrac{\mu_o \boldsymbol{\epsilon}}{w_1} - w_1\left(\tfrac{v^2}{\boldsymbol{\epsilon}} - 1\right)\right] = 0 \\
&\Rightarrow \quad \mu_o \boldsymbol{\epsilon}^2 - w_1^2 \boldsymbol{\epsilon} + w_1^2 v^2 = 0 \\
&\Rightarrow \quad \epsilon^o = \tfrac{\epsilon_a}{2\cos\theta}\left[\tfrac{1}{\cos\theta} \pm \sqrt{\tfrac{1}{cos^2\theta} - 4\sin^2\theta}\right] \\
&\Rightarrow \quad \epsilon^o = \begin{cases} \epsilon_a \\ \epsilon_a \tan^2\theta \end{cases}
\end{aligned}$$

These expressions for the input that zero the output have physically meaningful interpretations. In the one case, if we grow a "film" composed of the same material as the ambient, then the measurement at the "interface" between ambient and film is zero. On the other hand, we may grow a film composed of a different material, but its permittivity must be related to the angle of incidence, θ, in a special way for it to zero the output. In particular, ϵ^o is a material such that θ is precisely the Brewster angle, known to yield zero reflectance for this polarization.

To understand the zero dynamics, however, we now study the stability of r_{22} when driven by the same input that zeros r_{11}. We recall that for any constant input $\boldsymbol{\epsilon}$ we found two equilibria that were asymptotically stable and unstable, respectively, when $\epsilon_i < 0$. In this case, however, ϵ^o is real-valued, so $\epsilon_i = 0$ and the linearization approach employed earlier is inconclusive.

Substituting $\boldsymbol{\epsilon} = \epsilon_a \tan^2\theta$ into the dynamics governing r_{22}, we find equilibria $\mathrm{r}^{\{+,-\}}$. We restrict our attention to the physically meaningful equilibrium, r^-, and, defining $\mathrm{x} = \mathrm{r} - \mathrm{r}^-$, find local dynamics of the form

$$\dot{\mathrm{x}} = ia\mathrm{x} + ib\mathrm{x}^2.$$

where a and b are real-valued functions of ω, θ, and ϵ_a (not to be confused with the Fourier coefficients on the measured voltage as discussed earlier). Since the linearization approach is not conclusive for understanding the stability of this system, we will study the solutions of this system directly to understand its stability properties. Integrating, we find

$$
\begin{aligned}
-\frac{i}{a}\int\frac{\partial \mathrm{x}}{\mathrm{x}}+\frac{ib}{a}\int\frac{\partial \mathrm{x}}{a+b\mathrm{x}} &= t+\mathrm{c} \\
\Rightarrow \ln\frac{\mathrm{x}}{a+b\mathrm{x}} &= iat+\mathrm{c} \\
\Rightarrow \frac{\mathrm{x}}{a+b\mathrm{x}} &= \tilde{\mathrm{c}}e^{iat}
\end{aligned}
$$

where $\tilde{\mathrm{c}}$ is specified by the initial condition. This implies that

$$|\mathrm{x}| = |\tilde{\mathrm{c}}||\mathrm{a}+\mathrm{bx}|$$

$$\Rightarrow x_r^2+x_i^2 = |\tilde{\mathrm{c}}|^2\left[(a+bx_r)^2+(bx_i)^2\right]$$

$$\Rightarrow (1-|\tilde{\mathrm{c}}|^2b^2)x_r^2-2|\tilde{\mathrm{c}}|^2abx_r+(1-|\tilde{\mathrm{c}}|^2b^2)x_i^2 = |\tilde{\mathrm{c}}|^2a^2$$

and if $|\tilde{\mathrm{c}}|^2b^2 \neq 1$, then we have

$$\left(x_r-\frac{|\tilde{\mathrm{c}}|^2ab}{1-|\tilde{\mathrm{c}}|^2b^2}\right)^2+x_i^2 = \frac{|\tilde{\mathrm{c}}|^2a^2}{(1-|\tilde{\mathrm{c}}|^2b^2)^2}$$

This is the equation of a closed curve (specifically, a circle), which does not pass through the equilibria (the origin, r^-, or $x=-a/b$, which is r^+ in the local coordinate system). Since the solutions are constrained to a closed curve, we note that the zero dynamics are marginally stable, except when the initial condition is on the line $x_r = -a/2b$. In this case, $|\tilde{\mathrm{c}}|^2b^2 = 1$, the equation for the circle degenerates into a line, and the dynamics run off to infinity, hence yielding unstable zero dynamics.

Zero dynamics that are not asymptotically stable suggest caution is needed if one hopes to invert the system as part of a feedback design. In particular, such inversion renders some at-least-marginally-unstable mode unobservable, removing it beyond the reach of any feedback system for stabilization. If using "built-in" software

designed to invert the ellipsometer dynamics, one must be careful to ensure that the resulting unobservable oscillations in r_{22} can not inadvertently lead to an undetected periodicity in the material structure of the growing film.

Chapter 3

Control of Epitaxy using Ellipsometric Feedback

The production of many compound semiconductor devices relies on the ability to continuously control the alloy composition of a growing, high-quality crystalline film. A range of epitaxial processes and specialized reactors have been developed that are capable of depositing high quality films uniformly over a wafer substrate, but the ability to reproducibly and precisely control this growth with first-pass success has been an open problem.

Ellipsometry provides a mechanism to continuously monitor the changing polarization state of reflected light from a growing film. Using this measurement, the question of how to adjust process variables to realize a desired device profile becomes meaningful.

This chapter formulates the control problem for epitaxial growth of a desired device structure with ellipsometry as a sensor technology. First, epitaxy is described as a deposition process, and simple models are presented reflecting a range of techniques used in production. Next, the growth objective is discussed as a relaxation of typical tracking problems. Finally, a simulation study is presented, demonstrating that even when approaching the control problem as a standard tracking problem, currently ungrowable structures are realizable.

3.1 Epitaxial Processes

Epitaxy is a specific deposition process where the crystallographic order of the film is inherited from the substrate. By depositing on high quality crystalline substrates, single crystalline films can be grown. This section describes the growth dynamics and actuation devices, discusses the relationship between different production processes, and presents a model that roughly captures the behavior of this spectrum of epitaxial techniques.

3.1.1 Growth Dynamics

The dynamics of adsorption, nucleation, and coalescence associated with epitaxy are similar to those of more general deposition processes as described in Section 2.2.2. Regardless of the specific chemistry of a particular epitaxial process, then, the structure of these idealized growth dynamics is somewhat generic, parametrized by variables that depend on the specific process and materials being deposited [55, 62]. This idealization hypothesizes a growth regime that satisfies certain conditions:

1. There are no parasitic reactions, i.e. all group III source molecules reach the growth interface,

2. Ideal gas behavior, if applicable i.e. in vapor phase processes reactants are sufficiently diluted to justify use of the ideal gas law,

3. Group III limited growth i.e. all group III molecules at the growth interface are incorporated uniformly into the growing film,

4. Stoichiometric growth, i.e. no group V reactant is incorporated into the solid phase without the corresponding minority component,

5. No undesired impurities are introduced into the film.

From these assumptions about the generic nature of the growth regime, a simple model of the growth dynamics can be derived from first principles, as described in

[67, 62, 61]. For a quaternary system such as $In_{y1}Ga_{1-y1}As_{y2}P_{1-y2}$, where (In, Ga) are group III elements and (As, P) are group V elements, this model is given by

$$\frac{\partial d}{\partial t} = |k_1 v_1 + k_2 v_2|$$
$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \frac{k_5 v_1}{k_5 v_1 + k_6 v_2} \\ \frac{k_7 v_3}{k_7 v_3 + k_8 v_4} \end{bmatrix} \tag{3.1}$$

where d is the film thickness; $v_{1,2,3,4}$ are the molecular fluxes of the input species In, Ga, As, and P, respectively; and $k_{1,2,3,4,5,6,7,8}$ are parameters defined by the specific epitaxy process and material system.

This generic description of growth simply implies that the film grows thicker as more group III flux is applied, and that the weighted relative mixing between group III flux and group V flux define alloy composition. Besides parametric affects captured by k_{1-8}, differences between specific epitaxy processes then appear primarily by different actuation mechanisms for generating these material fluxes, $v_{1,2,3,4}$, from whatever actual control touch-points, $u_{1,2,3,4}$, are available to that particular process.

3.1.2 Actuation Devices

There are two types of actuation devices used by the spectrum of processes considered in this study. First, there are gas sources that are actuated by mass flow controllers to regulate the flow and hence the flux of the desired material. The other actuation device is an elemental source, or effusion cell, that heats a charge of some input material to evaporate either individual atoms or molecules, generating molecular flux. Both of these kinds of sources are discussed below.

Gas Sources

The simplest of the gas sources are hydride sources. These sources produce a steady group V flux by generating a consistent stream of group V hydride gas. The actuation point for these sources is a mass flow controller, which is easily manipulated to yield

the desired molecular flux. Dangers of using these sources result from the extreme toxicity of various hydrides, such as arsine and phosphine.

Group III gas sources are more complicated than those of group V elements. Alkyl group III metal-organic compounds are used as the precursor materials to produce the flux for these elements since these compounds are 1) liquid at room temperature (except for trimethylindium, which is solid), 2) stable up to at least 200°C, and 3) have vapor pressures that are generally in the one to tens of Torr at temperatures between -10 and 100°C. These properties suggest a simple mechanism for generating the desired flux by passing hydrogen through the liquid metal-organic compound (or over the solid, in the case of trimethylindium) to carry specific amounts of precursor molecules. By fixing the temperatures and pressures of the containers, called bubblers, holding these precursors, the resulting partial pressure of the precursor elements define the ammount of metalorganics carried by the hydrogen gas. In this way, the actuation points become the mass flow controllers regulating the hydrogen flow through the bubblers. Although the concentration of precursor molecules will change over time as the bubbler is depleted, this effect can be largely ignored, leaving the input flux some fixed percentage of the commanded flow. Purity may also be an issue, compared with elemental sources for the same material. Also, the flow control is sensitive to the ambient temperature and pressure across the flow controllers.

The deposition chemistry is usually more involved when gas sources are used, since the approaching molecule must disassociate, often in a complex process (especially when metal-organics are involved) as part of the reaction mechanism. Nevertheless, under the right conditions, the growth behaves in the "mass transport limited regime", as opposed to other regimes where reaction kinetics play a more visible role, and the resulting dynamics approximate 3.1. Moreover, the use of mass flow controllers as actuation points results in steady flow over a broad range of flux commands, and we model this flow with the relation $v_i = u_i$, although a slight first order lag could be used to capture fast dynamic effects if desired. The scaling effects associated with different partial pressures in bubblers, etc. are captured by parameters k_i associated with the relevant source.

Elemental Sources

Effusion cells are arguably the most crucial component of MBE systems, used to generate ultra-pure molecular beams for a variety of materials. These containers are usually made of Boron Nitride, slightly conical in shape and open on one end. Around the container, heating elements control the temperature over a broad range, sometimes higher than 1400°C. Inside the container is an ultra-pure charge of a selected material, called the melt. As the melt is heated, molecules evaporate and a molecular beam effuses from the open end. A shutter can close over the orifice, allowing for abrupt transitions in molecular flux within the growth chamber.

One of the most obvious restrictions of using effusion cells is the line-of-sight requirement between the cell and substrate. Nevertheless, other difficulties, such as slow dynamics associated with temperature as an actuation point, changes in flux as the melt decays with use, or sensitivity to slight configuration adjustments as interface components expand and contract with temperature can make use of the cells trying. Some research has focused on the flux transients that result when the cell shutter opens and the melt temperature drops [64], and various flux sensors can be employed to better understand these temporal behaviors for a particular cell.

While these issues are explored, however, a simple idealization of a cell's behavior can be developed from first principles and empirical observation. Knudsen evaporation considers a nearly closed box, inside of which the vapor and liquid phases are maintained in equilibrium. A small amount of vapor leaves the reservoir through an opening without any disruption to equilibrium, and the resulting flux approximately follows the Arrhenius behavior

$$v = k \exp^{-w/T}$$

where w and k are constants and T is the cell temperature. Noting that local PID loops drive cell temperatures to commanded values, we model the effusion cell by the relationship

$$v_i = \exp^{-w_i/u_i} \tag{3.2}$$

where k has been absorbed in the system model coefficient k_i, and w_i is a constant that is fit from prior growth data and depends on the material and particular cell being considered.

3.1.3 Production Processes

Epitaxial processes use different environments and growth conditions to induce the reactions driving deposition. This study focuses on a spectrum of processes that are closely related, in that each can operate in a particular growth regime, defined by specific choices of growth temperatures and pressures, resulting in dynamics described by the general model above 3.1. Although the underlying chemistry and thermodynamic drivers resulting in deposition are quite different between these processes, operating in the appropriate regime relegates these differences, at an operational level, to differences of actuation, either by gas sources (either metal-organic or hydride) or elemental sources (effusion cells).

Metal-Organic Chemical Vapor Deposition

MOCVD has been the epitaxy workhorse for industrial applications. The system uses gas sources composed of metal-organic bubblers for group III sources, and hydride gases for group V sources. There is no line-of-site requirement between the sources and the growing film, so gases are typically mixed in a central manifold before being introduced into the reaction chamber.

One of the main advantages of MOCVD is that the reaction chamber operates at atmospheric pressures, as opposed to a vacuum environment, which makes the machine easier to maintain and cheaper to operate. One of the key difficulties, however, is controlling the fluid dynamics in the system to ensure steady laminar flow over the wafer surface. Different techniques are used to accomplish this, including varying reactor geometries e.g. vertical chamber vs. horizontal chamber, cold-wall vs. warm wall environments, and screening input gases through a mesh grating to modulate flow.

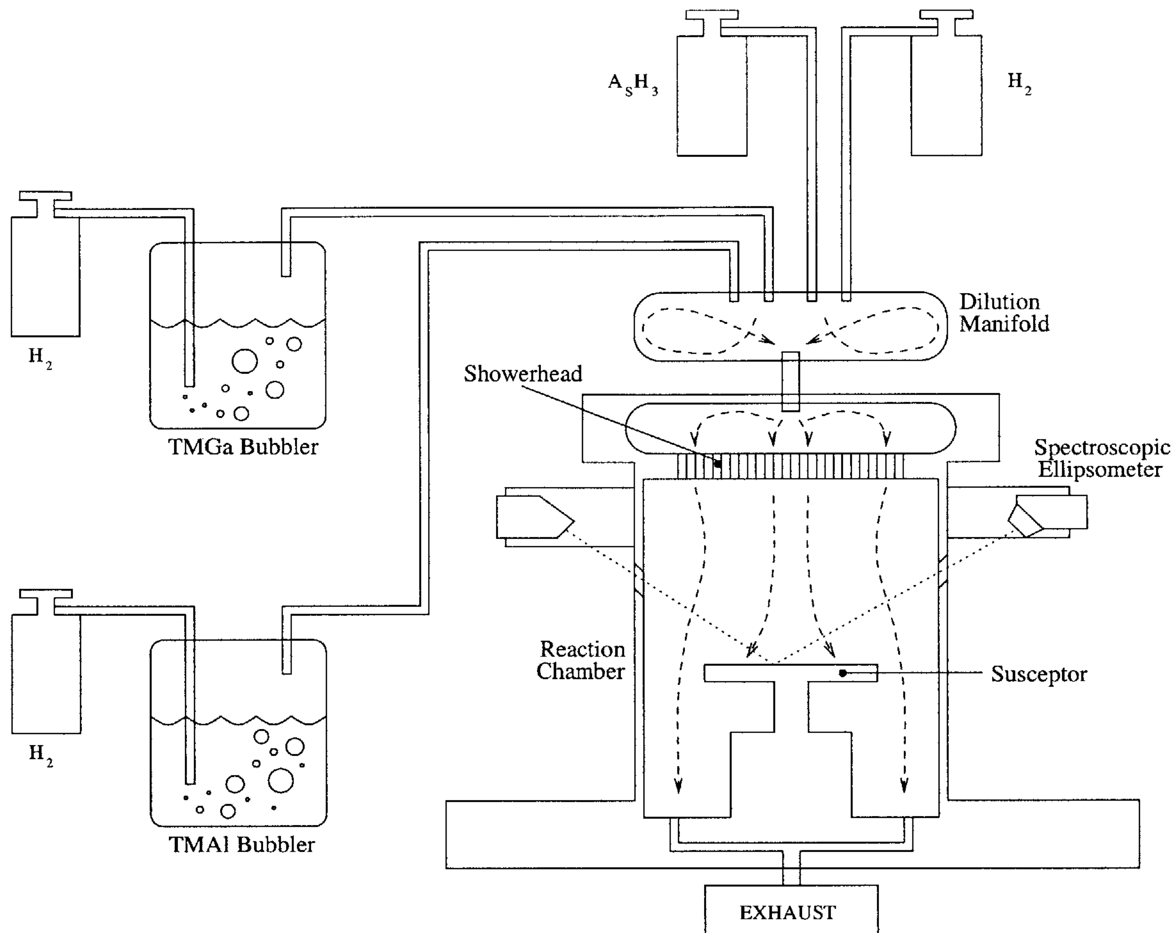


Figure 3-1: Vertical warm-wall MOCVD apparatus for AlGaAs, illustrating the geometry of sources, reaction chamber, and in-situ ellipsometer during growth.

Other difficulties include parasitic reactions, particularly in the warm-wall systems; problems making sharp transitions, especially if a large volume of plumbing separates the sources from the chamber; and trouble with impurities. Nevertheless, the ability to scale production to larger wafers and even batches of wafers create strong economic incentives to make MOCVD work well.

Chemical Beam Epitaxy

CBE brings together the gas-phase precursors of MOCVD with the vacuum environment of MBE. Advantages of the process include the ease of using flows as control inputs, the ability to leverage the vacuum environment to grow at lower temperatures (allowing, for example, greater thicknesses of strained material to be deposited

e.g. three times the number of quantum wells can be grown in the InAsP system compared to MOCVD, or growth with less dopant redistribution. Note that lower temperature growth is not always an advantage, as higher temperatures may yield better crystal quality.), and the ability to perform selective area epitaxy (depositing on exposed substrate while preventing growth on masked regions). Additional advantages include the ability to replenish sources without opening the vacuum chamber (as in MBE, which uses elemental sources), a significant reduction in material consumption (which impacts material costs in addition to the cost of disposing toxic wastes), and the ability to easily scale up production to multi-wafer systems. The downside of CBE include deposition purity, safety of its sources (compared with other vacuum techniques), and the complexity involved with operating and maintaining the vacuum growth environment.

Note that the vacuum processes are generally equipped with a Reflection High Energy Electron Diffraction (RHEED) sensor. RHEED uses an electron beam to measure the surface structure of a growing film by firing a beam of high energy electrons at grazing incidence on the surface of the growing film. The beam diffracts and reflects depending on the crystal structure, and a suitably located phosphor screen displays the results of this reflective diffraction. To the trained eye, this diffraction pattern reveals much about the surface structure.

Metal-Organic Molecular Beam Epitaxy

MOMBE replaces the hydride sources of CBE with elemental group V sources (Note this notation is not universal–sometimes MOMBE is used for all processes with any metal-organic source). The primary motivation for replacing these sources in CBE is to eliminate safety concerns working with arsine and phosphine. Purity is also typically a problem for this technique, as carbon contaminated growth is difficult to avoid [55].

Hydride Source Molecular Beam Epitaxy

HSMBE uses the hydride sources of CBE (thus retaining the safety problems associated with handling hydrides), but replaces the metal-organics with elemental group III sources. These sources, or effusion cells, evaporate elemental charges of the source material to form an atomic or molecular beam. A "line-of-sight" geometry from the source to substrate is thus essential (see Figure 3-2). Also, temperatures of the cells, which control the beam flux, become primary control points, replacing the convenience of flow control in CBE, MOMBE, or MOCVD. The key advantage of HSMBE over MOMBE, CBE, or MOCVD is the purity associated with the elemental sources. Moreover, HSMBE can effectively grow phosphorus compounds, in contrast to MOMBE or MBE, which are key to many optoelectronic devices. This combination of purity and applicability have made HSMBE a popular high-tier tool for research environments.

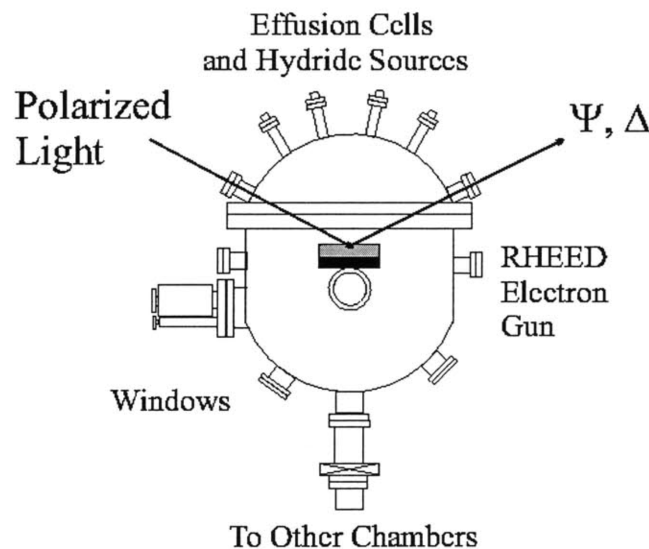


Figure 3-2: Elemental sources of a HSMBE system introduce line-of-sight constraints on reactor geometry and ellipsometer configuration (figure courtesy of Gale Petrich).

Molecular Beam Epitaxy

MBE replaces the hydride sources of HSMBE with elemental sources. This alleviates the danger of working with hydrides. The elemental sources introduce line-of-sight constraints on reactor geometry, but also deliver the highest purity and the ability to make very sharp transitions with shutter control.

3.1.4 Process Model

Appealing to 3.1, then, the process model for any of these processes becomes

$$\frac{\partial d}{\partial t} = |k_1 v_1 + k_2 v_2| \tag{3.3}$$

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \frac{k_5 v_1}{k_5 v_1 + k_6 v_2} \\ \frac{k_7 v_3}{k_7 v_3 + k_8 v_4} \end{bmatrix} \tag{3.4}$$

where

$$v_i = \begin{cases} u_i & \text{Gas Sources} \\ \exp -w_i/u_i & \text{Elemental Sources} \end{cases} \tag{3.5}$$

Although the parameters in this epitaxy model carry some physical meaning (e.g. k_i is related to the sticking coefficient of the ith source, w_i is related to the activation energy of the ith material, etc.), we will not presume our idealizations used to justify and explain the model to be accurate enough to predict the appropriate values of these parameters. Instead, we note that system operators typically keep a database of growth rates and compositions achieved with various process settings, and we propose to fit the values of these parameters with this historical data for the specific machine being considered. Note that this model ignores shutter operation, which will nevertheless play an important role in achieving sharp discontinuities between materials.

3.2 Growth Objective

The high quality crystalline films produced by epitaxy exhibit a particular periodicity in their atomic lattices, thereby generating an electrical band structure. This band structure is a function of both the structural properties of the lattice (size, shape, etc.) as well as the valence structure of the elemental atoms occupying sites throughout the lattice. Choosing alloy mixtures that approximate the same lattice structure but change the elemental components on lattice sites provides a mechanism for engineering film profiles with a desired band structure. It is the shaping of this band structure, as a function of film thickness, that results in devices with remarkable qualities and creates the incentive for high performance epitaxial processes.

This section defines the performance objective for epitaxial growth by first presenting an overview of band-gap engineering, motivating why a film with a certain composition profile may contribute to interesting electric or optical devices. A measure of film quality is then suggested, providing a mechanism to evaluate post-growth product quality. This measure motivates the performance criteria used to govern feedback design, and the relationship of this criteria to standard design specifications is then discussed.

3.2.1 Band-gap Engineering

Band-gap engineering uses the energy band structure of lattice structures to guide device design. This section will introduce this band model in the context of crystalline semiconductor materials, and show how various properties of such films may result in a desired band structure.

Semiconductors form a category of materials with electrical conductivity between that of conductors and insulators. As a result, the electrical properties of semiconductors can change significantly, becoming either more or less conducting as trace amounts of impurities are introduced. The physical structure of semiconductors in solid form range from amorphous, with no recognizable order, to polycrystalline, with segments of orderly arrays, to crystalline structures (Figure 3-3).

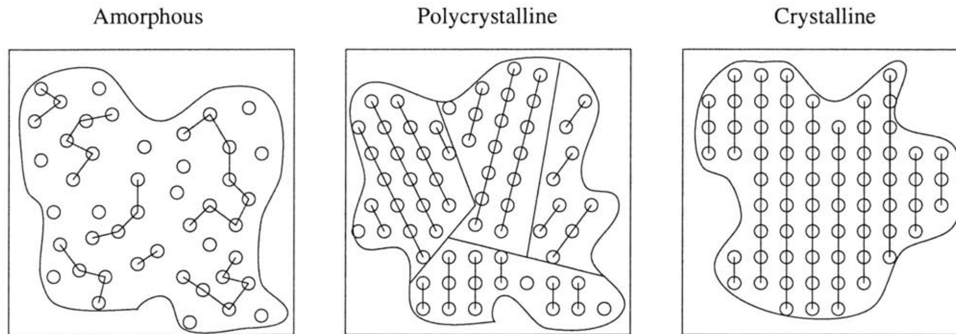


Figure 3-3: Classification of the physical structure of solids.

Crystalline solids form spatially periodic structures, with the type of lattice forming a single period being determined by the materials involved. Elemental semiconductors, such as silicon or germanium, form a "diamond lattice," so called because it characterizes the typical crystalline structure of Column IV elements such as carbon (see Figure 3-4). Compound semiconductors, such as those composed of elements from Columns III and V, typically form a "zincblende lattice" (see Figure 3-4). The cubic shape of these lattice structures suggest that its electrical properties would be isotropic in bulk. Moreover, both of these lattice structures exhibit chemical bonding dominated by the attraction between any atom and its four nearest neighbors.

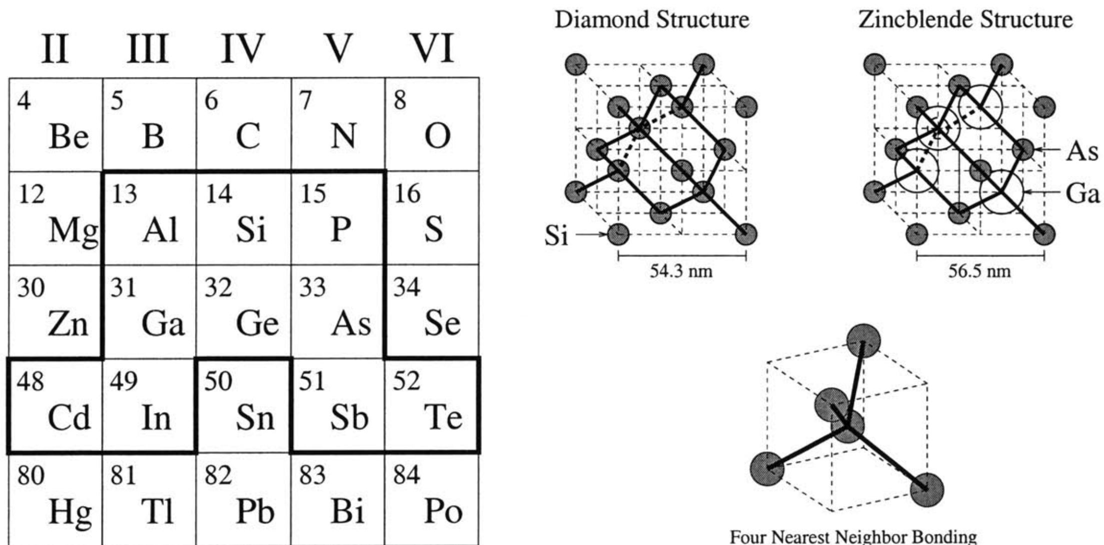


Figure 3-4: Columns II through VI of the periodic table, highlighting the most common materials used in semiconductor devices (left). Diamond and Zincblende lattice structures characterized by four-nearest-neighbor bonding (right).

This bonding between an atom and its four nearest neighbors leads to the "bonding model" of the lattice (Figure 3-5). This representation is a convenient tool to visualize lattice defects, impurities, or broken bonds resulting in free carriers for electrical conductivity. These carriers can either be free electrons or holes, the absence of a shared valence electron in the normal lattice structure. Holes may be harder to visualize as "particles", but they can be thought to move around the lattice as adjacent electrons jump in the hole leaving a vacancy elsewhere. Both of these carriers are important in defining the conductivity of the material, thus the bonding model facilitates a modest understanding of how lattice properties affect conductivity. For example, an increase in lattice temperature results in an increase in broken bonds as electrons become thermally excited. From the bonding model, we then can visualize how this will increase the number of free electrons and holes in the lattice, and predict a corresponding increase in conductivity. Such an increase in conductivity with temperature is an important characteristic of semiconductors that distinguishes them from other materials (metals, for example).

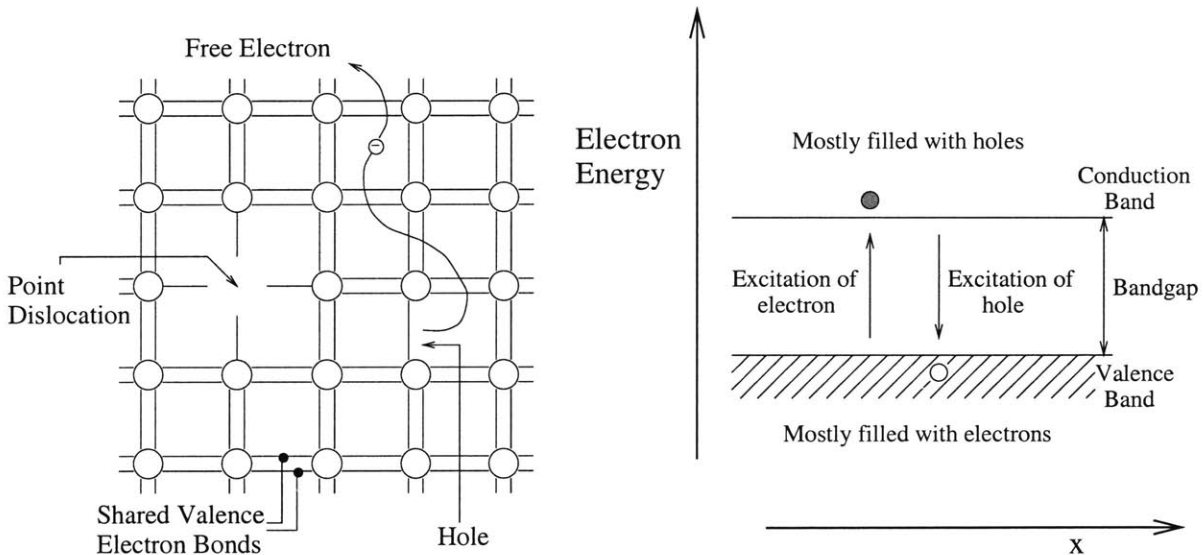


Figure 3-5: The Bond (left) and Band models (right) serve to visualize different features of the lattice.

A certain amount of energy is needed to free an electron from its lattice bond, though, depending on the chemistry of the materials composing the lattice structure. This fact leads to the (energy) band model of the lattice. In this model, energy of

carriers is plotted against one spatial direction (Figure 3-5). Such a plot is dominated by a gap of forbidden energies, i.e. energies that no carriers in the lattice can have, sandwiched between the *conduction band* (top) and *valence band* (bottom). The valence band represents the allowed electron states that serve as bonds in the lattice, while the conduction band represents the allowed higher-energy, or excited, electron states that would be occupied by a free electron. Thus, a perfect lattice approaching 0° K has a valence band completely filled with electrons (since electrons seek out and fill the lowest available energy states) and a conduction band completely filled with holes (since holes seek out and fill the highest available energy states). As energy is introduced to the lattice, though, the lattice absorbs the energy until enough energy is introduced to break a bond. When this happens an electron is freed, introducing an electron and hole pair in the lattice. In the band model we represent this event by showing an electron jumping from the valence band to the conduction band, and a hole jumping from the conduction band to the valence band. The amount of energy needed to break the bond defines the width of the *bandgap*, which is one of the most important properties of a material. Conductors, in fact, turn out to simply be materials with bandgaps that are either nonexistent (overlapping valence and conduction bands) or very narrow (yielding a large number of free carriers at room temperature), while insulators are wide gap materials. The band model becomes a useful tool, then, for determining the action of carriers over available energy states.

The idea of device design, then, is to shape the bands in ways that will induce carrier dynamics leading to the desired device characteristics. Since the band model only considers one spatial dimension, most simple devices engineered from the band model are one dimensional i.e. they are composed of uniform films deposited sequentially with variation in only one direction (film thickness). This allows the band shape to vary as different materials are deposited on the lattice substrate. Of course, the band model may also play a key role in the design of more complicated structures, and it is usually the primary design tool for the epitaxial processing step in any device.

The important task of bandgap engineering, then, is to find materials with compatible lattice structures and the necessary range of bandgaps, and then determine

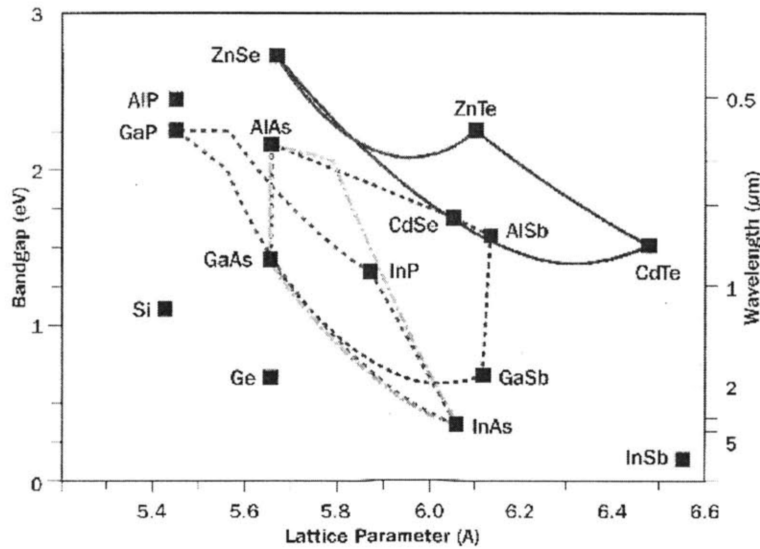


Figure 3-6: Lattice constant versus bandgap for a variety of material systems.

a growing sequence from these materials that will result in the desired band shape. Figure 3-6 shows the lattice constants, or lengths of a single period in the lattice structures, and the associated bandgaps for a variety of compound semiconductor materials and their alloys. Materials should be lattice matched, or have approximately the same spacing between atomic sites of the crystalline structure, to avoid serious defects in the crystal quality of the resulting structure. Doping profiles also play a central role in shaping the band structure and must be considered in the device design. Each of these considerations play an important role in determining a particular "recipe," or desired profiles, to be delivered by epitaxy.

3.2.2 Measuring Film Quality

As discussed above, there are many aspects of the resulting film that matter critically to the performance of a particular device, including wafer uniformity, purity and doping, and crystal quality. Nevertheless, this study is focused on the specific problem of delivering a film with the desired compositional profile as a function of thickness.

In general, material composition of a fixed film of thickness d is a vector function

given by

$$y(z) = \begin{bmatrix} y_1(z) \\ y_2(z) \end{bmatrix} \quad 0 \leq z \leq d,$$

where $y_i(z) : [0\ d] \rightarrow [0\ 1]$, $i = 1,\ 2$. The range of each component function is bounded on the unit interval since composition represents the percentage, on average, of available lattice sites filled by a particular species. The vector index indicates different components of the lattice that are available for mixing. For example, a quaternary alloy $In_{y1(z)}Ga_{1-y1(z)}As_{y2(z)}P_{1-y2(z)}$ has composition $(y_1(z),\ y_2(z))$, meaning that the percentage of indium atoms occupying group III lattice sites, averaged over the planar surface at the height z, is $y_1(z)$. Likewise, the percentage of arsine atoms occupying the group V lattice sites, averaged over the plane at that height, is $y_2(z)$. For ternary alloys, such as $Al_{y(z)}Ga_{1-y(z)}As$, the composition vector has only one component since there is no mixing, in this case, on the group V lattice sites.

Once a film is grown, we can evaluate its quality by comparing its composition profile $y(z)$, $0 \leq z \leq d$, to a reference profile, $y^{ref}(z)$, $0 \leq z \leq d^{ref}$, characterizing the desired structure. Making this comparison in a meaningful way can be difficult, however, since the film thicknesses may differ. We propose the following measure as a metric of quality

$$e = \left\| W_1(z) \left(y^{ref}(z) - y(z) \right) \right\|_{\mathcal{L}2,[0\ min(d,d^{ref})]} + W_2 |d^{ref} - d|,$$

where $W_1(z)$, a positive definite matrix for every z, and the scalar W_2 are weights that would depend on the particular device. This measure computes the weighted 2-norm through the height of the shortest film, and then penalizes any difference in thickness between the actual and desired profiles.

To justify this choice of quality metric, consider the multiple quantum well laser structure [40] shown in Figure 3-7. Slight deviations in thickness of the cladding layers will have minimal impact on device performance, so W_1 can be made small near those interfaces. On the other hand, slight deviations in thickness or composition in the quantum wells would critically effect the band structure and resulting emission

wavelength, so W_1 could be made very large over that range. The accumulated error, then, over the height of the film would seem to capture a relevant notion of performance. Contrasting this with a sup-norm metric, which would be triggered by slight errors near the discontinuous interfaces but then be blind to other deviations over the profile, the integral affect of this metric does seem more relevant to device performance.

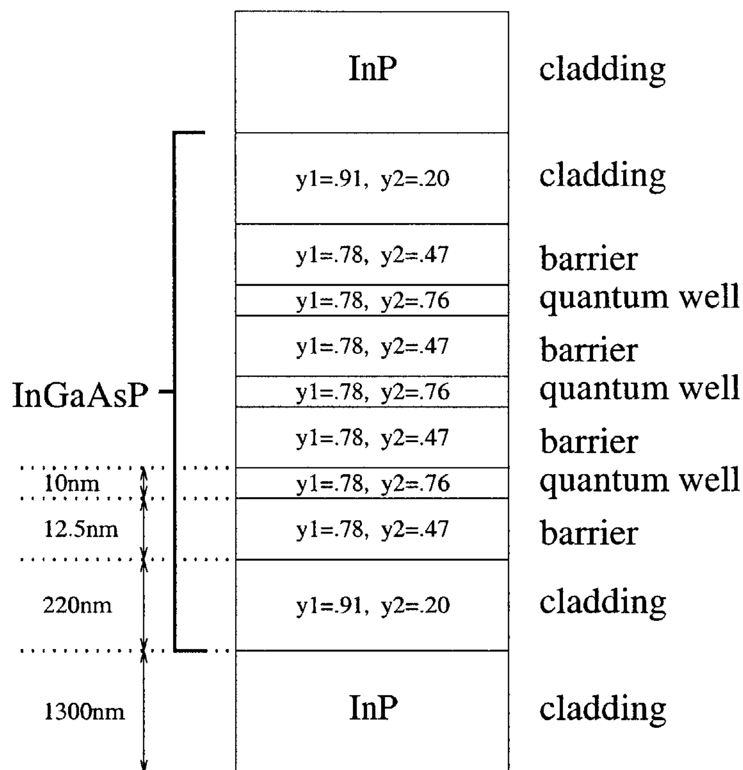


Figure 3-7: Desired composition profile for a multiple quantum well laser design.

Nevertheless, this measure of product quality does fail to be meaningful in two situations. First, sharp spikes in a composition profile would fail to register as a large error even though they could be devastating to device performance. Such spikes do not occur in general, but unsynchronized shutter operation in MBE processes can lead to mono-layer jumps near profile discontinuities. We thus assume that any shuttering mechanism employed is well timed to eliminate this possibility. The second way this measure would fail to be meaningful is if constant composition profiles did not converge to their desired values. In this case, a small error in composition can

persist without registering a significant quality error for the resulting film. To prevent these persistent errors, we require asymptotic convergence of composition profiles to their desired values as a necessary performance criterion. Under the assumption that spikes will not occur, and with the asymptotic-tracking design specification, the 2-norm metric with design-specific weights does capture a meaningful measure of film quality.

3.2.3 Performance Criterion

Establishing a post-growth measure of film quality motivated the need for asymptotic convergence of composition profiles as a performance criterion. This criterion can be viewed as the dominant feedback design specification, requiring that

$$y(d) \rightarrow y^{ref}(d) \text{ as } d \rightarrow \infty.$$

Nevertheless, this criterion is not a typical tracking specification since the growth function, $d(t)$, is itself a controlled variable. Abusing notation to associate $y(t)$ as the functional composition of $y(d)$ and $d(t)$, we note that the epitaxial process generates $d(t)$ and $y(t)$ as outputs. This leads to the observation that the performance criterion

$$y(d) \rightarrow y^{ref}(d) \text{ as } d \rightarrow \infty \quad \Rightarrow \quad y(t) \rightarrow y^{ref}(d(t)) \text{ as } t \rightarrow \infty$$

can be viewed as the *synchronization* of $y(t)$ with $d(t)$, as specified by y^{ref}.

This synchronization criterion reduces to a typical tracking specification if $d(t)$ is fixed. Having the degrees of freedom to independently control $d(t)$, however, implies that this synchronization criterion is a relaxation of the standard tracking specification. Driving $d(t)$ in particular ways can exploit this extra degree of freedom to improve the system's response to high performance maneuvers. For example, Figure 3-8 demonstrates how slowing growth for thin layers, such as quantum wells, can result in an easier profile to track in real time.

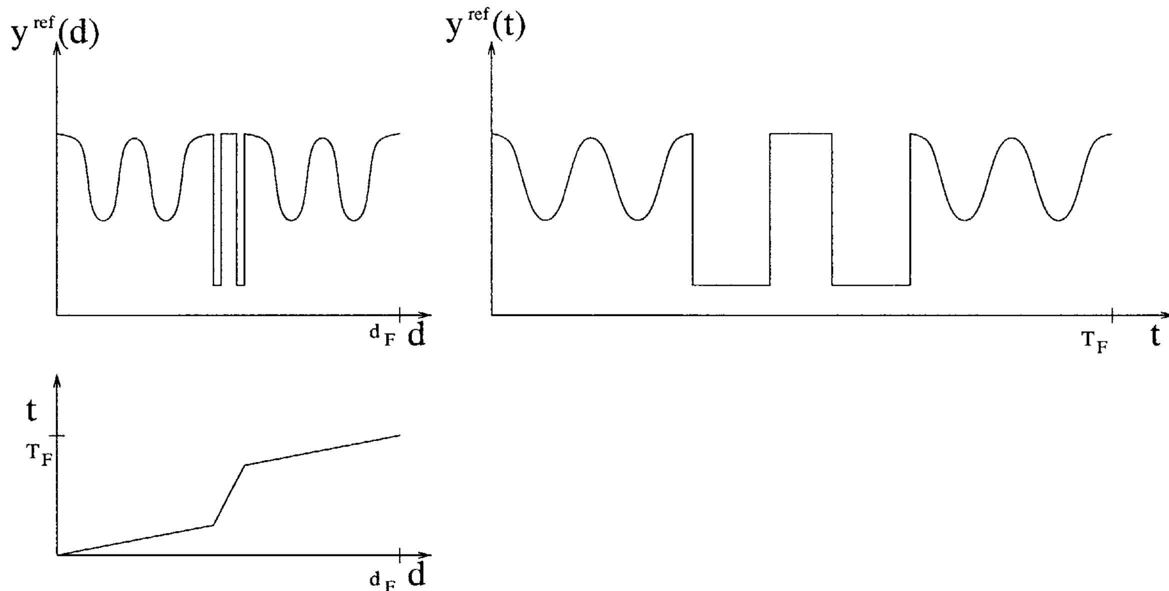


Figure 3-8: Synchronization as a performance objective allows one to choose different growth-time profiles d(t) to make difficult maneuvers easier to track.

3.3 Tracking Graded Quaternaries

Equipped with models for the epitaxy process and the ellipsometric sensor, and with an understanding of the performance objective and relevant measures of quality, we are now in a position to formulate a feedback design problem. This section discusses a formulation and outlines the synthesis steps leading to a working design and closed-loop simulation results. These results suggest that currently ungrowable structures, such as graded quaternary films, can be realized using ellipsometric feedback.

3.3.1 Problem Formulation

A number of issues make formulating a solvable problem challenging for this application. First, manipulating the synchronization criterion directly requires knowledge of the film thickness, so we present an indirect tracking criterion. Next, we show the relationship between convergence of material composition vs. convergence of a film's optical constants, highlighting that measurements at multiple wavelengths may be necessary for some material systems. Finally, we discuss necessary conditions for achieving the desired tracking performance and propose a subsequent design problem

that meaningfully captures the growth objective for epitaxy.

From Synchronization to Tracking

The first difficulty in formulating a design problem emerges from the dependence of the synchronization objective

$$y(t) \to y^{ref}(d(t))$$

on knowledge of $d(t)$. Two approaches we consider for addressing this issue include building a thickness estimator using available measurements, or using a tracking objective as a proxy for synchronization.

Building a thickness estimator involves constructing a dynamic system $\hat{S}$ with state $\hat{d}$ such that the thickness estimate $\hat{d}(t)$ converges to the actual film thickness $d(t)$, regardless of the values of other variables in the system. Recall that a standard state estimator assumes structural information about the system, such as A, B, C, D matrices for a linear system, is known, while the initial state value is unknown. The estimator is then constructed such that the estimation error, $e = x - \hat{x}$, where x represents the system state and $\hat{x}$ represents the state estimate, has stable dynamics, implying that the error approaches zero. This estimation problem differs considerably, however, since the initial thickness is always known to be zero. Thus, the need for estimation would imply that the structural information governing the evolution of $\dot{d}$ is what is unknown. Estimating with a lack of structural information is a robustness problem, and we don't expect to recover the same kinds of results under such circumstances. In particular, if the system inputs are persistent, then we can't expect our error estimates to converge to zero since the driving terms can't be exactly canceled. Thus, the applicability of a thickness estimator may be limited.

In the absence of a thickness estimate, we use our best guess of the current thickness. This guess is the value we compute from our growth model driven by the known system inputs, and we note that any errors in our model will result in errors in the thickness guess. Nevertheless, in the absence of more information, such an approach

is reasonable in that experiments can later verify whether this "best guess" is good enough, or if independent thickness-oriented sensor technology is necessary.

The degrees of freedom of this problem, however, suggest that the growth function, $d(t)$ can be driven independently from material compositions. This means that we can select an arbitrary reference growth function $d^{ref}(t)$ and choose control inputs that will drive our model along this desired trajectory while leaving control authority to compensate for errors in composition. In principle, this desired growth rate can be adjusted in real-time based on the additional information acquired through measurements. By selecting the desired growth function a priori, however, the synchronization problem reduces to a standard tracking problem, given by

$$y(t) \rightarrow y^{ref}(d^{ref}(t)),$$

where we abuse notation to associate $y^{ref}(d^{ref}(t))$ with $y^{ref}(t)$ to write

$$y(t) \rightarrow y^{ref}(t).$$

The Optical Function Database

A critical link between performance, given by the structure of a film's composition profile, and our real-time measurements is the optical function that relates a specified composition to its complex permittivity, or dielectric function, ϵ, at a particular wavelength. Our assumptions about the nature of growing films modeled the dielectric function of a homogeneous layer as being independent of the film thickness. Thus, the dielectric function at the interface of two distinct media would rapidly transition from the "bulk" value of the first medium to the "bulk" value of the second medium. In the event that this rapid transition is simply a discontinuous jump, the necessary information to relate measurements to performance is a database of the bulk dielectric function for all possible alloys within the relevant material system. Strain resulting from lattice mismatch can also affect a material's dielectric function, suggesting that the optical constants depend not only on the composition of the current film but

also on the material substructure. Nevertheless, we will ignore these second order affects and consider the map from composition to ϵ to be a function represented by an empirical database.

Although such databases are readily available for ternary systems like AlGaAs or InGaAs, collecting this information for quaternary systems has been more difficult. A first approximation to such a quaternary database, however, is the linear interpolation of the complex permittivities of the relevant binary constituents. Thus, for a quaternary compound given by (y_1, y_2), where $\mathrm{A}_{y1}\mathrm{B}_{1-y1}\mathrm{C}_{y2}\mathrm{D}_{1-y2}$, we assume the values of the binary constituents are known to be ϵ_{AC},ϵ_{AD},ϵ_{BC}, and ϵ_{BD}. The complex permittivity of the alloy given by (y_1, y_2) then becomes

$$\epsilon_{\mathrm{A}_{y1}\mathrm{B}_{1-y1}\mathrm{C}_{y2}\mathrm{D}_{1-y2}} = y_1 y_2 \epsilon_{AC} + y_1 (1 - y_2)\, \epsilon_{AD} + (1 - y_1)\, y_2 \epsilon_{BC} + (1 - y_1)(1 - y_2)\, \epsilon_{BD}.$$

It can be shown that driving the ellipsometer equations with any value of ϵ in this region of interest yields a unique value of $\boldsymbol{\rho}$ as the film becomes optically thick. That is, there is a one-to-one correspondence between each admissible value of the complex permittivity and the associated equilibrated ellipsometric measurement. This bijective relationship is important because it suggests that we can map any criteria of interest defined on ϵ into a corresponding criterion on $\boldsymbol{\rho}$.

We note, however, that a such a bijective correspondence does not exist between alloy composition, (y_1, y_2), and ϵ for this material system. Figure 3-9 demonstrates how the complex permittivities fold onto themselves, yielding multiple compositions with the same permittivity. Note that the relative permittivity is related to permittivity by $\epsilon_{relative} = \epsilon/\epsilon_o$. The fold occurs in different places for different wavelengths, however, suggesting that a one-to-one correspondence can be established between a particular alloy characterized by its composition, (y_1, y_2), and a *set* of permittivities collected over multiple wavelengths.

With a bijective relationship between y and a vector of permittivities collected at different wavelengths, and between this vector ϵ and $\boldsymbol{\rho}$, performance criteria, such as asymptotic convergence, defined on composition map directly into associated criteria

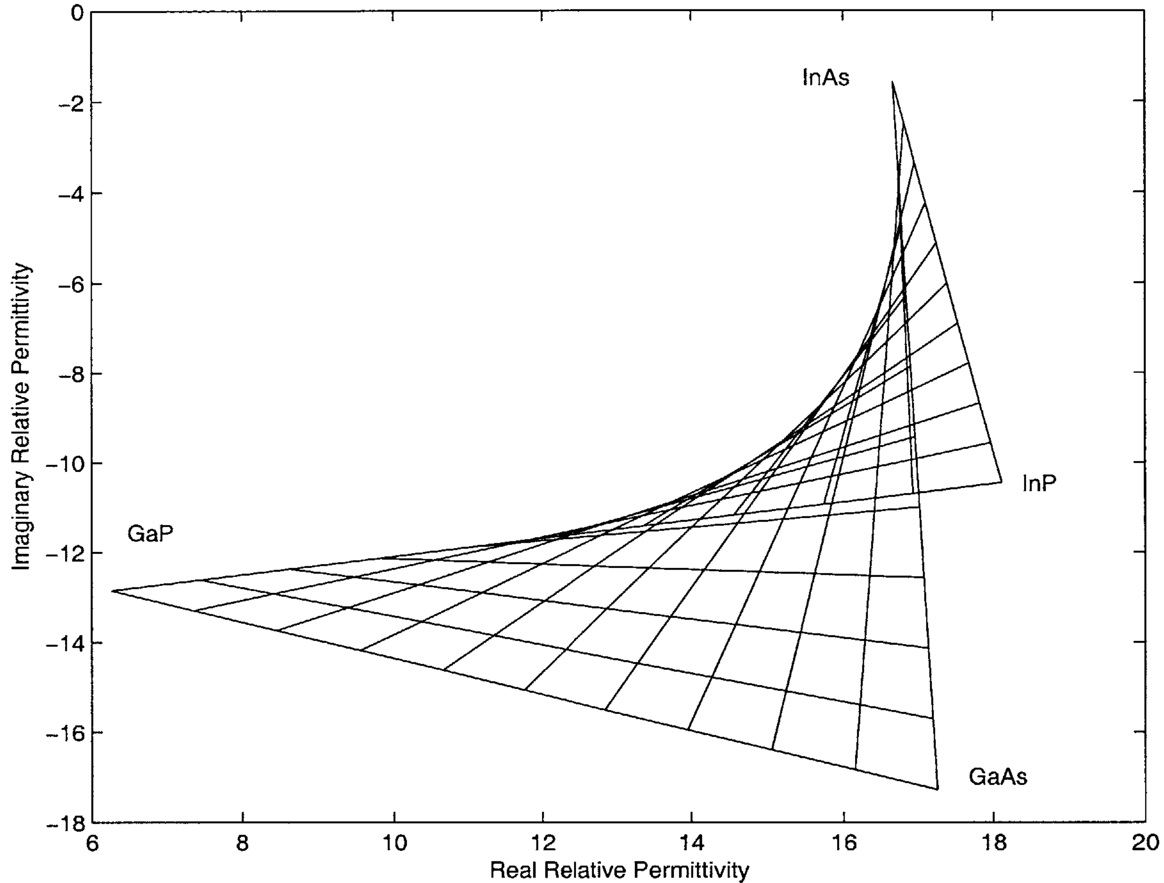


Figure 3-9: Different compositions can yield the same dielectric constants due to a "fold" in the map from composition to bulk permittivity. Maps for different wavelengths fold in different places, however, indicating that each composition maps to a unique set of permittivities collected over multiple wavelengths.

defined on $\boldsymbol{\epsilon}$ or $\boldsymbol{\rho}$. This flexibility will be important in formulating a meaningful tracking problem.

Generating Tracking Error

Driving material composition to track the desired reference command demands that we explicitly construct the tracking error within the feedback loop. Constructing this error, however, is challenging since we measure $\boldsymbol{\rho}$, not y. One approach to resolving this dilemma would be to approximately invert the sensor map, E, from y to $\boldsymbol{\rho}$. Advantages to this approach include the fact the the remaining controller design only needs to consider the process dynamics, which are relatively simple compared to the

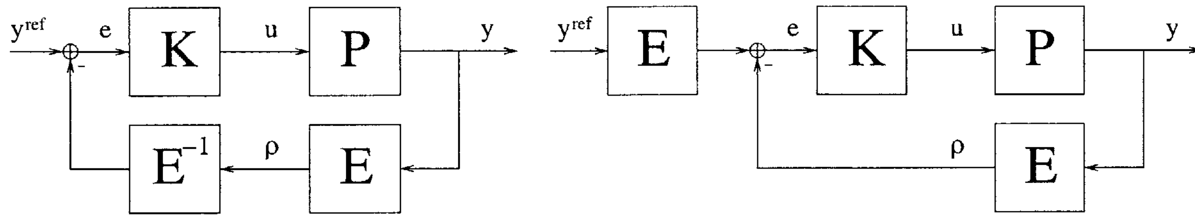


Figure 3-10: Alternate approaches for generating tracking error, where P represents the growth process, E represents the sensor dynamics, and K is the feedback algorithm responsible for delivering the desired performance.

sensor dynamics. We noted in earlier analysis, however, that since the zero dynamics of E are not asymptotically stable, such an inversion would render an unstable mode unobservable. No feedback system K could then stabilize this mode. Although such instability may not have practical consequences for some applications, we believe it is important to understand the alternatives to this approach.

Another approach would be to use our model of the sensor dynamics to translate the reference composition trajectory into a corresponding reference $\boldsymbol{\rho}$ trajectory, and then compare this desired trajectory with the actual measurement. This approach is viable only because of the bijective relationship between y and equilibrated $\boldsymbol{\rho}$; otherwise the $\boldsymbol{\rho}$-error may go to zero even though y may not converge to y^{ref}. Although constructing $\boldsymbol{\rho}$-error resolves the instability problem associated with inverting E, the process of translating y^{ref} into a reference $\boldsymbol{\rho}$ trajectory is sensitive to errors in the sensor model, E. That is, since the translation occurs outside of the feedback loop, no feedback system K could compensate for errors in the map E. Moreover, the feedback design for K becomes more difficult as E must be considered explicitly in the design.

In general, given a factorization of the sensor map given by E = MN, we can generate a meaningful comparison by generating the error given by

$$e = \mathrm{N}y^{ref} - \mathrm{M}^{-1}\boldsymbol{\rho},$$

provided M^{-1} exists. The two alternatives discussed above consider the extreme cases $(\mathrm{M} = \mathrm{E}, \mathrm{N} = \mathrm{I})$ and $(\mathrm{M} = \mathrm{I}, \mathrm{N} = \mathrm{E})$. Another possibility would be to let N be the optical function mapping composition to permittivity, and let M then be the map

from permittivity to $\boldsymbol{\rho}$. This decomposition would remove the optical function from feedback while inverting the ellipsometer dynamics. The part of E that we would most like to have within the feedback loop, however, is the optical function, since that is where we expect the most errors in our sensor model. Likewise, the part of E that we would most like to forward is the dynamics associate with the ellipsometer, mapping permittivity to $\boldsymbol{\rho}$. Unfortunately, we can not factor E to meet both these objectives simultaneously.

For the design problem in this study, we decided to avoid the stability problems caused by deploying E^{-1} and instead risked sensitivity to errors in E by generating tracking error from forwarding the reference command through our sensor model. We also interpreted the tracking criteria to be tracking to step commands, noting that $\boldsymbol{\rho} \rightarrow \boldsymbol{\rho}^{ref}$ implies that $y \rightarrow y^{ref}$ excepting errors in E. The problem formulation then becomes as follows.

Design Problem: *Given a reference trajectory $\boldsymbol{\rho}^{ref}(t)$, and the system* EP : u $\rightarrow \boldsymbol{\rho}$ *characterized by*

$$\frac{\partial d}{\partial t} = |k_1 e^{\frac{-w_1}{u_1}} + k_2 e^{\frac{-w_2}{u_2}}| \tag{3.6}$$

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \frac{k_5 e^{\frac{-w_1}{u_1}}}{k_5 e^{\frac{-w_1}{u_1}} + k_6 e^{\frac{-w_2}{u_2}}} \\ \frac{k_7 u_3}{k_7 u_3 + k_8 u_4} \end{bmatrix} \tag{3.7}$$

$$\boldsymbol{\epsilon} = y_1 y_2 \boldsymbol{\epsilon}_{AC} + y_1 (1 - y_2) \boldsymbol{\epsilon}_{AD} + (1 - y_1) y_2 \boldsymbol{\epsilon}_{BC} + (1 - y_1)(1 - y_2) \boldsymbol{\epsilon}_{BD} \tag{3.8}$$

$$\frac{\partial x_1}{\partial t} = \left[f_1 x_1 + g_1 x_1 \boldsymbol{\epsilon} + h_1(x_1) \frac{1}{\boldsymbol{\epsilon}} \right] \frac{\partial d}{\partial t} \tag{3.9}$$

$$\frac{\partial x_2}{\partial t} = [f_2 x_2 + g_2 x_2 \boldsymbol{\epsilon}] \frac{\partial d}{\partial t} \tag{3.10}$$

$$\boldsymbol{\rho} = \frac{x_1}{x_2} \tag{3.11}$$

$$e = \boldsymbol{\rho}^{ref} - \boldsymbol{\rho} \tag{3.12}$$

design a feedback controller K : e $\rightarrow$ u *such that the closed loop system is stable and* $e \rightarrow 0$ *for constant inputs* $\boldsymbol{\rho}^{ref}$.

3.3.2 Linearization

The pronounced nonlinearities in the system EP, along with the multi-input multi-output nature of the controller, make the design problem difficult. We attempt to mitigate much of this difficulty by linearizing the plant around a nominal input and its associated equilibrium point. Although the procedure is standard, it is included here for the sake of completeness.

Before linearizing the system per se, however, we first note that since some of the system variables are complex-valued, while others are real-valued, it may conceptually simplify the description to convert all system variables to real-valued quantities. This is accomplished by defining two real-valued quantities for every complex-valued variable, as $x_r + ix_i = x$, where x_r and x_i are real-valued variables. The ellipsometer dynamics are then given by:

$$\begin{aligned}\frac{\partial x_{1r}}{\partial t} = & \Big[f_{1r}(x_{1r}, x_{1i}) + g_{1r}(x_{1r}, x_{1i})\epsilon_r(y) - g_{1i}(x_{1r}, x_{1i})\epsilon_i(y) + \\ & \frac{h_{1r}(x_{1r},x_{1i})\epsilon_r(y)}{\epsilon_r^2(y)+\epsilon_i^2(y)} + \frac{h_{1i}(x_{1r},x_{1i})\epsilon_i(y)}{\epsilon_r^2(y)+\epsilon_i^2(y)}\Big] \frac{\partial d}{\partial t}\end{aligned} \tag{3.13}$$

$$\begin{aligned}\frac{\partial x_{1i}}{\partial t} = & \Big[f_{1i}(x_{1r}, x_{1i}) + g_{1i}(x_{1r}, x_{1i})\epsilon_r(y) + g_{1r}(x_{1r}, x_{1i})\epsilon_i(y) + \\ & \frac{h_{1i}(x_{1r},x_{1i})\epsilon_r(y)}{\epsilon_r^2(y)+\epsilon_i^2(y)} - \frac{h_{1r}(x_{1r},x_{1i})\epsilon_i(y)}{\epsilon_r^2(y)+\epsilon_i^2(y)}\Big] \frac{\partial d}{\partial t}\end{aligned} \tag{3.14}$$

$$\frac{\partial x_{2r}}{\partial t} = [f_{2r}(x_{2r}, x_{2i}) + g_{2r}(x_{2r}, x_{2i})\epsilon_r(y) - g_{2i}(x_{2r}, x_{2i})\epsilon_i(y)] \frac{\partial d}{\partial t} \tag{3.15}$$

$$\frac{\partial x_{2i}}{\partial t} = [f_{2i}(x_{2r}, x_{2i}) + g_{2i}(x_{2r}, x_{2i})\epsilon_r(y) + g_{2r}(x_{2r}, x_{2i})\epsilon_i(y)] \frac{\partial d}{\partial t} \tag{3.16}$$

$$\rho_r = \frac{x_{1r}x_{2r} + x_{1i}x_{2i}}{x_{2r}^2 + x_{2i}^2} \tag{3.17}$$

$$\rho_i = \frac{x_{1i}x_{2r} - x_{1r}x_{2i}}{x_{2r}^2 + x_{2i}^2} \tag{3.18}$$

and the decomposition of ϵ becomes

$$\epsilon_r = y_1y_2\Re e\{\epsilon_A\} + y_1(1-y_2)\Re e\{\epsilon_B\} + (1-y_1)y_2\Re e\{\epsilon_C\} + (1-y_1)(1-y_2)\Re e\{\epsilon_D\} \tag{3.19}$$

$$\epsilon_i = y_1 y_2 \Im m\{\epsilon_A\} + y_1(1-y_2)\Im m\{\epsilon_B\} + (1-y_1)y_2 \Im m\{\epsilon_C\} + (1-y_1)(1-y_2)\Im m\{\epsilon_D\}. \tag{3.20}$$

Linearizing these equations around (x_o, u_o) lead to a system of the form

$$\frac{\partial \delta x}{\partial t} = A\delta x + B\delta u$$

$$\rho = C\delta x$$

where $u = u_o + \delta u$, $x = x_o + \delta x$, and (A, B, C) are 4×4, 4×4, 2×4 matrices, respectively, with entries:

$$A_{11} = \left[\frac{\partial f_{2r}}{\partial x_{2r}} + \frac{\partial g_{2r}}{\partial x_{2r}}\epsilon_r - \frac{\partial g_{2i}}{\partial x_{2r}}\epsilon_i + \frac{\partial h_{2r}}{\partial x_{2r}}\frac{\epsilon_r}{\epsilon_r^2+\epsilon_i^2} - \frac{\partial h_{2i}}{\partial x_{2r}}\frac{\epsilon_i}{\epsilon_r^2+\epsilon_i^2}\right]\frac{\partial d}{\partial t} \tag{3.21}$$

$$A_{12} = \left[\frac{\partial f_{2r}}{\partial x_{2i}} + \frac{\partial g_{2r}}{\partial x_{2i}}\epsilon_r - \frac{\partial g_{2i}}{\partial x_{2i}}\epsilon_i + \frac{\partial h_{2r}}{\partial x_{2i}}\frac{\epsilon_r}{\epsilon_r^2+\epsilon_i^2} - \frac{\partial h_{2i}}{\partial x_{2i}}\frac{\epsilon_i}{\epsilon_r^2+\epsilon_i^2}\right]\frac{\partial d}{\partial t} \tag{3.22}$$

$$A_{13} = 0 \tag{3.23}$$

$$A_{14} = 0 \tag{3.24}$$

$$A_{21} = \left[\frac{\partial f_{2i}}{\partial x_{2r}} + \frac{\partial g_{2i}}{\partial x_{2r}}\epsilon_r + \frac{\partial g_{2r}}{\partial x_{2r}}\epsilon_i + \frac{\partial h_{2i}}{\partial x_{2r}}\frac{\epsilon_r}{\epsilon_r^2+\epsilon_i^2} - \frac{\partial h_{2r}}{\partial x_{2r}}\frac{\epsilon_i}{\epsilon_r^2+\epsilon_i^2}\right]\frac{\partial d}{\partial t} \tag{3.25}$$

$$A_{22} = \left[\frac{\partial f_{2i}}{\partial x_{2i}} + \frac{\partial g_{2i}}{\partial x_{2i}}\epsilon_r + \frac{\partial g_{2r}}{\partial x_{2i}}\epsilon_i + \frac{\partial h_{2i}}{\partial x_{2i}}\frac{\epsilon_r}{\epsilon_r^2+\epsilon_i^2} - \frac{\partial h_{2r}}{\partial x_{2i}}\frac{\epsilon_i}{\epsilon_r^2+\epsilon_i^2}\right]\frac{\partial d}{\partial t} \tag{3.26}$$

$$A_{23} = 0 \tag{3.27}$$

$$A_{24} = 0 \tag{3.28}$$

$$A_{31} = 0 \tag{3.29}$$

$$A_{32} = 0 \tag{3.30}$$

$$A_{33} = \left[\frac{\partial f_{1r}}{\partial x_{1r}} + \frac{\partial g_{1r}}{\partial x_{1r}}\epsilon_r - \frac{\partial g_{1i}}{\partial x_{1r}}\epsilon_i\right]\frac{\partial d}{\partial t} \tag{3.31}$$

$$A_{34} = \left[\frac{\partial f_{1r}}{\partial x_{1i}} + \frac{\partial g_{1r}}{\partial x_{1i}}\epsilon_r - \frac{\partial g_{1i}}{\partial x_{1i}}\epsilon_i\right]\frac{\partial d}{\partial t} \tag{3.32}$$

$$A_{41} = 0 \tag{3.33}$$

$$A_{42} = 0 \tag{3.34}$$

$$A_{43} = \left[\frac{\partial f_{1i}}{\partial x_{1r}} + \frac{\partial g_{1i}}{\partial x_{1r}}\epsilon_r + \frac{\partial g_{1r}}{\partial x_{1r}}\epsilon_i\right]\frac{\partial d}{\partial t} \tag{3.35}$$

$$A_{44} = \left[\frac{\partial f_{1i}}{\partial x_{1i}} + \frac{\partial g_{1i}}{\partial x_{1i}}\epsilon_r + \frac{\partial g_{1r}}{\partial x_{1i}}\epsilon_i\right]\frac{\partial d}{\partial t} \tag{3.36}$$

$$\begin{aligned} B_{11} \;=\; & \Big[g_{2r}(x_{2r},x_{2i})\tfrac{\partial \epsilon_r}{\partial y_1}\tfrac{\partial y_1}{\partial u_1} - g_{2i}(x_{2r},x_{2i})\tfrac{\partial \epsilon_i}{\partial y_1}\tfrac{\partial y_1}{\partial u_1} + \\ & h_{2r}(x_{2r},x_{2i})\left(\tfrac{1}{\epsilon_r^2+\epsilon_i^2}\tfrac{\partial \epsilon_r}{\partial y_1} - \tfrac{2\epsilon_r}{\left(\epsilon_r^2+\epsilon_i^2\right)^2}\left[\epsilon_r\tfrac{\partial \epsilon_r}{\partial y_1} + \epsilon_i\tfrac{\partial \epsilon_i}{\partial y_1}\right]\right)\tfrac{\partial y_1}{\partial u_1} + \\ & h_{2i}(x_{2r},x_{2i})\left(\tfrac{1}{\epsilon_r^2+\epsilon_i^2}\tfrac{\partial \epsilon_i}{\partial y_1} - \tfrac{2\epsilon_i}{\left(\epsilon_r^2+\epsilon_i^2\right)^2}\left[\epsilon_r\tfrac{\partial \epsilon_r}{\partial y_1} + \epsilon_i\tfrac{\partial \epsilon_i}{\partial y_1}\right]\right)\tfrac{\partial y_1}{\partial u_1}\Big]\tfrac{\partial d}{\partial t} \end{aligned} \tag{3.37}$$

$$\begin{aligned} B_{12} \;=\; & \Big[g_{2r}(x_{2r},x_{2i})\tfrac{\partial \epsilon_r}{\partial y_1}\tfrac{\partial y_1}{\partial u_2} - g_{2i}(x_{2r},x_{2i})\tfrac{\partial \epsilon_i}{\partial y_1}\tfrac{\partial y_1}{\partial u_2} + \\ & h_{2r}(x_{2r},x_{2i})\left(\tfrac{1}{\epsilon_r^2+\epsilon_i^2}\tfrac{\partial \epsilon_r}{\partial y_1} - \tfrac{2\epsilon_r}{\left(\epsilon_r^2+\epsilon_i^2\right)^2}\left[\epsilon_r\tfrac{\partial \epsilon_r}{\partial y_1} + \epsilon_i\tfrac{\partial \epsilon_i}{\partial y_1}\right]\right)\tfrac{\partial y_1}{\partial u_2} + \\ & h_{2i}(x_{2r},x_{2i})\left(\tfrac{1}{\epsilon_r^2+\epsilon_i^2}\tfrac{\partial \epsilon_i}{\partial y_1} - \tfrac{2\epsilon_i}{\left(\epsilon_r^2+\epsilon_i^2\right)^2}\left[\epsilon_r\tfrac{\partial \epsilon_r}{\partial y_1} + \epsilon_i\tfrac{\partial \epsilon_i}{\partial y_1}\right]\right)\tfrac{\partial y_1}{\partial u_2}\Big]\tfrac{\partial d}{\partial t} \end{aligned} \tag{3.38}$$

$$\begin{aligned} B_{13} \;=\; & \Big[g_{2r}(x_{2r},x_{2i})\tfrac{\partial \epsilon_r}{\partial y_2}\tfrac{\partial y_2}{\partial u_3} - g_{2i}(x_{2r},x_{2i})\tfrac{\partial \epsilon_i}{\partial y_2}\tfrac{\partial y_2}{\partial u_3} + \\ & h_{2r}(x_{2r},x_{2i})\left(\tfrac{1}{\epsilon_r^2+\epsilon_i^2}\tfrac{\partial \epsilon_r}{\partial y_2} - \tfrac{2\epsilon_r}{\left(\epsilon_r^2+\epsilon_i^2\right)^2}\left[\epsilon_r\tfrac{\partial \epsilon_r}{\partial y_2} + \epsilon_i\tfrac{\partial \epsilon_i}{\partial y_2}\right]\right)\tfrac{\partial y_2}{\partial u_3} + \\ & h_{2i}(x_{2r},x_{2i})\left(\tfrac{1}{\epsilon_r^2+\epsilon_i^2}\tfrac{\partial \epsilon_i}{\partial y_2} - \tfrac{2\epsilon_i}{\left(\epsilon_r^2+\epsilon_i^2\right)^2}\left[\epsilon_r\tfrac{\partial \epsilon_r}{\partial y_2} + \epsilon_i\tfrac{\partial \epsilon_i}{\partial y_2}\right]\right)\tfrac{\partial y_2}{\partial u_3}\Big]\tfrac{\partial d}{\partial t} \end{aligned} \tag{3.39}$$

$$\begin{aligned} B_{14} \;=\; & \Big[g_{2r}(x_{2r},x_{2i})\tfrac{\partial \epsilon_r}{\partial y_2}\tfrac{\partial y_2}{\partial u_4} - g_{2i}(x_{2r},x_{2i})\tfrac{\partial \epsilon_i}{\partial y_1}\tfrac{\partial y_2}{\partial u_4} + \\ & h_{2r}(x_{2r},x_{2i})\left(\tfrac{1}{\epsilon_r^2+\epsilon_i^2}\tfrac{\partial \epsilon_r}{\partial y_2} - \tfrac{2\epsilon_r}{\left(\epsilon_r^2+\epsilon_i^2\right)^2}\left[\epsilon_r\tfrac{\partial \epsilon_r}{\partial y_2} + \epsilon_i\tfrac{\partial \epsilon_i}{\partial y_2}\right]\right)\tfrac{\partial y_2}{\partial u_4} + \\ & h_{2i}(x_{2r},x_{2i})\left(\tfrac{1}{\epsilon_r^2+\epsilon_i^2}\tfrac{\partial \epsilon_i}{\partial y_2} - \tfrac{2\epsilon_i}{\left(\epsilon_r^2+\epsilon_i^2\right)^2}\left[\epsilon_r\tfrac{\partial \epsilon_r}{\partial y_2} + \epsilon_i\tfrac{\partial \epsilon_i}{\partial y_2}\right]\right)\tfrac{\partial y_2}{\partial u_4}\Big]\tfrac{\partial d}{\partial t} \end{aligned} \tag{3.40}$$

$$\begin{aligned} B_{21} \;=\; & \Big[g_{2i}(x_{2r},x_{2i})\tfrac{\partial \epsilon_r}{\partial y_1}\tfrac{\partial y_1}{\partial u_1} + g_{2r}(x_{2r},x_{2i})\tfrac{\partial \epsilon_i}{\partial y_1}\tfrac{\partial y_1}{\partial u_1} + \\ & h_{2i}(x_{2r},x_{2i})\left(\tfrac{1}{\epsilon_r^2+\epsilon_i^2}\tfrac{\partial \epsilon_r}{\partial y_1} - \tfrac{2\epsilon_r}{\left(\epsilon_r^2+\epsilon_i^2\right)^2}\left[\epsilon_r\tfrac{\partial \epsilon_r}{\partial y_1} + \epsilon_i\tfrac{\partial \epsilon_i}{\partial y_1}\right]\right)\tfrac{\partial y_1}{\partial u_1} - \\ & h_{2r}(x_{2r},x_{2i})\left(\tfrac{1}{\epsilon_r^2+\epsilon_i^2}\tfrac{\partial \epsilon_i}{\partial y_1} - \tfrac{2\epsilon_i}{\left(\epsilon_r^2+\epsilon_i^2\right)^2}\left[\epsilon_r\tfrac{\partial \epsilon_r}{\partial y_1} + \epsilon_i\tfrac{\partial \epsilon_i}{\partial y_1}\right]\right)\tfrac{\partial y_1}{\partial u_1}\Big]\tfrac{\partial d}{\partial t} \end{aligned} \tag{3.41}$$

$$\begin{aligned} B_{22} \;=\; & \Big[g_{2i}(x_{2r},x_{2i})\tfrac{\partial \epsilon_r}{\partial y_1}\tfrac{\partial y_1}{\partial u_2} + g_{2r}(x_{2r},x_{2i})\tfrac{\partial \epsilon_i}{\partial y_1}\tfrac{\partial y_1}{\partial u_2} + \\ & h_{2i}(x_{2r},x_{2i})\left(\tfrac{1}{\epsilon_r^2+\epsilon_i^2}\tfrac{\partial \epsilon_r}{\partial y_1} - \tfrac{2\epsilon_r}{\left(\epsilon_r^2+\epsilon_i^2\right)^2}\left[\epsilon_r\tfrac{\partial \epsilon_r}{\partial y_1} + \epsilon_i\tfrac{\partial \epsilon_i}{\partial y_1}\right]\right)\tfrac{\partial y_1}{\partial u_2} - \\ & h_{2r}(x_{2r},x_{2i})\left(\tfrac{1}{\epsilon_r^2+\epsilon_i^2}\tfrac{\partial \epsilon_i}{\partial y_1} - \tfrac{2\epsilon_i}{\left(\epsilon_r^2+\epsilon_i^2\right)^2}\left[\epsilon_r\tfrac{\partial \epsilon_r}{\partial y_1} + \epsilon_i\tfrac{\partial \epsilon_i}{\partial y_1}\right]\right)\tfrac{\partial y_1}{\partial u_2}\Big]\tfrac{\partial d}{\partial t} \end{aligned} \tag{3.42}$$

$$
\begin{aligned}
B_{23} \;=\; & \left[g_{2i}(x_{2r}, x_{2i}) \frac{\partial \epsilon_r}{\partial y_2} \frac{\partial y_2}{\partial u_3} + g_{2r}(x_{2r}, x_{2i}) \frac{\partial \epsilon_i}{\partial y_2} \frac{\partial y_2}{\partial u_3} + \right. \\
& h_{2i}(x_{2r}, x_{2i}) \left(\frac{1}{\epsilon_r^2+\epsilon_i^2} \frac{\partial \epsilon_r}{\partial y_2} - \frac{2\epsilon_r}{\left(\epsilon_r^2+\epsilon_i^2\right)^2} \left[\epsilon_r \frac{\partial \epsilon_r}{\partial y_2} + \epsilon_i \frac{\partial \epsilon_i}{\partial y_2} \right] \right) \frac{\partial y_2}{\partial u_3} - \\
& \left. h_{2r}(x_{2r}, x_{2i}) \left(\frac{1}{\epsilon_r^2+\epsilon_i^2} \frac{\partial \epsilon_i}{\partial y_2} - \frac{2\epsilon_i}{\left(\epsilon_r^2+\epsilon_i^2\right)^2} \left[\epsilon_r \frac{\partial \epsilon_r}{\partial y_2} + \epsilon_i \frac{\partial \epsilon_i}{\partial y_2} \right] \right) \frac{\partial y_2}{\partial u_3} \right] \frac{\partial d}{\partial t}
\end{aligned} \tag{3.43}
$$

$$
\begin{aligned}
B_{24} \;=\; & \left[g_{2i}(x_{2r}, x_{2i}) \frac{\partial \epsilon_r}{\partial y_2} \frac{\partial y_2}{\partial u_4} + g_{2r}(x_{2r}, x_{2i}) \frac{\partial \epsilon_i}{\partial y_2} \frac{\partial y_2}{\partial u_4} + \right. \\
& h_{2i}(x_{2r}, x_{2i}) \left(\frac{1}{\epsilon_r^2+\epsilon_i^2} \frac{\partial \epsilon_r}{\partial y_2} - \frac{2\epsilon_r}{\left(\epsilon_r^2+\epsilon_i^2\right)^2} \left[\epsilon_r \frac{\partial \epsilon_r}{\partial y_2} + \epsilon_i \frac{\partial \epsilon_i}{\partial y_2} \right] \right) \frac{\partial y_2}{\partial u_4} - \\
& \left. h_{2r}(x_{2r}, x_{2i}) \left(\frac{1}{\epsilon_r^2+\epsilon_i^2} \frac{\partial \epsilon_i}{\partial y_2} - \frac{2\epsilon_i}{\left(\epsilon_r^2+\epsilon_i^2\right)^2} \left[\epsilon_r \frac{\partial \epsilon_r}{\partial y_2} + \epsilon_i \frac{\partial \epsilon_i}{\partial y_2} \right] \right) \frac{\partial y_2}{\partial u_4} \right] \frac{\partial d}{\partial t}
\end{aligned} \tag{3.44}
$$

$$B_{31} = \left[g_{1r}(x_{1r}, x_{1i}) \frac{\partial \epsilon_r}{\partial y_1} \frac{\partial y_1}{\partial u_1} - g_{1i}(x_{1r}, x_{1i}) \frac{\partial \epsilon_i}{\partial y_1} \frac{\partial y_1}{\partial u_1} \right] \frac{\partial d}{\partial t} \tag{3.45}$$

$$B_{32} = \left[g_{1r}(x_{1r}, x_{1i}) \frac{\partial \epsilon_r}{\partial y_1} \frac{\partial y_1}{\partial u_2} - g_{1i}(x_{1r}, x_{1i}) \frac{\partial \epsilon_i}{\partial y_1} \frac{\partial y_1}{\partial u_2} \right] \frac{\partial d}{\partial t} \tag{3.46}$$

$$B_{33} = \left[g_{1r}(x_{1r}, x_{1i}) \frac{\partial \epsilon_r}{\partial y_2} \frac{\partial y_2}{\partial u_3} - g_{1i}(x_{1r}, x_{1i}) \frac{\partial \epsilon_i}{\partial y_2} \frac{\partial y_2}{\partial u_3} \right] \frac{\partial d}{\partial t} \tag{3.47}$$

$$B_{34} = \left[g_{1r}(x_{1r}, x_{1i}) \frac{\partial \epsilon_r}{\partial y_2} \frac{\partial y_2}{\partial u_4} - g_{1i}(x_{1r}, x_{1i}) \frac{\partial \epsilon_i}{\partial y_2} \frac{\partial y_2}{\partial u_4} \right] \frac{\partial d}{\partial t} \tag{3.48}$$

$$B_{41} = \left[g_{1i}(x_{1r}, x_{1i}) \frac{\partial \epsilon_r}{\partial y_1} \frac{\partial y_1}{\partial u_1} + g_{1r}(x_{1r}, x_{1i}) \frac{\partial \epsilon_i}{\partial y_1} \frac{\partial y_1}{\partial u_1} \right] \frac{\partial d}{\partial t} \tag{3.49}$$

$$B_{42} = \left[g_{1i}(x_{1r}, x_{1i}) \frac{\partial \epsilon_r}{\partial y_1} \frac{\partial y_1}{\partial u_2} + g_{1r}(x_{1r}, x_{1i}) \frac{\partial \epsilon_i}{\partial y_1} \frac{\partial y_1}{\partial u_2} \right] \frac{\partial d}{\partial t} \tag{3.50}$$

$$B_{43} = \left[g_{1i}(x_{1r}, x_{1i}) \frac{\partial \epsilon_r}{\partial y_2} \frac{\partial y_2}{\partial u_3} + g_{1r}(x_{1r}, x_{1i}) \frac{\partial \epsilon_i}{\partial y_2} \frac{\partial y_2}{\partial u_3} \right] \frac{\partial d}{\partial t} \tag{3.51}$$

$$B_{44} = \left[g_{1i}(x_{1r}, x_{1i}) \frac{\partial \epsilon_r}{\partial y_2} \frac{\partial y_2}{\partial u_4} + g_{1r}(x_{1r}, x_{1i}) \frac{\partial \epsilon_i}{\partial y_2} \frac{\partial y_2}{\partial u_4} \right] \frac{\partial d}{\partial t} \tag{3.52}$$

$$C_{11} = \left[\frac{x_{2r}}{x_{2r}^2 + x_{2i}^2} \right] \tag{3.53}$$

$$C_{12} = \left[\frac{x_{2i}}{x_{2r}^2 + x_{2i}^2} \right] \tag{3.54}$$

$$C_{13} = \left[\frac{x_{1r}}{x_{2r}^2 + x_{2i}^2} - \frac{2x_{2r}\left(x_{1r}x_{2r} + x_{1i}x_{2i}\right)}{\left(x_{2r}^2 + x_{2i}^2\right)^2} \right] \tag{3.55}$$

$$C_{14} = \left[\frac{x_{1i}}{x_{2r}^2 + x_{2i}^2} - \frac{2x_{2i}\left(x_{1r}x_{2r} + x_{1i}x_{2i}\right)}{\left(x_{2r}^2 + x_{2i}^2\right)^2} \right] \tag{3.56}$$

$$C_{21} = \left[-\frac{x_{2i}}{x_{2r}^2 + x_{2i}^2} \right] \tag{3.57}$$

$$C_{22} = \left[\frac{x_{2r}}{x_{2r}^2 + x_{2i}^2} \right] \tag{3.58}$$

$$C_{23} = \left[\frac{x_{1i}}{x_{2r}^2 + x_{2i}^2} - \frac{2x_{2r}\left(x_{1i}x_{2r} - x_{1r}x_{2i}\right)}{\left(x_{2r}^2 + x_{2i}^2\right)^2} \right] \tag{3.59}$$

$$C_{24} = \left[-\frac{x_{1r}}{x_{2r}^2 + x_{2i}^2} - \frac{2x_{2i}\left(x_{1i}x_{2r} - x_{1r}x_{2i}\right)}{\left(x_{2r}^2 + x_{2i}^2\right)^2} \right] \tag{3.60}$$

Note that all these entries are evaluated at the nominal value of the input, u_o, and the associated equilibrium x_o. It is apparent, then, that B and C are full matrices, while A is block diagonal, as we would expect.

3.3.3 Simulation Results

Given the linearized system, we achieve steady-state tracking by augmenting the plant with integrators at the plant output and then designing a stabilizing controller for this augmented system. The actual controller includes these integrators at the control input, and with the integral action achieves the desired tracking performance.

For the control design we used H_2 design tools that computed an optimal H_2 controller with respect to a quadratic cost function. The cost function itself was manipulated as a design parameter until a satisfactory response was achieved. The objective here is not to present the absolute best controller performance, however, but to demonstrate that standard linear design techniques can be employed to deliver unprecedented first-pass success for difficult device designs.

Figures ?? and 3-12 show how a particular controller design [70] performed tracking different graded quaternary structures. On these plots the composition, (y_1, y_2), is plotted on the same figure. For these designs, y_1 ranges from 0.6 to 1.0 and y_2 ranges from 0.0 to 0.4. The dashed line indicates the reference trajectory, while the solid line indicates the actual film composition. Nominal control inputs were chosen so the growth rate would be about 3 angstroms per second. The simulation param-

eters k_i and w_i were perturbed 5-10% in these simulations to capture the affects of uncertainty in the system, and white Gaussian measurement noise with variance on the order of 15% of the signal variation over the entire structure was added to the $\boldsymbol{\rho}$ measurements.

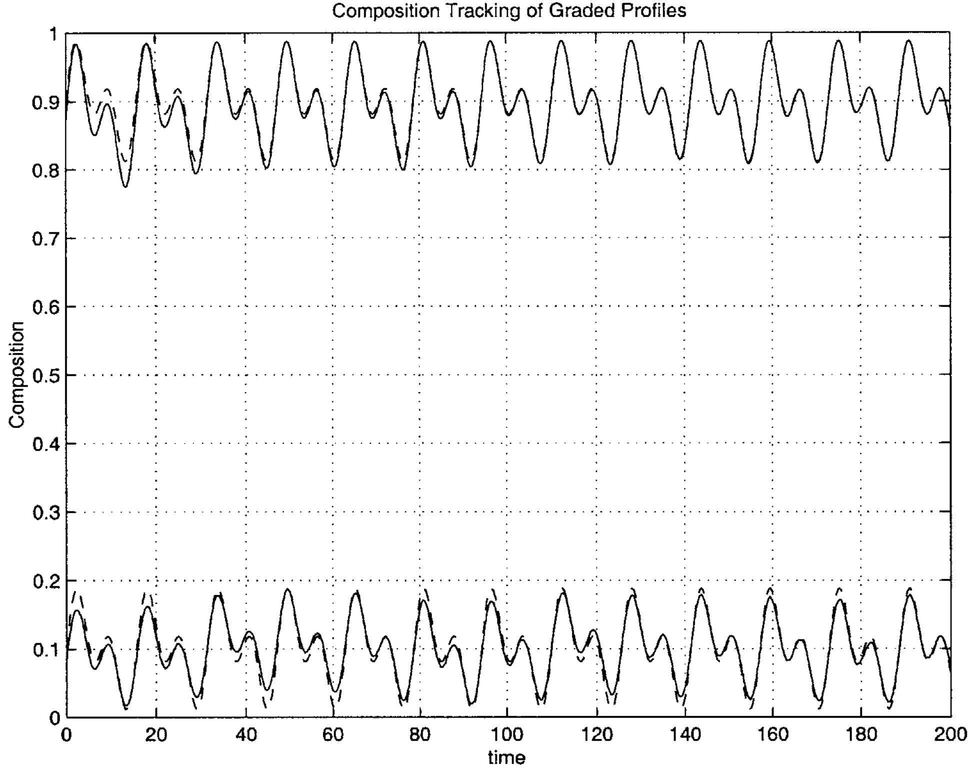


Figure 3-11: Simulated closed-loop HSMBE growth of a graded InGaAsP structure.

Figure 3-11 is a highly unusual periodic structure with a period of about 50 Angstroms. The design is not motivated from a particular device, but rather as an example of the utility of feedback to reproduce currently ungrowable structures with first-pass success.

Figure 3-12 is a design reminiscent of a vertical cavity surface emitting laser (VCSEL) with bi-parabolic mirrors and multiple quantum wells. Such a VCSEL design has recently been shown to have enhanced performance over other VCSEL designs that avoid the graded profiles, yet it is difficult to manufacture since the graded mirrors need strong reproducibility over multiple periods to yield the desired effect. These results indicate, however, that such reproducibility is possible even on the first

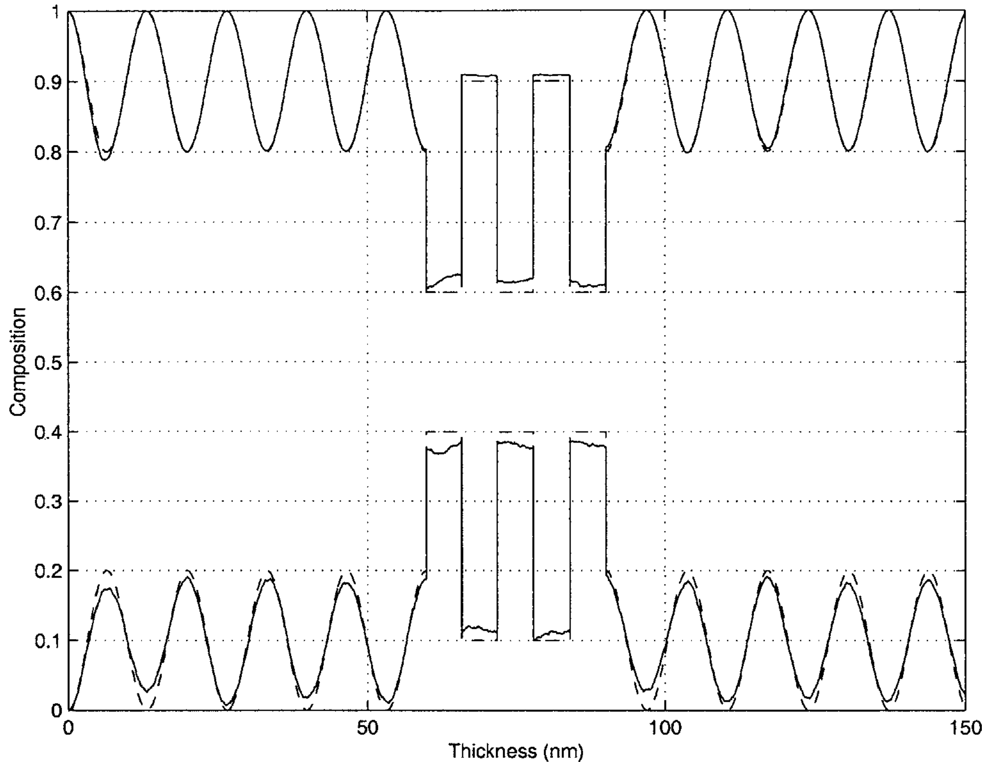


Figure 3-12: Simulated closed-loop HSMBE growth of a graded structure with multiple quantum wells.

growth, before system parameters have been calibrated and tuned. Even the quantum wells were reproduced with some level of accuracy without making special efforts to improve their tracking performance.

Note that the performance level in the quantum wells may not be acceptable, and we might with to do better. Part of what makes them so hard to track is the fact that those compositions are very far away from the linearization point. To do better, one may need to linearize around multiple points and schedule the controllers together to switch smoothly between them. Even without scheduling, however, the controller maintains stability in this case.

Overall, this level of performance for the first-pass growth of such difficult device designs is an encouraging demonstration of the value of ellipsometric feedback for epitaxy. Using calibration runs to estimate the parameters of the system to better than 5-10% will only improve performance, realizing reproducible growth for even the most difficult of device structures.

Chapter 4

Piloting Epitaxy

Although simulations provide encouraging motivation to the impact feedback may have on epitaxy, the real test will be reliable implementation. In an industry where so much down-time is caused by software failure, it will be the experimental proof that a desired profile can be grown accurately every time that will prove the concept.

This chapter concludes our study by discussing some of the hurdles yet to be crossed. First, we report our current state of implementation [56], then we highlight issues for future research.

4.1 Implementation

Evolving the laboratory in preparation for closed-loop operation is a tedious and often expensive operation. Equipment that was never designed for a particular application is used by necessity, often introducing constraints that were not anticipated. Sometimes reformulation is necessary.

Our focus on a simple feedback design objective and standard design tools has allowed us to develop a virtual epitaxy laboratory in a simulink environment that expedites control design and analysis. This streamlining allows great flexibility as implementation issues force unanticipated constraints without having to start from scratch.

As we began this project, our hope was to tap into the ellipsometer operating

software and measure $\rho(t)$, or at least the photodetector voltage $V(t)$, continuously. We also were unsure how to address substrate wobble during rotation, which can cause the reflected beam from the ellipsometer to move off and on the photodetector. Moreover, we anticipated operating the reactor and ellipsometer from the same control computer.

It turns out, however, that we could not bypass the ellipsometer control code and access the direct measurement continuously. Moreover, wobble was too severe on the existing machine to fix, so a new MBE system was ordered that anticipated the need for in-situ measurements, eliminating wobble. Finally, the particular ellipsometer we purchased only operates on a system using Windows 95, forcing operation on a machine distinct from the reactor control software.

These evolving needs motivated us to develop a pre-experiment testing platform where we could convince ourselves that the closed loop interconnection was functioning as anticipated. As this closed-loop infrastructure is fixed, we will test the robustness of our existing designs against any unmodeled dynamics the changes introduce. We can then model these infrastructural dynamics and incorporate them into the virtual laboratory so the feedback design will account for them. This section provides an overview of the existing experimental set-up and reports our current state of pre-experiment tests.

4.1.1 Experimental Set-up

The HSMBE lab at MIT revolves around a dual-chamber Riber 32p reactor equipped with ellipsometry ports on the III-V chamber. The machine is controlled from a Dell Precision 530 workstation using an Intel XEON dual processor, running Windows 2000, and equipped with a RS422 serial communications card and a National Instruments data acquisition card. A dedicated RS232 port is used to control source shutters, while the RS422 connection delivers setpoint commands to Eurotherm temperature controllers driving the effusion cells.

The National Instruments data acquisition card facilitates three communication channels. The first is an analog to digital connection that reads measurements from

pressure gauges, mass flow controllers, the RHEED sensor, etc. The next channel is a digital to analog connection that delivers the mass flow controller set points to the hydride sources. Finally, a digital port monitors the valve states, whether gases are routed to the vent pump or through the reaction chamber.

The control computer uses version 6 of National Instruments' Labview laboratory operation software as the primary dashboard. This dashboard is driven by a homemade reactor-control system, written in Labview, and tailored to the specific laboratory set-up.

The ellipsometer is a Wollam M-88 spectroscopic ellipsometer. Its control computer is a Zeos Pentium 1 running Windows 95 and specially equipped with Wollam's A/D and system communications cards. These cards run the ellipsometer hardware. The control software is WVASE 32, which drives the hardware but also provides characterization tools and an operating dashboard. The ellipsometer talks to the reactor control computer through a TCP/IP network connection.

4.1.2 Pre-Experiment Testing

To verify the feasibility of the closed-loop connection, a Dell X200 notebook and Sony Viao 505 notebook, both running Windows 2000 and Labview, were connected through their TCP/IP network connections. The Sony was set-up to emulate both the Riber machine and the ellipsometer control computer running WVASE using a simulation designed in Labview. The Dell was set-up running the actual reactor control software with the output drives turned off.

The reactor control software is designed to read in a Matlab file containing a linear feedback controller characterized by four matrices (A, B, C, D):

$$\dot{x} = Ax + Be$$

$$du = Cx + De$$

where du are the control updates and e is the rho-error given by $e = \rho^{ref} - \rho$. The

software initializes by sending an initialization key to the Emulator, which responds by giving an initial measurement, ρ_o, of the pre-growth substrate. The reactor control software then changes the setting in its status file indicating that shutters should open and growth should begin. At this time the reactor control software also begins running an ellipsometer simulation, using ρ_o as an initial condition. The Emulator continuously checks the reactor control's status file, and when it sees the change in status to begin growth, it simulates the opening of shutters and the corresponding growth/ellipsometry measurements. The reactor control software then requests an updated ellipsometer measurement every .5 seconds, and the emulator is designed to send a measurement every time it receives a request. Once the control software receives a measurement update, a digital implementation of the feedback controller is made by computing

$$e = \rho^{ref} - \rho$$

$$Xnew = (I + \Delta TA)Xold + \Delta TBe$$

$$du = CXnew$$

$$Xold = Xnew$$

where ΔT is the time step in the forward Euler approximation to the derivative (time between samples). The resulting change in the status file is made to reflect the new control inputs.

The experiment we are conducting linearizes a controller around a nominal composition, and then commands a composition near this nominal point. Initially when the controller was implemented, coding bugs caused the resulting experiment to be obviously unstable. We then found these bugs, corrected them, and although the resulting test was clearly not unstable as before, the composition still did not stabilize after 100nm.

Various conjectures were made as to the discrepancy between the simulated and tested responses. One thought was that a time delay was introduced in the measurement that was unanticipated in the controller design. We thus simulated the designed

controller against the plant with a time delay of 5 seconds introduced on the measurement signal. Figure 4-1 shows that the system stabilizes even with this unmodeled effect. The next idea was that in addition to this delay, the WVASE software doesn't

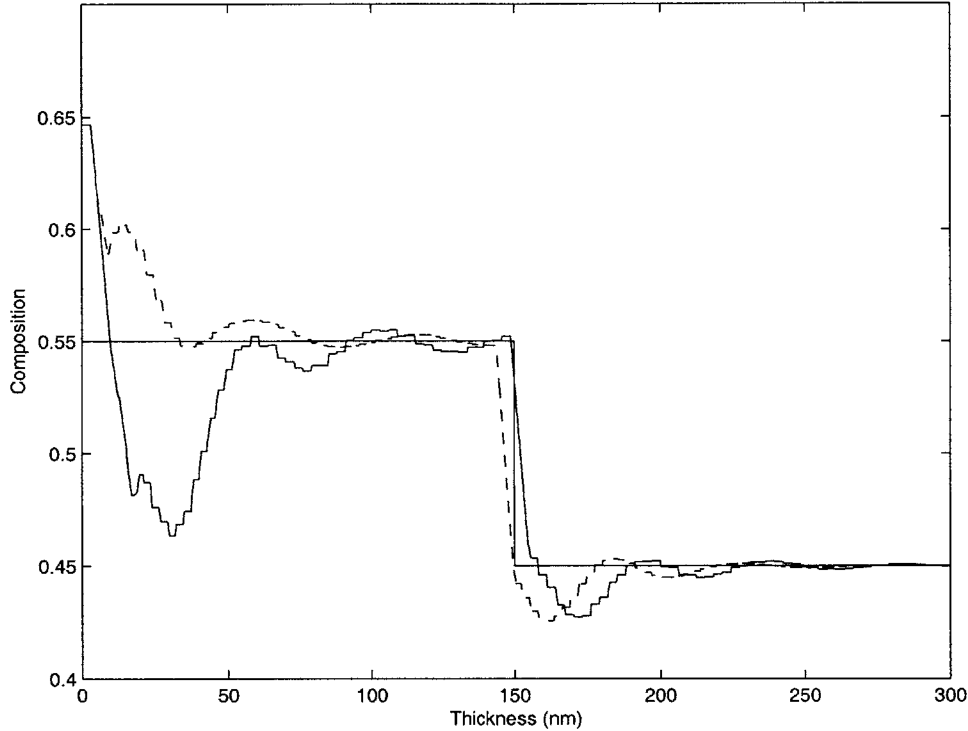


Figure 4-1: Simulated response with delay (dashed) and with delay and hold (solid).

update the measurement values as frequently as we expected. A 20-second zero-order hold, in addition to the delay, was then introduced. Note that both of these simulations were conducted with the same degree of parametric uncertainty as the earlier ones (2-5%).

Since neither of these effects seem to be causing the observed behavior in the tests, more work is needed to identify the unmodeled component of the pre-experiment setup. Some ideas are the lack of synchronization between the ellipsometer simulation in the control computer and that of the emulator, and the resulting skew in constructing the measurement error. Additionally, the digital implementation of the continuous controller should be modeled to capture the discretization effects, or the controller design can be conducted in discrete time from the start. Also, the effect of network delay could be modeled on both the control input and measurement output to note

the combined impact on the system. Whatever the reason, what is clear to date is that the something about the current implementation remains unmodeled in the virtual laboratory.

4.2 Future Work

This project laid the foundation for the design of closed-loop deposition processes by focusing on the generic aspects of control design for these types of problems. The ellipsometer model is easily extended to other optical sensors, which are preferred for these applications due to their non-invasive nature. The process model is coarse and covers a range of specific processes.

The nature of synchronization as an objective function applies broadly to problems where the objective is to produce a specific product. The resulting separation between product quality and process timing is an important feature of this problem that has not been fully explored. In particular, the idea of using feedback to choose a better d^{ref} function in real time, as opposed to choosing the function a priori, introduces the idea of feedback synchronization as a way of trading off desired quality with total throughput of the process. This trade off yields a measure for the manufactuability of a given device design. Some work was conducted in this direction that relied on the existence of an additional sensor to independently measure film thickness [71].

Moreover, leveraging feedback to deliver the desired compositional profiles introduces a level of process automation that substantially changes the role of the machine operator. Currently, the operator of an epitaxy machine fills a variety of roles that have evolved from necessity more than design. For example, it is not uncommon for an operator to not only conduct calibration experiments, decide process settings, monitor growth, and maintain the data record, but also maintain the machine, replace parts, order source gases, keep track of hazardous waste, ensure appropriate disposal of toxic materials, assess production capabilities, and schedule jobs. Sometimes, since different composition profiles may yield the same device performance, the machine operator is even consulted by device physicists about the proposed design to help

choose a structure that would be easiest to manufacture.

With ellipsometric feedback, the operator may be relieved of most calibration, process setting, monitoring, and recording responsibilities. Nevertheless, these tasks may be replaced by sensor hardware integration and maintenance, process linearization with each new device profile, and subsequent controller design and testing. Effectively, expensive calibration experiments are replaced by less expensive controller design iterations, where each new design can be simulated and tested in a virtual epitaxy laboratory before being deployed and incurring processing costs.

With a globally stabilizing controller, however, the demands of redesign are relieved. Then one might ask what should the role of the operator be? While the operator currently focuses on the growing machine, automation may lift attention from the machine onto the desired product, or even to production with the scheduling of jobs and coordination of multiple processes. Should automation replace operators altogether, or does it simply change the nature of their jobs? If operators are to remain, what will they do and how will their efforts be integrated with process automation? These questions will always be important to engineering problems where feedback is concerned: just how should epitaxy be piloted?

Bibliography

[1] TRW's new integrated HEMT/HBT circuits. *Compound Semiconductor*, 1(2):6, September/October 1995.

[2] Discrete inquiry into the transistor market. *Compound Semiconductor*, 2(4):24, July/August 1996.

[3] First release of the MAFET design environment. *Compound Semiconductor*, 3(5):34, September/October 1997.

[4] Hall elements from Asahi Chemical. *Compound Semiconductor*, page 38, Special Issue 1997.

[5] HP introduces highest-intensity LEDs. *Compound Semiconductor*, 3(5):4, September/October 1997.

[6] Integrated laser/HBT structure. *Compound Semiconductor*, 3(3):41, May/June 1997.

[7] LEDs on tour. *Compound Semiconductor*, 3(4):23, July/August 1997.

[8] Planar processing of HBTs. *Compound Semiconductor*, 3(3):41, May/June 1997.

[9] White light burning bright at familiar site. *Compound Semiconductor*, page 28, Special Issue 1997.

[10] Ibm demonstrates world's first 1GHz microprocessor. *Compound Semiconductor*, 4(2):12, March 1998.

[11] Optically pumped red and blue organic lasers. *Compound Semiconductor*, 4(2):36, March 1998.

[12] Planar self-aligned RTD fabrication. *Compound Semiconductor*, 4(3):40, April 1998.

[13] D.E. Aspnes. Minimal-data approaches for determining outer-layer dielectric responses of films from kinetic reflectometric and ellipsometric measurements. *Journal of the Optical Society of America*, 10(5):974–983, May 1993.

[14] D.E. Aspnes. Optical approaches to the determination of composition of semiconductor alloys during epitaxy. *IEEE Journal of Selected Topics in Quantum Electronics*, 1(4):1054–1063, December 1995.

[15] D.E. Aspnes, W.E. Quinn, and S. Gregory. Application of ellipsometry to crystal growth by organometallic molecular beam epitaxy. *Applied Physics Letters*, 56(25):2569–2571, June 1990.

[16] D.E. Aspnes, W.E. Quinn, M.C. Tamargo, S. Gregory, S.A. Schwarz, M.A.A. Pudensi, M.J.S.P. Brasil, and R.E. Nahory. Growth of $Al_xGa_{1-x}As$ parabolic quantum wells by real-time feedback control of composition. *Applied Physics Letters*, 60(10):1244–1246, March 1992.

[17] R.M.A. Azzam and N.M. Bashara. *Ellipsometry and Polarized Light*. North-Holland, NY, 1977.

[18] D. Barton, J.C. Comfort, and F.K. Urban. Numerical ellipsometry: Real-time solutions using mapping onto the complex index plane. *Journal of Vacuum Science Technology, A*, pages 167–17, May/June 1996.

[19] M. Born and E. Wolf. *Principles of Optics: Electromagnetic Theory of Propagation, Interference, and Defraction of Light*. Pergamon Press, 1959.

[20] F.G. Celii, Y.C. Kao, T.S. Moise, M. Woosley, T.B. Harton, and K. Haberman. Closed-loop thickness control of resonant-tunneling diode MBE growth using spectroscopic ellipsometry. *MRS Symposium Proceedings*, 1996.

[21] M.E. Coltrin, R.J. Kee, and J.A. Miller. A mathematical model of the coupled fluid mechanics and chemical kinetics in a chemical vapor deposition reactor. *Journal of the Electrochemical Society*, 131(2), February 1984.

[22] M.E. Coltrin, R.J. Kee, and J.A. Miller. A mathematical model of silicon chemical vapor deposition: Further refinements and the effects of thermal diffusion. *Journal of the Electrochemical Society*, 133(6), June 1986.

[23] Spire Corporation. Annual report, 1993.

[24] M. George Craford. Wide bandgap LED overview. *Compound Semiconductor*, 1(1):48, July/August 1995.

[25] J.N. Damask. *Integrated-Optic Grating-Based Filters for Optical Communication Systems*. PhD thesis, Massachusetts Institute of Technology, 1997.

[26] Sh.A. Furman and A.V. Tikhonravov. *Basics of Optics of Multilayer Systems*. Editions Frontieres, Gif-sur-Yvette, 1992.

[27] M.S. Gaffney, C.M. Reaves, R.S. Smith, A.L. Holmes, and S.P. DenBaars. Control of III-V epitaxy in a metalorganic chemical vapor deposition process: Impact of source flow control on composition and thickness. *Journal of Crystal Growth*, September 1996.

[28] Cecilia Gabriela Galarza. Learning and estimation theory for manufacturing systems applied to microelectronics manufacturing. *Ph.D. Dissertation, The University of Michigan*, 1999.

[29] M.A. Gevelber, M. Bufano, M. Toledo-Quinones, D. Brown, and R. Pasaro. Dynamic modelling analysis for control of chemical vapor deposition. *Proceedings of the American Control Conference*, 2, 1994.

[30] S. He, S. Strom, and V.H. Weston. *Time Domain Wave-Splittings and Inverse Problems*. Oxford University Press, 1998.

[31] E Hecht and Zajac. A. *Optics*. Addison-Wesley Publishing Company, 1979.

[32] B. Johs, D. Doerr, S. Pittal, I.B. Bhat, and S. Dakshinamurthy. Real-time monitoring and control during MOVPE growth of CdTe using multiwavelength ellipsometry. *Thin Solid Films*, 233:293–296, 1993.

[33] B. Johs, C. Herzinger, J.H. Dinan, A. Cornfeld, J.D. Benson, D. Doctor, G. Olson, I. Ferguson, M. Pelczynski, P. Chow, C.H. Kuo, and S. Johnson. Real-time monitoring and control of epitaxial semiconductor growth in a production environment by in situ spectroscopic ellipsometry. *Thin solid films*, 313-314:490–495, 1998.

[34] Michael Kenward. Fantastic plastic, part one: Polymer-based displays. *Compound Semiconductor*, 2(6):30, November/December 1996.

[35] Jin Au Kong. *Electromagnetic Wave Theory.* John Wiley and Sons, Inc., 1990.

[36] C.H. Kuo, M. Boonzaayer, M. DeHerrera, T. Kyong, Y.H. Zhang, B. Johs, and J.S. Hale. Real time in-situ composition control of InGaAs lattice matched to InP by 88 wavelength ellipsometer. *16th North American Conference on Molecular Beam Epitaxy*, 1997.

[37] John Lekner. *Theory of Reflection of Electromagnetic and Particle Waves.* Kluwer Academic Publishers, 1987.

[38] H.A. Macleod. *Thin-Film Optical Coating Design*, pages 1–39. Optical Engineering. Marcel Dekker, Inc., 1995.

[39] G.N. Maracas, J.L. Edwards, D.S. Gerber, and R. Droopad. In situ spectroscopic ellipsometry in MBE for photonic devices. *Applied Surface Science*, 1993.

[40] E.A. Marley. The development of inp-based optoelectronic devices. *Masters Thesis, Massachusetts Institute of Technology*, 1996.

[41] Elisabeth A. Marley. The development of InP-based optoelectronic devices. Master's thesis, Massachusetts Institute of Technology, 1996.

[42] Robert A. Metzger. Cutting the cord. *Compound Semiconductor*, 1(2):33, September/October 1995.

[43] Robert A. Metzger. Turning blue to green. *Compound Semiconductor*, 1(1):26, July/August 1995.

[44] Robert A. Metzger. Manufacturing III-V solar cells for space applications. *Compound Semiconductor*, 2(6):25, November/December 1996.

[45] Robert A. Metzger. Progress and challanges in the diamond semiconductor field. *Compound Semiconductor*, 2(4):30, July/August 1996.

[46] Robert A. Metzger. Even in microelectronics GaN was big news in 1997. *Compound Semiconductor*, 3(6):21, November/December 1997.

[47] Marie Meyer. Infared LEDs. *Compound Semiconductor*, 2(3):39, May/June 1996.

[48] Marie Meyer. 9 things to watch in 1997. *Compound Semiconductor*, 3(1):21, January/February 1997.

[49] Marie Meyer. Full color LED displays. *Compound Semiconductor*, page 49, Special Issue 1997.

[50] Marie Meyer. Gallium nitride device forcast: $3 billion by 2006. *Compound Semiconductor*, 3(6):8, November/December 1997.

[51] Marie Meyer. High temperature semiconductors: Evolution, not revolution. *Compound Semiconductor*, page 42, Special Issue 1997.

[52] Marie Meyer. InP for the information society. *Compound Semiconductor*, 3(4):8, July/August 1997.

[53] Marie Meyer. Will HBTs be the next compound semiconductor success story? *Compound Semiconductor*, 3(3):30, May/June 1997.

[54] J.M. Olsen and D.J. Friedman. High efficiency III-V solar cells. *Compound Semiconductor*, 2(6):27, November/December 1996.

[55] M.B. Panish and H. Temkin. *Gas Source Molecular Beam Epitaxy: Growth and Properties of Phosphorus Containing III-V Heterostructures*, volume 26 of *Springer Series in Marterial Science*. Springer-Verlag, 1993.

[56] Dr. Gale Petrich. Research Scientist. personal communication with the author, July 2003.

[57] S. Pittal, P. He, J.A. Woolam, S.D. Murthy, and I.B. Bhat. In-situ monitoring and control of MOCVD growth using multiwavelength ellipsometry. *Proceedings on 21st International Symposium on Compound Semiconductors*, February 1994.

[58] J. Ramer, B. Patel, A. Patel, V. Boguslavskiy, and A. Gurary. Real time substrate temperature control by emissivity compensated pyrometry during $In_xGa_{1-x}As_{1-y}P_y/InP$ growth on production scale rotating disc MOVPE reactors. *Proceedings of the MRS Symposium on New-Methods, Mechanisms and Models of Vapor Deposition*, 2000.

[59] Professor Linton Salmon. Director of the Integrated Microelectronics Laboratory at Brigham Young University. personal communication with the author, July 1996.

[60] Peter Singer. Cvd technology trends. *Semiconductor International*, pages 55–62, February 1995.

[61] D. L. Smith. *Thin Film Deposition: Principles and Practice.* McGraw Hill, Inc., 1995.

[62] G.B. Stringfellow. *Organometallic Vapor-Phase Epitaxy: Theory and Practice.* Academic Press, Boston, 1989.

[63] Toby Strite and Walter Riess. Organic LEDs vs III-V's: Basics, status and outlook. *Compound Semiconductor*, 3(6):34, November/December 1997.

[64] M.K. Tucker and D.G. Meyer. Nonlinear modeling, identification, and feedback control design for the modern effusion cell. Technical report, Nonlinear and Real-Time Control Laboratory, University of Colorado, Boulder, 1997.

[65] F.K. Urban and M.F. Tabet. Real time in-situ ellipsometry solutions using artificial neural network preprocessing. *Thin Solid Films*, pages 167–173, 1994.

[66] James R. Wait. *Electromagnetic Waves in Stratified Media.* Pergamon Press, 1970.

[67] S.C. Warnick. Feedback control of organometallic vapor-phase epitaxial growth of aluminum gallium arsenide devices. *Masters Thesis LIDS-TH-2313, Massachusetts Institute of Technology*, 1995.

[68] S.C. Warnick and M.A. Dahleh. Feedback control of MOCVD growth of submicron compound semiconductor films. *IEEE Transactions on Control System Technology*, pages 62–71, January 1998.

[69] S.C. Warnick and M.A. Dahleh. Piloting epitaxy through ellipsometric feedback. *Proceedings of the 37th Conference on Decision and Control*, 1998.

[70] S.C. Warnick and M.A. Dahleh. Ellipsometry as a sensor technology for the control of deposition processes. *IEEE International Conference on Control Applications*, 1999.

[71] S.C. Warnick and M.A. Dahleh. Feedback synchronization: A relaxation of tracking. *Proceedings of the American Controls Conference*, 2000.

[72] Asahi kasei electronics homepage. www.asahi-kasei.co.jp/ake, 1997.

[73] Tim Whitaker. Tunnel diodes break through at last. *Compound Semiconductor*, 4(3):36, April 1998.

[74] J. A. Inc. Woolam. *WVASE User's Manual.* J. A. Woolam, Inc., 1995.