MEMS Reference Shelf

Olav Solgaard

Photonic Microsystems

Micro and Nanotechnology Applied to Optical Devices and Systems

Springer

Photonic Microsystems

Micro and Nanotechnology Applied to Optical Devices and Systems

MEMS Reference Shelf

Series Editors:

Stephen D. Senturia
Professor of Electrical Engineering, Emeritus
Massachusetts Institute of Technology
Cambridge, Massachusetts

Robert T. Howe
Department of Electrical Engineering
Stanford University
Stanford, California

Antonio J. Ricco
Small Satellite Division
NASA Ames Research Center
Moffett Field, California

Photonic Microsystems: Micro and Nanotechnology Applied to Optical Devices and Systems
Olav Solgaard
ISBN: 978-0-387-29022-5

MEMS Vibratory Gyroscopes Structural Approaches to Improve Robustness
Cenk Acar and Andrei Shkel
ISBN: 978-0-387-09535-6

BioNanoFluidic MEMS
Peter Hesketh, ed.
ISBN 978-0-387-46281-3

Microfluidic Technologies for Miniaturized Analysis Systems
Edited by Steffen Hardt and Friedhelm Schöenfeld, eds.
ISBN 978-0-387-28597-9

Forthcoming Titles

Self-assembly from Nano to Milli Scales
Karl F. Böhringer
ISBN 978-0-387-30062-7

Micro Electro Mechanical Systems: A Design Approach
Kanakasabapathi Subramanian
ISBN 978-0-387-32476-0

Experimental Characterization Techniques for Micro-Nanoscale Devices
Kimberly L. Turner and Peter G. Hartwell
ISBN 978-0-387-30862-3

Microelectroacoustics: Sensing and Actuation
Mark Sheplak and Peter V. Loeppert
ISBN 978-0-387-32471-5

Inertial Microsensors
Andrei M. Shkel
ISBN 978-0-387-35540-5

Olav Solgaard

Photonic Microsystems

Micro and Nanotechnology Applied to Optical Devices and Systems

Springer

Olav Solgaard
Stanford University
E.L. Ginzton Laboratory
Stanford, CA 94305

MEMS Reference Shelf ISSN 1936-4407

ISBN 978-0-387-29022-5 e-ISBN 978-0-387-68351-5

Library of Congress Control Number: 2008937947

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

To Terri, Jenni, and Nikolai

Preface

Like many other engineers and inventors, I believe that the boundaries between traditional fields offer unique and exciting opportunities for innovation and new developments. This is almost self evident when one considers complex systems that integrate functions from several domains. It is also natural that the boundaries between fields are less understood, simply because their study requires expertise in two or more fields.

From this last observation, it follows that interdisciplinary research is hard. It requires dedicated individuals who are willing to make the heavy investments necessary to master several fields of inquiry, or, something even more extraordinary, teams that are able to smoothly communicate across disciplinary boundaries. This is the defining problem of the book. It is written to encourage and facilitate interdisciplinary research on optical microsystems, by which we mean optics created using microfabrication technology, i.e. the tools and techniques developed to fabricate Integrated Circuits (ICs) and MicroElectroMechanical (MEMS).

Innovation and design of modern optical systems requires input from many fields, as well as specific application knowledge. Examples include optical interconnects, optical-fiber communication networks, digital projectors, and imagers for photography and microscopy. The design of these systems depends on seamless integration of optics with electronics and mechanics. The best solutions are optimized over all these domains to meet application demands. In the case of micro-optics, the interdisciplinary requirements are even stricter; these systems must be optimized for the Integrated Circuit (IC) and MicroElectroMechanical (MEMS) fabrication environment. A large part of that optimization is to reduce the dimensions of the optical-systems designs so that they can be practically and economically fabricated using IC and MEMS techniques.

This book gives students, researchers, and developers the tools they need to analyze and design micro-optical devices systems. Design is the ultimate “inverse” problem, so the emphasis is on analytical models that can be turned into design equations. The point is to enable interdisciplinary research, so very little background in optics, MEMS, or fabrication is assumed. The first part on optics fundamentals is accessible to readers with an understanding of first-year, university-level physics. The book is self-contained in that the concepts developed in the first part give the necessary background for understanding the detailed descriptions of the second and third parts.

Acknowledgements

This book would not be possible if it were not for my collaborators at Stanford, UC Berkeley, UC Davis, the University of Oslo, and at SINTEF in Oslo. Their brilliant insights and stimulating discussions have been a constant source of inspiration, and I am forever grateful for being able to work in the exciting environment that they create. For as much as I have learned from my colleagues, I believe my students taught me more. Working with such a talented group has been a true privilege and I thank all of them for the time and effort they invested and for their many contributions. A special thanks also goes to the reviewers of this text. They made it better in many ways and therefore more enjoyable for the reader.

Being a teacher, I believe in the power of good mentors, and I have been lucky to learn from some of the best. During my years as a post doc at Berkeley, I worked with Professors Kam Lau and Richard Muller. Between these two leading experts, the fields of semiconductor lasers and MEMS were opened to me. In addition to their technical advice, I owe them both for creating an inspiring and demanding environment and for encouraging me to following my own ideas. But my biggest debt of gratitude goes to my PhD advisor, Professor David Bloom. He, more than anyone else, taught me that it is always possible to improve the status quo, that even crazy ideas can be harnessed, and that the best solutions are often found in unlikely places.

Contents

1: Introduction to Optical Microsystems

1.1 Scaling of Optics

The main theme of this book is miniaturization of optics. We ask the question: "How small can we make an optical system?", and we explore how size affects optical characteristics, and how performance changes as we scale optics to the micrometer and nanometer scales. The goals are to present the fundamental limits and illuminate the advantages and challenges of scaling optics down in size, and ultimately to teach how to design miniaturized optics.

Modern optics is primarily used for information capture, communication, and presentation[a]. The motivation for miniaturization of optics is therefore the same as for electronics; to create cheaper and more functional information-technology (IT) systems, and to gain access to regions where bulk equipment will not fit. Examples of the latter include remote sensing, and, increasingly, *in-vivo* microscopy. Optical microscopy is used extensively for determining health and pathology of biopsy samples. By miniaturizing optics, we will be able to take the microscope to the patient, instead of taking (pieces of) the patient to the microscope.

The second theme of the book, closely related to the first, is integration. Our premise is that highly-functional IT systems require both electronics (for computations) and optics (for communication), so these two should be closely integrated. That leads us to consider Integrated Circuits (IC) and MicroElectroMechancial System (MEMS) as platforms for optical systems. Both ICs and MEMS are largely based on Silicon, so our focus is on "silicon optics", i.e. optical devices and systems that can be realized in silicon fabrication technology, AND that can enhance existing IT systems[b].

It is always challenging to apply a technology outside its intended field of use. Silicon technology is, however, fundamentally very flexible, and the tools that are

---

[a] Much important research in optics is directed at information processing, but in practice this area of IT is still squarely in to domain of electronics.

[b] Many of the techniques and solutions described in this book are also applicable to solar cells, but our focus is on IT applications.

developed for IC and MEMS processing are very powerful, so viable solutions can be found for almost any fabrication challenge presented by optical devices and systems. Often it is necessary, however, to use the tools in unconventional ways. Significant parts of the book tool are therefore dedicated to descriptions of unorthodox silicon processing.

At first sight, it may appear strange that this book, which is mostly about optical device concepts and design, puts such emphasis on a single fabrication methodology. It is the revolutionary capabilities of ICs and MEMS that justifies this approach. Silicon technology allows vast numbers of devices to be integrated and aligned with great precision on a common substrate (chip). This makes practical a large number of optical systems that rely on interaction between different individual devices. For example, Texas Instruments DLP® technology, which integrates more than 1 million moving mirrors, would be a practical impossibility with any fabrication technology that does not leverage the parallel-processing advantage of modern lithography. This will be a reoccurring theme of this book: ICs and MEMS provide a flexible and practical fabrication technology for realizing optical systems that are cumbersome or prohibitively expensive when using traditional fabrication technology.

A direct consequence of integration with electronics is the availability of cheap and abundant signal processing. That undermines one of the traditional tenets of optical design that says that optics should be designed to have the minimum required number of degrees of freedom to improve stability and to simplify control and calibration. Many of the systems presented in this book take the opposite view: If we can reduce overall size, then it is advantageous to use many parallel systems to do the job of a single large device, in spite of the fact that the complexity of control increases dramatically. Examples of systems based on this design philosophy include projectors and imagers, as well as switches and adaptive optics.

An integral part of IT is information capture, so optical sensors, or rather sensors with optical output signals, are treated in detail in this book. The main advantages of optical-output sensors are that they can be designed for many important measurands (e.g. pressure, acceleration, rotation, temperature, and bio-molecular association), that they are thermally and chemically robust, and that their output signals are immune to ElectroMagnetic Interference (EMI). These advantages would be of little consequence, however, if the sensors didn't also have better sensitivity, specificity, and reliability than the competition. We therefore compare optical-output sensors to other classes of measurement systems, e.g. capacitive and piezoresitive sensors, putting the reader in position to draw conclusions about the optimum choice of sensors for a given measurement application.

1.2 Design of miniaturized optics

The main goals of this book are to teach the reader how to design miniaturized optics and to stimulate new inventions in this area. Design and innovation require intuition and simple physical models. The treatment is therefore focused on conceptual understanding and analytical calculations, as opposed to numerical analysis.

All design starts with a concept, and much of the innovation and significant contributions are in the conceptual design, rather than in the detailed plans that follow. Almost all important scientific breakthroughs and technological innovation follow from conceptual thinking. Conceptual descriptions develop intuition and inspire researchers to find new solutions and new applications. They also allow technical experts to share their insights with laymen, and provide a way to convey the range of opportunities that a technology has to offer.

Conceptual design by itself is, however, not sufficient. It must be complemented by fundamental understanding and qualitative models that allow concepts to be tested for viability and detailed implementation plans to be drawn up. The first part, test for viability, is extremely important in optics, because many aspects of optics are counterintuitive. This fact is amply demonstrated by our difficulty in explaining everyday optical phenomena[c] and the longevity of erroneous scientific concepts like the Luminiferous aether. The patent literature is also full of optical devices that are conceived by smart people and deemed sound by competent patent reviewers, but that nevertheless are in violation of fundamental laws of physics.

We can of course avoid submitting patent applications on unphysical "inventions" by simply analyzing our specific optical designs, using one of the many numerical optical analysis software packages that are now commercially available. Showing that a specific implementation doesn't work does not invalidate a concept, however. For that we need fundamental understanding of the physics that govern propagation of electromagnetic waves. For this purpose it is helpful at the conceptual design stage to ask the question: Where does the energy go? It can be a surprisingly difficult question to answer for many optical components. A significant part of this book is therefore dedicated to understanding how energy flows in optical systems and how insight about energy flow can help us avoid pursuing solutions that are unphysical and therefore doomed to fail.

Once the conceptual design is found to be viable, the real work begins. Now we must decide if it is practical and if can be scaled to manageable dimensions given the technology we plan to use. Again numerical analysis tools fall short. All

[c] Try to explain to a child how the sky is blue, how the rainbow looks the way it does, how things look bigger under water, or how stars twinkle.

practical optical systems are too complex to be designed by guessing a solution and then analyzing it to see if it works. That approach is invaluable for fine tuning of sophisticated designs, but it cannot answer questions about the ultimate scaling limits of a conceptual design, about how to integrate different technologies, about the best trade-offs of scaling and complexity (degrees of freedom), and about how to optimize the fabrication technology.

For these higher level design questions we need analytical models that allow parametrical answers to design ("inverse") questions. In other words, we must have the tools to clarify how a given implementation parameter should be chosen to give a desired value to a specific operational characteristic. In addition to inspiring innovation in microoptics, a major goal of the book is to contribute such analytical modeling tools that can be used for design of a wide range of microoptics and nano-photonics. The intention is to develop the tools to the point where they are simple enough that they can be used in the conceptual-design phase, yet powerful enough to bring the designs close enough to completion that they can successfully be refined by numerical methods.

Practical implementations of optical microsystems require interdisciplinary teams that collectively provide knowledge of many fields, including semiconductor processing, mechanical engineering, optical-device design, optical-system fabrication and packaging, as well as application-specific expertise. This book is therefore written to be useful to scientists and engineers of a wide range of backgrounds. No attempt is made at a comprehensive coverage of all optical MEMS systems. For that the reader is referred to a series of well-written books [1], special issues [2,3,4,5], and review articles [6]. Instead the emphasis is on optics fundamentals and on the specific challenges of miniaturization and semiconductor implementation, thereby addressing needs of both semiconductor and MEMS experts who are interested in optical applications, as well as for optics researchers who want to understand how to add microfabrication to their tool chest.

1.3 Roadmap

To answer questions about scaling of optics and to understand the design of miniaturized systems, we must understand how light spreads, or diffracts, in three-dimensional space. Electromagnetic waves, like all wave phenomena, are subject to diffraction during wave propagation. Our starting point for an accurate description of optical diffraction is Maxwell's equations in Chapter 2. Maxwell's equations allow us to derive the Poynting theorem and the Reciprocity Theorem, both of which give valuable insights into the scaling of optical devices. In combination, these theorems simplify the analysis of optical devices, because they allow us to calculate coupling and contrast to be carried out at one, well-chosen physical location with the certainty that these quantities does not change within the device.

We also use Maxwell's equations to derive the wave equation for electromagnetic waves. In Chapter 3 we use it to find the laws of reflection and refraction of plane waves, and to explain important concepts like Total Internal Reflection and evanescent fields. The formalism we develop also lends itself to the analysis of multiple reflections, so we are able to calculate the reflection and transmission of structures that are periodic in one dimension. These structures can be thought of as one-dimensional Photonic Crystals, so our formalism gives us the first glimpse of the functions and operations of this important class of materials.

Plane waves give insight into many optical devices and systems, but do not tell us how small they can be. For that we need to understand diffraction which is the effects that almost always determines the ultimate limit on scaling. In Chapter 4 we therefore study the fundamental properties and propagation of Gaussian beams. Gaussian beams are solutions to the paraxial wave equation, i.e. they are strictly speaking only valid for optical fields that propagate at small angles to the optical axis. In many optical devices, this restriction is of no significance and even for those diffraction problems where non-paraxial effects must be considered for accurate solutions, we still get good physical insight by using the Gaussian-beam model.

Once we understand how light diffracts, we are in a position to realize how to control the spread. Lenses, the traditional tools for controlling light propagation, are well understood and described in numerous text books, so we give them no more than a cursory treatment. Instead, we focus on waveguides, diffraction gratings, and Photonic Crystals. In Chapter 5 we derive the fundamentals of optical waveguides. Again we rely on Maxwell's Equations, and use them to calculate the eigenmodes of waveguides, and study the details of pulse propagation on optical fibers.

In Chapter 6 we describe a set of waveguide devices. The set is chosen to be illustrative of important concepts, rather than to be comprehensive. The chapter starts out with a discussion of the simple but important problem of how to couple light into waveguides. It continues with the development of coupled-mode theory, which enables us to analyze several ubiquitous waveguide components, including directional couplers and Bragg reflectors. The latter is our first detailed look at the concept of Photonic Crystals. Chapter 6 wraps up our treatment of optics fundamentals, which constitute part one of the book.

Part two of the book takes a close look at Optical MEMS. Perhaps the most fundamental of all optical MEMS devices, the optical scanner, is the focus of Chapter 7. MEMS scanners appear in a multitude of complex optical systems with very different applications and requirements. They therefore seem to defy all attempts at classification and systematization of their design. It turns out, however, that optical scanners follow very general scaling laws that greatly simplify the optical design of scanning systems. The mechanical design of optical scanners is more difficult to treat in a general manner, so that subject is discussed through a series of

examples. Optimization of the mechanical design depends on the actuator technology that is applied. A comprehensive coverage of actuator technologies are outside the scope of this book, but the most common and practical types, electrostatic actuators, are covered in Appendix B.

One application of the miniaturized optical scanners is the fiber switch, described in Chapter 8. We classify MEMS fiber switches into two broad groups based on their principle of operation; matrix switches and beam steering switches. Both of these switch types have their own set of implementation challenges, and their range of use depend closely on the available fabrication and packaging technology, but as a general conclusion we find that matrix switches function better for small port counts, while beam steering switches scale better to large port counts.

In Chapter 9 we move on from single devices to arrays. In the first part of the chapter, we derive models for scaling of arrays of rotating micromirrors that operate much like the scanners of Chapter 7 and 8. We think of the operation of rotating mirrors as amplitude modulation, because these mirrors control the amount (amplitude) of the reflected light that is picked up by the output optical system.

In the second part of Chapter 9 we turn our attention to phase-modulating mirrors. These are conceptually more difficult than amplitude-modulating mirrors, because their response depends on the setting of its neighboring mirrors, but we also find that under certain circumstances they scale better; For a given optical projection system, phase-modulating mirrors can create finer detail in the projected image than can be achieved by amplitude-modulating mirrors. We conclude that amplitude modulating-mirrors, due to their simplicity, are better suited to magnifying projection systems (e.g. video projectors), in which resolution is typically limited by the projection lens, while phase modulation is superior in systems that need to create the finest possible detail (e.g. optical lithography systems).

The efficiency and superior scaling of phase-modulating microoptics is a reoccurring theme in the rest of the book. Chapter 9 wraps up with a qualitative description of several optical MEMS systems that use phase modulation to advantage. These systems can be categorized as diffractive optical MEMS, because their operation relies on phase modulation combined with diffraction. Chapter 10 describes the Grating Light Modulator in detail, and derives mathematical models for its operation, with a focus on display applications. Chapter 11 extends the treatment of the Grating Light Modulator to fiber optics, and derives models that are appropriate for this field of use.

In Chapters 7 through 11 we model how microoptics shape and control optical fields. In Chapter 12 we take the opposite perspective and consider how optical fields can be used to measure the state of a microoptical sensor. In other words, we are modeling sensors with optical outputs. Again we compare and contrast amplitude-modulating (optical levers) and phase-modulating (interferometers) systems. We find that the ultimate limits on sensitivity are slightly better for phase-

modulating sensors, but also point out the significant practical advantages of amplitude modulation.

The last chapter in the Optical MEMS part is Chapter 13 on optical filters. As for many other classes of optical MEMS, filters can be based on amplitude or phase modulation, but here we exclusively consider phase-modulating filters[d], because they scale better to small sizes and because they are easier to tune with MEMS actuators.

The last two chapters of the book are dedicated to Nanophotonics, or more specifically Photonic Crystals (PCs), as they relate to optical microsystems. The first part of Chapter 14 describes the fundamentals of PCs, demonstrate how they control optical fields over sub-wavelength distances, and shows how they enable optical devices with improved scaling and functionality compared to traditional optics. Photonic Crystals are complex and a comprehensive treatment require a whole book (see the references in Chapter 14 for suggestions), so here we are giving just a bare-bones introduction to the central concepts.

While the first part of Chapter 14 is general, the second part is focused specifically on a class of devices that are of great utility in microsystems and optical MEMS; two-dimensional, thin-film PCs. We develop a model for the reflection and transmission of 2-D PCs and show that they can be used as filters and high-reflectivity mirrors.

The last chapter of the book covers Photonic Crystal Devices and Systems. We focus on free-space systems, giving only a cursory treatment of waveguides and waveguide devices. Even with this restriction, it is not possible to give a comprehensive coverage. Instead, we make the case for integration of PCs and MEMS, as well as for integration of PCs and optical fibers.

The chapter starts with a description of PC fabrication techniques that are compatible with ICs and MEMS. It continues with the descriptions of a series of PC optical components that are enabled by silicon PCs. The last part of the chapter gives examples of systems based on integration of PCs with MEMS and/or optical fibers.

Of all the chapters of the book, the last one is the closest to the research forefront. Unlike some of the earlier, more fundamental chapters, it will therefore quickly be outdated. That is as it should be. The intention and the hope is that some of the

[d] One could argue that the filters of Fig. 13.6, 13.13, and 13.14 are based on amplitude modulation, because the micromirrors array in the Fourier plane can be amplitude modulators. The key frequency-discriminating component of the filter is the diffraction grating, however, and it is based on interference, i.e. a phase-modulation effect.

innovation and improvements that will make irrelevant this last chapter, or any other part of the book, will have been inspired by this very text.

References

1 M.E. Motamedi (editor), "MOEMS - Micro-Opto-Electro-Mechanical Systems", SPIE Press, 2005.

2 IEEE Journal of Selected Topics in Quantum Electronics, Issue on Optical Micro- and Nanosystems (OMNS), vol. 13, no. 2, March/April, 2007.

3 IEEE Journal of Selected Topics in Quantum Electronics, Issue on Optical Microsystems, vol. 10, no. 3, May/June, 2004.

4 IEEE Journal of Lightwave Technology, Special Issue on Optical MEMS, vol. 21, no. 3, March 2003.

5 IEEE Journal of Selected Topics in Quantum Electronics, Issue on MicroOptoElectroMechanical System (MOEMS), vol. 5, no. 1, Jan/Feb 1999.

6 M. Wu, O. Solgaard, J. Ford, "Optical MEMS for Light Wave Communication" (Invited Paper), Journal of Lightwave Technology, vol. 24, no. 12, December 2006, pp. 4433-4454.

2: Electromagnetic Fields and Energy

2.1 Introduction to Fields and Energy

The study of light is the study of wave propagation. A working knowledge of the wave nature of light is necessary for design and analysis of all optical systems. In some cases we must also add the concept of the photon, i.e. quantization of the optical field, to get a complete understanding, but wave propagation is the foundation that all of optics is built on.

In this chapter we start with Maxwell's equations for electromagnetic fields and from them we derive the wave equation for electromagnetic waves. We then investigate a simple yet important solution to the wave equation in a homogeneous medium: the plane wave. Part of the plane wave solution is the dispersion relation, which is the relationship between the wave vector and the frequency of the optical field. The dispersion relation for plane waves is particularly simple, almost to the point of seeming obvious and therefore of little utility. Later on, however, we'll study optical devices, e.g. optical fibers and photonic crystals that have complex structures. For such devices, the dispersion relationship helps visualizing optical propagation characteristics, and it is therefore an important tool for conceptualizing and designing optics.

In addition to the dispersion relation, the plane-wave solution to the electromagnetic wave equation also demonstrates the importance of the phase of the optical field. We emphasize this by using phasor representation to describe an important optical device: the Michelson interferometer. The phasor representation is another much used graphical tool to conceptualize and design optical devices.

In the last part of the chapter, we go back to Maxwell's equations and derive the Poynting theorem, which we will use to use develop a set of restrictions on the characteristics of loss-less systems that combines optical fields. We will derive a simple matrix formulation for an optical two-port, and generalize the results to multiport systems. A number of examples will be presented to illustrate the power of energy-conservation arguments in optics, and to give a preview of how such arguments will be used in detailed designs appearing in later chapters.

2.2 From Maxwell's Equations to the Wave Equation

Our starting point is Maxwell's equations in their differential form. As in the rest of the book, we are using MKS units. (We attempt to consistently use MKS units throughout, only deviating in instances where common practice has made other units the standard choice). Maxwell's equations then take the following form:

$$\text{Faraday's law:}\quad \nabla\times\vec{E} = -\frac{\partial}{\partial t}\vec{B} \tag{2.1}$$

$$\text{Ampere's law:}\quad \nabla\times\vec{H} = \frac{\partial}{\partial t}\vec{D} + \vec{J} \tag{2.2}$$

$$\text{Gauss's laws:}\quad \nabla\cdot\vec{D} = \rho \tag{2.3}$$

$$\nabla\cdot\vec{B} = 0 \tag{2.4}$$

where E is the electric field, H the magnetic field, D the displacement or electric flux density, B the magnetic flux density, J the electric current density, and ρ the electric charge density.

The flux densities are related to the fields through the constitutive relations:

$$\vec{D} = \varepsilon_0\vec{E} + \vec{P} = \varepsilon\vec{E} \tag{2.5}$$

$$\vec{B} = \mu_0\vec{H} + \mu_0\vec{M} = \mu\vec{H} \tag{2.6}$$

where P is the electric polarization, M is the magnetization, ε is the permittivity, and μ is the permeability. In free space (vacuum) we have:

$$\text{Permittivity of free space:}\quad \varepsilon_0 = 8.8542\cdot 10^{-12}\ \text{[F/m]}$$

$$\text{Permeability of free space:}\quad \mu_0 = 4\pi\cdot 10^{-7}\ \text{[H/m]}$$

In general, both the permittivity and the permeability are functions of frequency in all materials. In practice, we may consider the permeability of most non-magnetic materials to be constant and equal to the permeability of free space. The frequency dependence of the permittivity, on the other hand, must be taken into consideration in many practical optical devices and systems.

Boundary Conditions

We'll now use Stoke's theorem and Gauss's divergence theorem to derive the integral forms of Maxwell's equations from the differential forms. The differential

forms are well suited for the derivations of the boundary conditions that we will use in reflection, transmission, and waveguide calculations.

$$\text{Stoke's theorem:} \quad \int_{Area} \nabla \times \vec{A} \cdot d\vec{S} = \int_{loop} \vec{A} \cdot d\vec{l} \tag{2.7}$$

$$\text{Gauss's divergence theorem:} \quad \int_{Surface} \vec{F} \cdot d\vec{S} = \int_{Volume} \nabla \cdot \vec{F} dv \tag{2.8}$$

where $\vec{S}$ is the surface and $\vec{l}$ is the loop vector.

Using these relations, we can write Maxwell's Equations in integral form:

$$\text{Faraday's law:} \quad \int_{loop} \vec{E} \cdot d\vec{l} = -\frac{\partial}{\partial t} \int_{area} \vec{B} \cdot d\vec{S} \tag{2.9}$$

$$\text{Ampere's law:} \quad \int_{loop} \vec{H} \cdot d\vec{l} = \int_{area} \vec{J} \cdot d\vec{S} + \frac{\partial}{\partial t} \int_{area} \vec{D} \cdot d\vec{S} \tag{2.10}$$

$$\text{Gauss's laws:} \quad \int_{surface} \vec{D} \cdot d\vec{S} = Q_{enclosed} \qquad \int_{surface} \vec{B} \cdot d\vec{S} = 0 \tag{2.11}$$

From Gauss's laws it follows that the normal components of the magnetic and electric flux densities are both continuous. (For the electric flux this requires that there are no surface charges). This, in combination with Faraday's law, shows that the tangential component of the electric field is continuous. Finally, we see from Ampere's law that the magnetic field is continuous if there are no surface currents.

For later reference we repeat these boundary conditions, which are valid in source-free media (ρ=0, J=0):

$$E_t \text{ is continuous:} \quad \vec{S} \times \left(\vec{E}_2 - \vec{E}_1\right) = 0 \tag{2.12}$$

$$H_t \text{ is continuous:} \quad \vec{S} \times \left(\vec{H}_2 - \vec{H}_1\right) = 0 \tag{2.13}$$

$$D_n \text{ is continuous:} \quad \vec{S} \cdot \left(\vec{D}_2 - \vec{D}_1\right) = 0 \tag{2.14}$$

$$B_n \text{ is continuous:} \quad \vec{S} \cdot \left(\vec{B}_2 - \vec{B}_1\right) = 0 \tag{2.15}$$

Wave Equation

The derivation of the wave equation starts with taking the curl of both sides of Faraday's law to get the following expression:

$$\nabla\times\left(\nabla\times\vec{E}\right)=\nabla\times\frac{-\partial}{\partial t}\vec{B}=\nabla\times\frac{-\partial}{\partial t}\mu\vec{H} \tag{2.16}$$

In almost all cases of practical interest it is a good assumption that μ is independent of time and position, so we can write:

$$\nabla\times\left(\nabla\times\vec{E}\right)=-\mu\nabla\times\frac{\partial}{\partial t}\vec{H} \tag{2.17}$$

For continuous functions we can reverse the order of the spatial and temporal derivatives:

$$\nabla\times\left(\nabla\times\vec{E}\right)=-\mu\frac{\partial}{\partial t}\left(\nabla\times\vec{H}\right) \tag{2.18}$$

We now use Ampere's law (assuming zero current density) and the constitutive relation for the electric displacement to find an equation for the electric field, assuming that the permittivity is time invariant:

$$\nabla\times\left(\nabla\times\vec{E}\right)=-\mu\frac{\partial}{\partial t}\left(\frac{\partial}{\partial t}\vec{D}\right)=-\mu\varepsilon\frac{\partial}{\partial t}\left(\frac{\partial}{\partial t}\vec{E}\right) \tag{2.19}$$

To simplify further, we need the vector identity:

$$\nabla\times\nabla\times\vec{A}=\nabla\left(\nabla\cdot\vec{A}\right)-\nabla^2\vec{A} \tag{2.20}$$

where ∇^2 is the linear, three-dimensional Laplacian operator, which in Cartesian coordinates is defined as

$$\nabla^2=\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2} \tag{2.21}$$

When applied to a vector field, the Laplacian must be applied to each vector component separately, i.e.:

$$\nabla^2\vec{A}=\nabla^2A_x\vec{x}+\nabla^2A_y\vec{y}+\nabla^2A_z\vec{z} \tag{2.21}$$

Application of the above vector identity to our equation for the electric field results in the following expression:

$$\nabla\left(\nabla\cdot\vec{E}\right)-\nabla^2\vec{E}=-\mu\varepsilon\frac{\partial}{\partial t}\left(\frac{\partial}{\partial t}\vec{E}\right) \tag{2.22}$$

Now we will assume that the divergence of the electric field is zero. This is not strictly true for fields in media with non-homogeneous permittivities, but in practice it is a good approximation. With this assumption, we arrive at the homogeneous wave equation:

$$\nabla^2\vec{E}-\mu\varepsilon\frac{\partial^2\vec{E}}{\partial t^2}=0 \tag{2.23}$$

Similarly, by starting with Ampere's law, we find the wave equation for the magnetic field:

$$\nabla^2\vec{H}-\mu\varepsilon\frac{\partial^2\vec{H}}{\partial t^2}=0 \tag{2.24}$$

2.3 Plane Waves

A simple solution to the wave equation is a plane wave of the form

$$\vec{E}=\vec{x}\cdot E_0\cdot\cos\left(\omega t-kz\right) \tag{2.25}$$

where $\omega=2\pi\cdot f$ is the natural frequency and k is the wave vector of the plane wave. This particular plane wave is uniform in the x-y plane, and it propagates in the positive z-direction. The E-field points in the x-direction, or in other words, the plane wave is polarized in the x-direction.

The wavevector, k, and the natural frequency, ω, are related as

$$k=\frac{\omega}{v}=\frac{2\pi f}{\lambda\cdot f}=\frac{2\pi}{\lambda} \tag{2.26}$$

where we have also introduced the wavelength $\lambda=\frac{v}{f}$. The corresponding magnetic field can be found from Faraday's or Ampere's laws. Using Faraday's law we find

$$\frac{\partial \vec{H}}{\partial t} = -\frac{1}{\mu_0}\nabla \times \vec{E} = -\vec{y}\frac{kE_0}{\mu_0}\sin(\omega t - kz) \Rightarrow$$
$$\vec{H} = \vec{y}\sqrt{\frac{\varepsilon_0}{\mu_0}}E_0\cos(\omega t - kz) \tag{2.27}$$

Notice that there is no phase variation of either the electric nor the magnetic fields in planes (x,y planes) perpendicular to the direction of propagation (*z* direction), justifying naming these solutions plane waves. The electric and magnetic fields of a uniform plane wave are illustrated in Fig. 2.1.

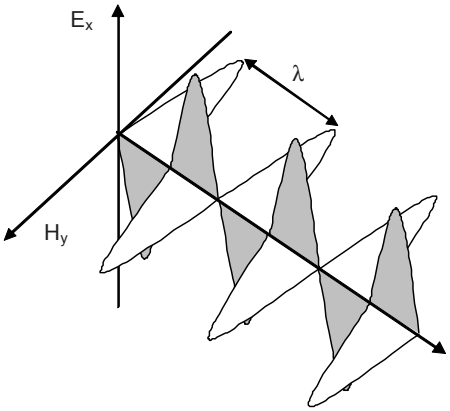


Figure 2.1. Electric and magnetic fields of a uniform plane wave propagating in the z-direction. The fields are mutually orthogonal, and orthogonal to the direction of propagation.

Invalid solutions:

Not all solutions to the wave equation are solutions to Maxwell's equations. Valid solutions must also satisfy Gauss's law for the electrical displacement. An example of a solution to the wave equation, that doesn't represent a valid electric field is

$$\vec{E} = \vec{z} \cdot E_0 \cdot \cos(\omega t - kz) \tag{2.28}$$

Direct substitution shows that this is indeed a solution to the wave equation, but we also observe that due to the fact that the polarization coincides with the propagation direction, we have $\nabla \cdot \vec{D} \neq 0$ in violation of Gauss's law. The plane waves described above do, however, fulfill the requirement $\nabla \cdot \vec{D} = 0$, and are therefore valid mathematical solutions.

We can generalize the plane wave solution to

$$\vec{E} = \vec{E}_0 \cdot f\left(\omega \cdot t - \vec{k} \cdot \vec{r}\right) \tag{2.29}$$

where E_0 is a constant vector orthogonal to the wave vector $\vec{k}$, and *f(x)* is an arbitrary function. This solution represents a wave that is polarized in a direction orthogonal to the direction of propagation as defined by the wave vector $\vec{k}$. The wave is uniform in planes orthogonal to $\vec{k}$, and has a shape given by *f(x)* in the direction of $\vec{k}$.

Substituting the general solution back into the wave equation yields

$$\left(-k_x^2 - k_y^2 - k_z^2 + \omega^2 \mu\varepsilon\right)\vec{E}_0 \cdot f'' = 0 \tag{2.30}$$

where *f''* is the second derivative of *f* with respect to its whole argument. Non-trivial solutions require

$$k^2 = k_x^2 + k_y^2 + k_z^2 = \omega^2 \mu\varepsilon \tag{2.31}$$

This is the dispersion relation for plane waves in a homogeneous medium. In free space or vacuum the dispersion relation is a simple straight line $\omega = \dfrac{k}{\sqrt{\mu_0 \varepsilon_0}}$.

In homogeneous materials, the dispersion relation is non-linear, due to the frequency dependence of the permittivity (and sometimes also the frequency dependence of the permeability). Practical optical devices are of course not made of homogeneous materials, but of complex combinations of materials of different properties. In such structures, the distributions of the optical fields are dependent on frequency, further complicating the dispersion relation. For complex optical devices, optical fibers, Bragg filters, and Photonic Crystals, the dispersion relation contains most of the important information about wave propagation in the structure, and its graphical representation is a valuable tool for visualizing device characteristics.

Phase Velocity

Going back to Eq. 2.25 describing a plane wave propagating in the positive *z* direction, we find that it propagates with a phase velocity given by

$$v = \frac{dz}{dt} = \frac{\omega}{k} = \frac{1}{\sqrt{\mu\varepsilon}} = \frac{c}{n} \tag{2.32}$$

Here we have introduced the reflective index *n*, which for a given material and frequency is defined as the ratio of the speed of light in vacuum to the speed of

light in the material at the given frequency. The speed of light in vacuum $c \equiv \frac{1}{\sqrt{\mu_0 \varepsilon_0}}$ evaluates to $2.998 \cdot 10^8$ m/s.

Group Velocity

From the above expression, we see that the speed of an optical wave at one specific frequency, i.e. a harmonic wave, is simply given by the value of the dispersion relation at that frequency. To understand the propagation of more complicated waves, we must consider superpositions of multiple harmonics.

A superposition of two optical fields at distinct frequencies can be described as

$$\begin{aligned} &E_1 + E_2 = \\ &E_0\left(\cos\left[(\omega + \Delta\omega)t - (k + \Delta k)z\right] + \cos\left[(\omega - \Delta\omega)t - (k - \Delta k)z\right]\right) \end{aligned} \tag{2.33}$$

Using the identity

$$\cos[x + y] + \cos[x - y] = 2\cos x \cdot \cos y \tag{2.34}$$

this can be rewritten

$$E_1 + E_2 = 2E_0 \cos(\omega t - kz) \cdot \cos(\Delta\omega t - \Delta k z) \tag{2.35}$$

We see that the superposition consists of an harmonic wave at the average frequency and average propagation constant, plus an envelope function that propagates at the velocity

$$v_g = \frac{\Delta\omega}{\Delta k} \tag{2.36}$$

This velocity of the envelope is called the group velocity.

In general we find that the group velocity for a superposition of several harmonics can be expressed

$$v_g = \frac{d\omega}{dk} \tag{2.37}$$

In contrast to the phase velocity, the group velocity is not given by the value of the dispersion relation, but the slope of the dispersion relation at a given frequency.

To relate the group velocity to material constants, we perform the following rewrite:

$$v_g = \frac{d\omega}{dk} = \left[\frac{dk}{d\omega}\right]^{-1} = \left[\frac{d}{d\omega}\left(\frac{\omega n}{c}\right)\right]^{-1} = \left[\frac{n}{c} + \frac{\omega}{c}\frac{dn}{d\omega}\right]^{-1}$$
$$= \frac{c}{n + \omega\frac{dn}{d\omega}} = \frac{c}{n + \omega\left(-\frac{\lambda}{\omega}\right)\frac{\partial\omega}{\partial\lambda}\frac{dn}{d\omega}} = \frac{c}{n - \lambda\frac{dn}{d\lambda}} \tag{2.38}$$

The last expression shows that the group velocity is close, but not exactly equal, to the phase velocity.

We say that a material has normal dispersion for those wavelengths where $dn/d\lambda<0$. In other words, we have regular dispersion in regions where the group velocity is less than the phase velocity, c/n. In regions of anomalous dispersion, we have $dn/d\lambda>0$, and the group velocity is larger than the phase velocity. Optical fibers are often used at wavelengths close to the dispersion minimum ($d^2n/d\lambda^2=0$) of glass, so that the group velocity dispersion changes sign in the wavelength range of interest.

Maxwell's equations are first-order differential equations in space and time. They are also linear, provided that the constitutive relations are linear. In this case, superpositions of solutions to Maxwell's equations are themselves solutions, and monochromatic plane waves can be added to form solutions of arbitrary time waveforms

$$\vec{E}(r,t) = \vec{x}\frac{1}{2\pi}\int_0^\infty E_x(\vec{r})\cos(\omega t - k(\omega)z + \phi(\omega))d\omega \tag{2.39}$$

The plane-wave solutions we have found are not square-integrable, so they cannot contain finite energy and can therefore not be physically implemented. Understanding plane waves is nevertheless very useful, because (1) they give us a very good 1st order understanding of wave propagation phenomena, and (2) it is possible to physically realize field distributions that are arbitrarily close to plane waves and (3) sums of plane waves can be used for accurate modeling of many optical systems.

2.4 Phasor Notation

The linearity of Maxwell's equations also let us use phasor notation, which greatly simplifies mathematical manipulation. The plane wave solutions we examined earlier can be written in the following form

$$\begin{aligned}&\vec{E}(\vec{r},t)=\vec{x}\cdot E_0(\vec{r})\cdot\cos(\omega t-\phi_x(\vec{r}))=\\&\vec{x}\cdot\frac{1}{2}\left[E_x(\vec{r})e^{j(\omega t-\phi_x(\vec{r}))}+c.c.\right]=\\&\vec{x}\cdot\mathrm{Re}\left[E_x(\vec{r})e^{j(\omega t-\phi_x(\vec{r}))}\right]=\vec{x}\cdot\mathrm{Re}\left[\hat{E}_x(\vec{r})e^{j\omega t}\right]\end{aligned}\tag{2.40}$$

Notice that in the last identity the phase factor, exp[-jϕx], has been included in the expression for the field amplitude component, which then becomes a complex quantity, a phasor.

In phasor notation we drop the explicit taking of the real part such that the plane wave solution is written

$$\vec{E}(\vec{r},t)=\vec{x}\cdot\hat{E}_x(\vec{r})e^{j\omega t}\tag{2.41}$$

In the remainder of this book we will not use any type of notational identification, but will instead rely on context to distinguish time-harmonic phasors from time-dependent field amplitudes. (An explicit time dependence rules out phasors, which are never time dependent). When calculating using phasors we must remember to always take the real part of the final answer to obtain the corresponding physical entity. We must also be careful to only use phasor notation in linear calculations.

We can combine the three vector-components of the electric field in a single phasor of the form

$$\vec{E}(\vec{r},t)=\vec{E}(\vec{r})e^{j\omega t}\tag{2.42}$$

where the phasor, $\vec{E}(\vec{r})$, has six components (three vector components each with amplitude and phase).

Maxwell's equations in phasor form are

$$\text{Faraday's law: } \nabla\times\vec{E}=-j\omega\vec{B}\tag{2.43}$$

$$\text{Ampere's law: } \nabla\times\vec{H}=j\omega\vec{D}+\vec{J}\tag{2.44}$$

$$\text{Gauss's laws: } \nabla\cdot\vec{D}=\rho\qquad\nabla\cdot\vec{B}=0\tag{2.45}$$

and the wave equation becomes the Helmholz equation

$$\left(\nabla^2+\omega^2\mu_0\varepsilon_0\right)\vec{E}=0\tag{2.46}$$

In phasor notation, the plane wave solution we considered before can be written

$$\vec{E}(z,t) = \mathrm{Re}\left[\vec{E}(z)e^{j\omega t}\right]$$
$$\vec{E}(z) = \vec{x} \cdot E_0 e^{-jkz} \tag{2.47}$$

The corresponding magnetic field is found from Faraday's law in phasor form

$$\vec{H}(z) = -\frac{\nabla \times \vec{E}}{j\omega\mu_0} = \vec{y}\frac{kE_0}{\omega\mu_0}e^{-jkz} \Rightarrow$$
$$\vec{H}(z) = \vec{z} \times \vec{E}/\eta_0 \tag{2.48}$$

where, $\eta_0 = \sqrt{\mu_0/\varepsilon_0} \approx 377\Omega$, is the wave impedance in free space.

2.4.1 Michelson Interferometer – Phasor notation

One of the advantages of the phasor representation is that it provides a simple and easy-to-understand method for explaining many important optical concepts and devices. Consider, for example, the Michelson interferometer shown in Fig. 2.2. The input beam is split into two beams that each propagates to a mirror where they are reflected back into the beam splitter. Upon propagation from the beam splitter to the mirrors and back, the two beams accumulates phase shifts.

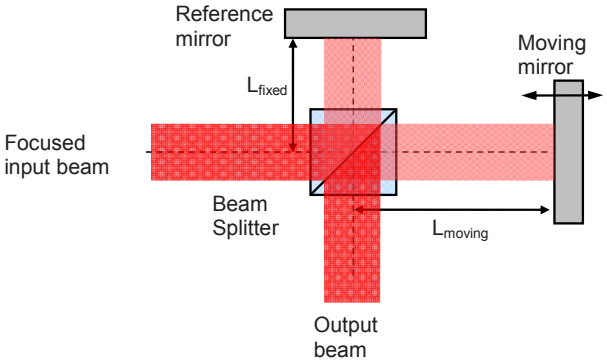


Figure 2.2 *The Michelson interferometer consists of a beam splitter and two mirrors; one fixed reference mirror and one movable target mirror. The incident light is split into the fixed and variable arms of the interferometer and recombined in the beam splitter after reflecting off the mirrors. The phase difference between the two beams upon recombination determines how much light is transferred to the output and how much is reflected backwards along the path of the input beam.*

According to the plane-wave solutions described above, the accumulated phase shift for the beam reflected off the fixed reference mirror is given by[1]:

$$\theta_{fixed} = k \cdot 2L_{reference} = \frac{2\pi \cdot 2L_{reference}}{\lambda} \tag{2.49}$$

The beam reflected from the moving mirror gets a similar phase shift, so the phase difference of the two beams can be expressed:

$$\theta_{\Delta} = k \cdot 2\left(L_{reference} - L_{moving}\right) = \frac{4\pi \cdot \left(L_{reference} - L_{moving}\right)}{\lambda} \tag{2.50}$$

The key to the operation of the Michelson interferometer is the recombination of the two beams in the beam splitter. To understand the result of that recombination we use phasors to represent the two beams in the reference arm and the measurement arm of the interferometer as shown in Fig. 2.3. Here the phasor representing the optical field in the reference arm is held fixed, while the phasor representing the optical field in the measurement arm is given a phase equal to the phase difference of the two beams.

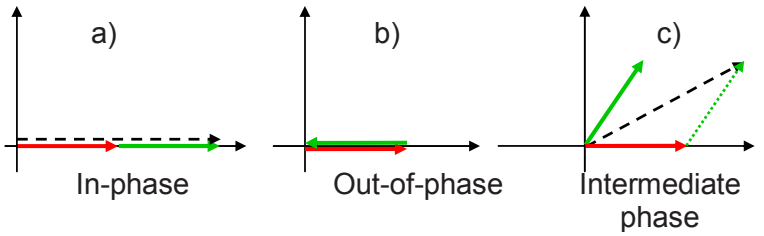


Figure 2.3. *Phasor representation of the output optical fields of a Michelson interferometer. The two phasors representing the light reflected from the fixed and moving mirrors are drawn as solid lines, while the total output field is drawn as a dashed line. In (a) the two reflected parts are in phase, resulting in a maximum value for the output (shown offset for clarity). In (b) the two reflected parts are exactly out of phase, so the output is zero. In (c) the phase difference between the two reflected parts is between zero and π radians, so the resulting output field is between zero and its maximum value.*

Figure 2.3a) is the phasor representation of the physical situation where the path length difference between the reference arm and the measurement arm is zero or an integer number of half wavelengths. The two phasors are then in phase and their sum attains its maximum value, which is the sum of the absolute values of

[1] Collimated or focused beams accumulated phase differently than plane waves as we shall see in the Chapter 4, but the difference is small and can be ignored in the Michelson interferometer.

the two parts. In this state, the relative size of the two phasors representing the parts of the field is unimportant.

In Fig. 2.3b) the path-length difference for the light propagating in the two arms of the interferometer is π radians, i.e. the two parts of the reflected light are in exactly opposite phase. The result is that there is no output light from the interferometer in this state. The incident light is therefore completely backreflected. It is clear from the figure that we only get complete suppression of reflection when the two phasors representing the two parts of the reflected light are equal so that they exactly cancel each other when they are in opposite phase. For the Michelson interferometer of Fig. 2.2, this means that the beam splitter must divide the incoming optical field in two equal parts.

The usefulness of the phasor representation becomes clear when we consider Fig. 2.3c that shows the reflected light when the two parts of the reflections have a relative phase between zero and π radians. The resulting output field now has a value that is somewhere in between zero for the out-of-phase configuration and the maximum value for the in-phase configuration.

We can find the resulting reflected field for an arbitrary relative phase, θ, by vector summation. Here we are not interested in the absolute phase of the reflected light, so we write:

$$\begin{aligned} &\vec{E}_{out} = \vec{E}_{reference} + \vec{E}_{measurement} \Rightarrow \\ &\left|E_{out}\right| = \sqrt{\left(\left|E_{reference}\right| + \left|E_{measurement}\right| \cdot \cos\theta\right)^2 + \left(\left|E_{measurement}\right| \cdot \sin\theta\right)^2} \end{aligned} \tag{2.51}$$

where $\vec{E}_{out}$ is the total output field, $\vec{E}_{reference}$ is the reflected field from the reference mirror, and $\vec{E}_{measurement}$ is the reflected field of the measurement arm.

To simplify the calculations, we assume that the reflected fields from the ribbons and from the substrate are of equal magnitude, i.e. $\left|E_{reference}\right| = \left|E_{measurement}\right|$. The output field then becomes

$$\begin{aligned} \left|E_{out}\right| &= \left|E_{reference}\right| \cdot \sqrt{1 + 2 \cdot \cos\theta + \cos^2\theta + \sin^2\theta} \\ \left|E_{out}\right| &= \left|E_{reference}\right| \cdot \sqrt{2 \cdot (1 + \cos\theta)} \\ \left|E_{out}\right| &= 2 \cdot \left|E_{reference}\right| \cdot \cos\frac{\theta}{2} \end{aligned} \tag{2.52}$$

As we will see in the next section, the optical power is proportional to the square of the optical field, so we can write the following expression for the optical output power from the interferometer:

$$P_{out} = P_{incident} \cdot \cos^2 \frac{\theta}{2} \tag{2.53}$$

where P_{out} and $P_{incident}$ are the reflected and incident optical powers, respectively. The light that is not transmitted to the output must be reflected, so the back-reflected power, $P_{reflected}$, can be expressed as:

$$P_{reflected} = P_{incident} \cdot \left(1 - \cos^2 \frac{\theta}{2}\right) = P_{incident} \cdot \sin^2 \frac{\theta}{2} \tag{2.54}$$

These two simple harmonic expressions for the output and back-reflected optical powers are shown graphically in Fig. 2.4. The curves confirm our intuition that says that to have zero output, the interfereing beams must have a relative phase of π radians.

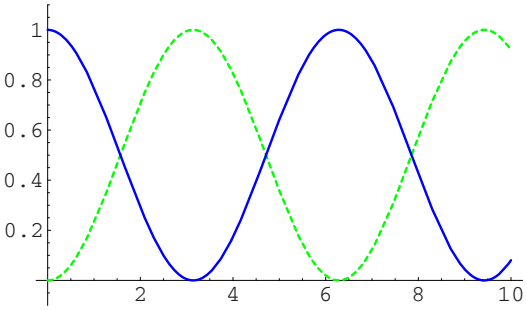


Figure 2.4. *Output (solid) and back-reflected (dashed) optical powers of the Michelson Interferometer as a function of relative phase difference of the two beams propagating through the interferometer. Both the output and back-reflected optical powers are harmonic functions of the phase difference.*

This example shows the usefulness of the phasor representation, not so much in calculations, but in explaining device operation. We will find that phasors are even more useful in the design process, where their simple and intuitive form helps clarify implementation and optimization strategies.

2.5 The Poynting Theorem

We will now use Maxwell's equations to derive expressions for energy transport, dissipation, and storage. Combining Faraday's law with the constitutive relation for the electric field, we get

$$\nabla \times \vec{E} = -\frac{\partial}{\partial t}\mu_0\left(\vec{H} + \vec{M}\right) \tag{2.55}$$

Taking the scalar product of H with this equation results in

$$\vec{H} \cdot \nabla \times \vec{E} = \vec{H} \cdot \frac{\partial}{\partial t}\mu_0\left(\vec{H} + \vec{M}\right) = -\frac{\mu_0}{2}\frac{\partial}{\partial t}\left(\vec{H} \cdot \vec{H}\right) - \mu_0 \vec{H} \cdot \frac{\partial}{\partial t}\vec{M} \tag{2.56}$$

Similarly, combining Ampere's law with the constitutive relation for the magnetic field, and forming the scalar product with E gives

$$\vec{E} \cdot \nabla \times \vec{H} = \frac{\varepsilon_0}{2}\frac{\partial}{\partial t}\left(\vec{E} \cdot \vec{E}\right) + \varepsilon_0 \vec{E}\frac{\partial}{\partial t}\vec{P} + \vec{E} \cdot \vec{J} \tag{2.57}$$

We now subtract these last two equations, and use the vector identity

$$\nabla \cdot \left(\vec{A} \times \vec{B}\right) = \vec{B} \cdot \nabla \times \vec{A} - \vec{A} \cdot \nabla \times \vec{B} \tag{2.58}$$

to get:

$$\begin{aligned} &-\nabla \cdot \left(\vec{E} \times \vec{H}\right) = \\ &\frac{\partial}{\partial t}\left(\frac{\varepsilon_0}{2}\vec{E} \cdot \vec{E} + \frac{\mu_0}{2}\vec{H} \cdot \vec{H}\right) + \vec{E}\frac{\partial}{\partial t}\vec{P} + \mu_0 \vec{H}\frac{\partial}{\partial t}\vec{M} + \vec{E} \cdot \vec{J} \end{aligned} \tag{2.59}$$

Now apply Gauss's dispersion theorem to arrive at the following expression

$$\begin{aligned} &-\int_{surface}\left(\vec{E} \times \vec{H}\right) \cdot d\vec{S} = \\ &\int_{volume}\left[\begin{array}{l} \frac{\partial}{\partial t}\left(\frac{\varepsilon_0}{2}\vec{E} \cdot \vec{E} + \frac{\mu_0}{2}\vec{H} \cdot \vec{H}\right) + \\ \vec{E}\frac{\partial}{\partial t}\vec{P} + \mu_0 \vec{H}\frac{\partial}{\partial t}\vec{M} + \vec{E} \cdot \vec{J} \end{array}\right] dv \end{aligned} \tag{2.60}$$

This is the Poynting theorem that specifies the flow and storage of electromagnetic power.

The different parts of the Poynting theorem can be interpreted as follows: The vector product $\vec{E} \times \vec{H}$ is the Poynting vector, so the surface integral

$$-\int_{surface}\left(\vec{E} \times \vec{H}\right) \cdot d\vec{S} \tag{2.61}$$

represents the power flowing into the volume enclosed by the surface.

The first term on the right hand side

$$\int_{volume}\left[\frac{\partial}{\partial t}\left(\frac{\varepsilon_0}{2}\vec{E}\cdot\vec{E}+\frac{\mu_0}{2}\vec{H}\cdot\vec{H}\right)\right]dv \tag{2.62}$$

is the rate of change of energy stored in the vacuum electromagnetic field.

The second term

$$\int_{volume}\vec{E}\frac{\partial}{\partial t}\vec{P}dv \tag{2.63}$$

is the power absorbed by the dielectric dipoles in the volume. This quantity is positive in materials in thermal equilibrium, but can be made negative to create optical gain in optical amplifiers and lasers.

The third term on the right hand side

$$\int_{volume}\mu_0\vec{H}\frac{\partial}{\partial t}\vec{M}dv \tag{2.64}$$

is the power dissipated by magnetic dipoles. This term can most often be neglected.

Finally, the last term

$$\int_{volume}\vec{E}\cdot\vec{J}dv \tag{2.65}$$

is simply the power lost to the moving charges.

To cast the Poynting theorem (Eq. 2.60) in terms of phasors, we, we use the fact that the product of two harmonic functions is

$$\begin{aligned}&A(t)\cdot B(t)=A\cos(\omega t+\phi)\cdot B\cos(\omega t+\theta)=\\&AB[\cos(\phi-\theta)+\cos(2\omega t+\phi-\theta)]\end{aligned} \tag{2.66}$$

The time average over one period of this product is

$$\begin{aligned}&\langle A(t)\cdot B(t)\rangle=\frac{1}{T}\int_0^T AB[\cos(\phi-\theta)+\cos(2\omega t+\phi-\theta)]dt=\frac{AB}{2}\cos(\phi-\theta)=\\&\mathrm{Re}\left[\frac{AB}{2}e^{j\phi}e^{-j\theta}\right]=\frac{1}{2}\mathrm{Re}\left[AB^*\right]\end{aligned} \tag{2.67}$$

This derivation is valid for vector products as well as scalar products, so we find for the time averaged Poynting vector

$$\left\langle \vec{E}(t)\times\vec{H}(t)\right\rangle = \frac{1}{2}\mathrm{Re}\left[\vec{E}\times\vec{H}^{*}\right] \tag{2.68}$$

Based on this equation, we define the complex Poynting vector as

$$\vec{E}\times\vec{H}^{*} \tag{2.69}$$

Here it should be emphasized that the complex Poynting vector is not the phasor representation of the real Poynting vector, i.e. we do not find the real Poynting vector by multiplying the complex vector by $e^{j\omega t}$ and taking the real part. Instead the real and complex Poynting vectors are related as defined in Eq. 2.68.

With these definitions, the complex Poynting theorem is

$$\begin{aligned} &-\int_{surface}\left(\vec{E}\times\vec{H}^{*}\right)\cdot d\vec{S} = \\ &\int_{volume}\left[j\omega\left(\frac{\varepsilon_0}{2}\vec{E}\cdot\vec{E}^{*}+\frac{\mu_0}{2}\vec{H}\cdot\vec{H}^{*}+\vec{E}\vec{P}^{*}+\mu_0\vec{H}\vec{M}^{*}\right)+\vec{E}\cdot\vec{J}^{*}\right]dv \end{aligned} \tag{2.70}$$

2.6 Combination of Optical Fields from Separate Sources

Now let's consider a simple, but important, consequence of the Poynting Theorem. Consider an optical field passing through a loss-less optical system as shown in Fig. 2.5. Let the input field consist of a sum of two separate distributions, so that we can write the input electric field as $E_{in1}+E_{in2}$, and the input magnetic field as $H_{in1}+H_{in2}$.

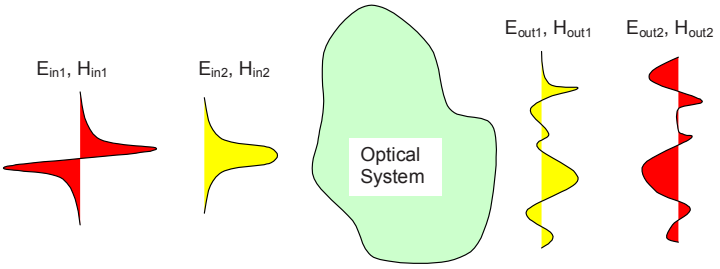


Figure 2.5. *Energy conservation requires that a loss-less optical system without storage cannot mix, or combine, modes, i.e. the overlap integral of two (or more) optical field distributions is unchanged by transmission through the system. Orthogonal modes remain orthogonal, and "optical communications channels" retain their cross talk when transitioning through the system.*

Because the optical system is loss less and without energy storage, the Poynting theorem simply states

$$\begin{aligned}&\int_{surface}\left[\left(\vec{E}_{in1}+\vec{E}_{in2}\right)\times\left(\vec{H}_{in1}+\vec{H}_{in2}\right)\right]\cdot d\vec{S}=\\&-\int_{surface}\left[\left(\vec{E}_{out1}+\vec{E}_{out2}\right)\times\left(\vec{H}_{out1}+\vec{H}_{out2}\right)\right]\cdot d\vec{S}\Rightarrow\end{aligned}\tag{2.71}$$

$$\begin{aligned}&\int_{surface}\left(\vec{E}_{in1}\times\vec{H}_{in1}+\vec{E}_{in1}\times\vec{H}_{in2}+\vec{E}_{in2}\times\vec{H}_{in1}+\vec{E}_{in2}\times\vec{H}_{in2}\right)\cdot d\vec{S}=\\&-\int_{surface}\begin{pmatrix}\vec{E}_{out1}\times\vec{H}_{out1}+\vec{E}_{out1}\times\vec{H}_{out2}\\+\vec{E}_{out2}\times\vec{H}_{out1}+\vec{E}_{out2}\times\vec{H}_{out2}\end{pmatrix}\cdot d\vec{S}\end{aligned}\tag{2.72}$$

This is true for arbitrary fields, so the energy in each field must be conserved, which means that the energy in the cross terms also must be conserved.

$$\begin{aligned}&\int_{surface}\left(\vec{E}_{in1}\times\vec{H}_{in2}+\vec{E}_{in2}\times\vec{H}_{in1}\right)\cdot d\vec{S}=\\&-\int_{surface}\left(\vec{E}_{out1}\times\vec{H}_{out2}+\vec{E}_{out2}\times\vec{H}_{out1}\right)\cdot d\vec{S}\end{aligned}\tag{2.73}$$

Using the complex Poynting theorem, this equation becomes:

$$\begin{aligned}&\int_{surface}\left(\vec{E}_{in1}\times\vec{H}^{*}_{in2}+\vec{E}_{in2}\times\vec{H}^{*}_{in1}\right)\cdot d\vec{S}=\\&-\int_{surface}\left(\vec{E}_{out1}\times\vec{H}^{*}_{out2}+\vec{E}_{out2}\times\vec{H}^{*}_{out1}\right)\cdot d\vec{S}\end{aligned}\tag{2.74}$$

This expression says that a loss-less optical system cannot combine optical fields. If the two input modes are orthogonal, in the sense that their cross-term Poynting vectors integrated over the surface of the optical system is zero, then so are the output modes. If the input modes are not orthogonal, then the Poynting-vector integral, or overlap integral, as defined above, is conserved.

This simple statement of energy conservation is very helpful in analyzing and designing optical communication devices and systems. In loss-less waveguides this theorem tells us that the power carried by orthogonal modes is the sum of the power in the individual modes. In this book we will use the conservation of the overlap integral to guide our analysis of waveguide couplers, optical scanners, fiber switches, displays and imaging devices, as well as other optical systems, in which cross talk between optical channels is important.

The second law of thermodynamics

We can also prove that loss-less optical systems cannot combine modes by considering incoherent sources in thermal equilibrium as shown in Fig. 2.6. Each of the inputs to the optical systems accepts a single mode from the surroundings, which are at a constant temperature. The energy that enters the device is therefore the same at each port. If the two single mode input could be combined into one single mode output, then there would be net energy flow from the input side to the output side in violation of the second law of thermodynamics.

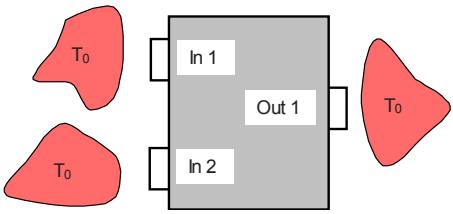


Figure 2.6. Hypothetical loss-less, linear optical device with two inputs and one output.

2.7 Analysis Based on Energy Conservation - Examples

To appreciate the energy-conservation argument derived in section 2.5, we will consider a few examples of practical value. Energy methods are not often used for detailed calculations, but, as we shall see, they can be very powerful conceptual tools for understanding the basics of optical devices.

2.7.1 "Collimated Optical Beam"

Let us now assume that there exists a field distribution of finite width that will propagate through free-space without changing its profile as shown in Fig. 2.7. Two such collimated beams that are intersecting, but propagating at slightly different angles, will then in general have a non-zero overlap integral in the region of intersection, but not outside.

This is clearly in violation of the Poynting theorem. We must therefore conclude that collimated beams of finite cross sections are impossible. In other words, diffraction of electromagnetic waves is a direct consequence of energy conservation.

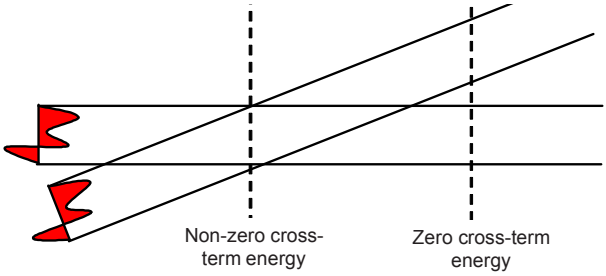


Figure 2.7. *Intersecting, collimated optical beams of finite cross sections have non-zero cross-term energy in the region of intersection, but not outside this region. This is in violation of the Poynting theorem and therefore impossible.*

2.7.2 Combination of optical beams – Fan-in

Now consider the hypothetical device shown in Fig. 2.8. It has two input ports and only one output port. All the energy entering the device must therefore exit through the same port, which means that the cross-term energy is different on the inputs and outputs. This is violating energy conservation, which means that the device is unphysical.

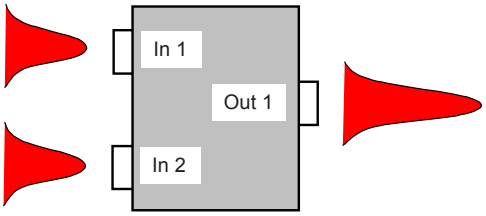


Figure 2.8. *Hypothetical loss-less, linear optical device with two inputs and one output. This operation is in violation of the Poynting theorem and therefore unphysical.*

A beam combiner with two outputs as shown in Fig. 2.9 is physically possible and realizable. In this device the energy from the two inputs is divided between the two outputs such that the cross-energy is zero, and the total energy on each output port depends on the relative phase and amplitude of the input modes.

It should be stressed that the conclusions drawn in these examples are strictly speaking valid only for loss less devices. Small amounts of loss will, however, not significantly change the conclusions. The maximum possible change in cross-term energy will be limited by the loss, so practical devices with some loss will in essence behave similarly to the loss less devices of these examples.

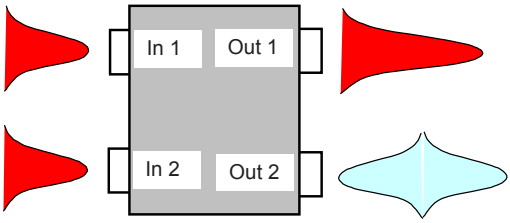


Figure 2.9. *Realizable loss-less, linear optical device with two inputs and two outputs. The distribution of the total output energy between the two output ports depends on the relative phase of the input modes. In the specific shown, the phases of the inputs are chosen such that the outputs are in phase on Output 1 and out-of phase on Output 2. Consequently, all the input power is directed to Output 1.*

2.7.3 Optical devices with two inputs and two outputs – General Case

It is useful to combine the concepts of energy conservation and reciprocity to arrive at a general description of loss-less, linear, optical two ports. Consider first a device with two inputs and two outputs, like the one shown in Fig. 2.9, in which the a fraction x (*0<x<1*) of the energy from Input 1 is coupled to Output 1 and the rest is coupled to Output 2. Similarly, a fraction y (*0<y<1*) of the energy from Input 2 is coupled to Output 2 and the rest is coupled to Output 1. Assume also that the cross-term energy is zero on the input side. The Poynting theorem requires that

$$\begin{aligned} &\int_{surface}\left(\vec{E}_{out1}\times\vec{H}^{*}_{out2}+\vec{E}_{out2}\times\vec{H}^{*}_{out1}\right)\cdot d\vec{S}= \\ &\int_{surface}\left(\vec{E}_{out1}\cdot\vec{E}^{*}_{out2}+\vec{E}_{out2}\cdot\vec{E}^{*}_{out1}\right)dS=0\Rightarrow \\ &\int_{surface}\left(\vec{E}_{out1}\cdot\vec{E}^{*}_{out2}\right)dS=0 \end{aligned} \tag{2.75}$$

where we have assumed without loss of generality that the Poynting vector is perpendicular to the surface enclosing the optical device at the outputs.

We assume that the two outputs are indeed separate, i.e. their overlap integral is zero, so we can write:

$$\begin{aligned}
&\int_{surface}\left(\vec{E}_{out1}\cdot\vec{E}^{*}_{out2}\right)dS = \\
&\int_{output1}\left(\sqrt{x}\vec{E}_{out1}\cdot\sqrt{1-y}\vec{E}^{*}_{out2}\right)\cdot dS \\
&+e^{j\phi}\int_{output2}\left(\sqrt{1-x}\vec{E}_{out1}\cdot\sqrt{y}\vec{E}^{*}_{out2}\right)dS = 0 \\
&\Rightarrow e^{j\phi_2}\sqrt{x}\cdot\sqrt{1-y}+e^{j\phi_2}\sqrt{1-x}\cdot\sqrt{y}=0
\end{aligned} \tag{2.76}$$

In this equation we have included phase terms that reflect the fact that we have not made stipulations about the relative phase of the two fields on each output. Given that both variables x and y are real quantities in the closed interval from zero to one, the only solutions to this equation are

$$\begin{aligned}
&\phi_1-\phi_2=\pi \\
&x=y
\end{aligned} \tag{2.77}$$

We conclude that a two-input, two-output optical device divides the optical power from each input symmetrically, and that the relative phases on the two outputs are different by π, i.e. if the two parts of the output are in-phase on Output 1, then they must be exactly out of phase on Output 2.

The special case of a symmetric device with $x=y=0.5$, is especially interesting because it allows all the energy to be directed to one or the other of the outputs depending on the relative phases of the input modes. This is the case which is depicted in Fig. 2.9. Reciprocity further guarantees that when the device is operated in reverse, the energy splitting from the outputs to the inputs matches that of the splitting from the inputs to the outputs.

2.7.4 Dielectric interface

Now we consider perhaps the simplest possible optical two-port; a planar dielectric interface as shown in Fig. 2.10. Here we have plane waves at normal incidence on either side a planar interface between two dielectrics e.g. air and glass. The plane waves are partially reflected and partially transmitted, so that the interface can be looked as an optical device with two inputs (plane waves normally incident from each side of the interface) and two outputs (the reflected waves from each side of the interface).

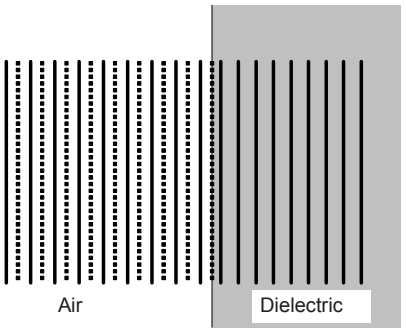


Figure 2.10. Lossless, dielectric interface with plane wave at normal incidence. Energy conservation requires that the field reflectivities from opposite sides of the interface have the same magnitude, but opposite phase when referred to the same plane of reference. In this illustration the incident (solid-line wave fronts) and reflected light (dashed-line wave fronts) on the left are out of phase (reflection from a high-index dielectric). A plane wave incident from the right would be in-phase with its reflection.

In the next section we will derive equations for the reflectivity of the electrical fields at the interface, but for now we will simply assume that the reflectivity for the incident planes is r_1. The relationship between the magnitudes of the E and H fields in the reflected plane wave is the same as for the incident planes wave, so the power reflectivity is given by $R_1 = r_1^2$.

It follows directly from the above calculations of the power distribution in a lossless optical two port that the power reflectivity of the plane wave incident from the right is the same, i.e. $R_2 = R_1$, while the field reflectivity has the same magnitude, but the exactly opposite sign, $r_2 = -r_1$. In other words, the reflectivities from opposite sides of an interface when referred to the same reference plane are of opposite signs. This simple consequence of energy conservation is important in all interferometric devices that include dielectric interfaces, and later in this book we will use it in modeling of many types of optical devices.

2.7.5 Y-coupler

With our new insight into the power flow in optical devices, we are now in a position to understand the operation of the y-coupler, which is a common integrated-optics device. A schematic design of a Y-coupler is shown in Fig. 2.11.

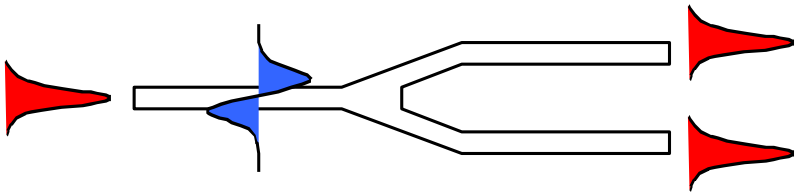


Figure 2.11. Integrated optics Y-coupler. The single mode waveguide on the input side is split into two single mode outputs. Energy conservation requires that the outputs couple to another input mode, which is depicted as an anti-symmetric higher-order mode on the input waveguide.

The Y-coupler consists of a single-mode waveguide, which is split into two single-mode waveguides. We know from energy conservation that the two outputs cannot both couple all their energy into the single-mode output, so we must postulate an input radiation mode that couples to both output modes. For the Y-coupler to work as a perfect power splitter, we must have that the two outputs both couple 50% of their energy into the same radiation mode in such a way that if the two output modes are in phase, the relative phase of the radiation modes to which they couple is exactly π radians out of phase.

2.7.6 Fan-in loss

The argument about the conservation of the cross-term energy can easily be extended to optical devices with arbitrary numbers of input and output ports. For a power splitter that splits the power of a single mode input into *n* single mode outputs, we must obviously have a splitting loss of *1/n*, as shown in Fig. 2.12. It follows from the preceding that the power loss from one of the outputs to the input must also be *1/n*, and that the lost power is coupled to radiation modes. The fan-out loss is obvious, but the fan-in loss is just as fundamental.

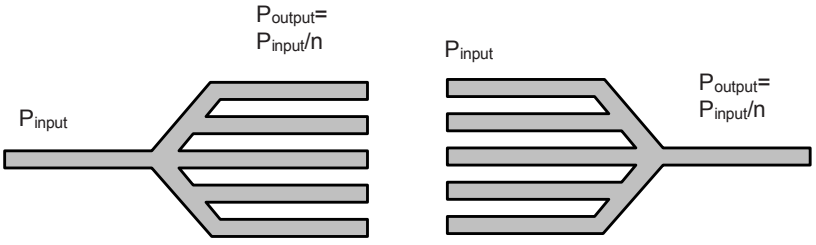


Figure 2.12. Schematic drawing of a n-way power splitter/combiner, which has a 1/n power loss in either direction.

2.7.7 Practical beam combiner

A polarizing beam splitter is often used to combine two optical beams as shown in Fig. 2.13. In this device we have two input modes and two output modes as indicated. Input mode 1 is vertically polarized, and Input mode 2 is horizontally polarized as shown. If we also chose the output modes to be the vertically and horizontally polarized beams (i.e. the outputs and inputs have the same polarizations), then we simply have that *Input 1* couples to *Output 1* and *Input 2* couples to *Output 2*. Clearly we cannot combine the power from both inputs in either of these outputs. If we chose the outputs polarizations to be at 45 degrees to the input polarizations, however, we get a splitting of each input between the two outputs. In this case the power of the two inputs can be combined into one output provided that the input phases are chosen correctly.

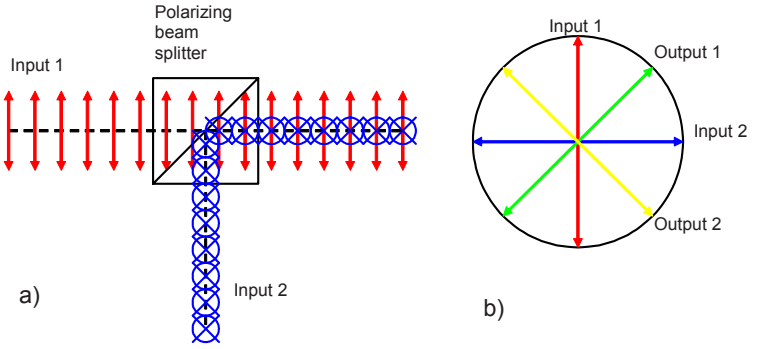


Figure 2.13. Optical power combiner based on polarizing beam splitter. The physical layout of the power combiner is shown in a), while the polarizations of the input and output channels are shown in b.

If the two inputs are incoherent, the output is best described as two linear polarizations with an arbitrary phase. This type of power combiner is often used with pump lasers for fiber amplifiers, because we want to pump the amplifier such that both polarizations of the fiber mode are amplified.

If the two inputs are in phase (this is the case if the inputs originate from the same source, if the input sources are phase locked, or if the relative phase of the input sources are monitored and carefully controlled), the power from the inputs can be combined in a single mode. This is called coherent beam combination.

2.7.8 Wavelength Division Multiplexing

In the preceding discussion, a mode is simply a degree of freedom in the description of an optical field. So far we have discussed spatial modes and polarization modes. Optical fields can clearly also be described in terms of spectral or tempo-

ral modes. There is nothing in our treatment that disallows the spatial combining of modes that are distinct both spatially and spectrally. This is not only possible, but also widespread, and the basis of Wavelength-Division-Multiplexed (WDM) fiber optic communication systems.

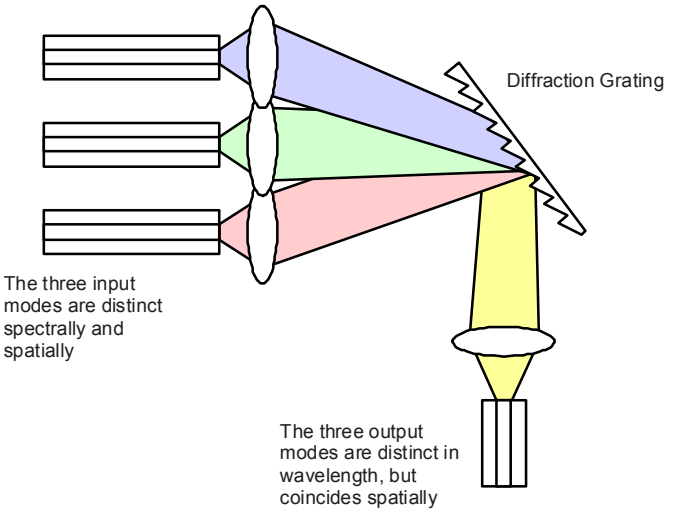


Figure 2.14. Free-space WDM wavelength channels combiner. The three different wavelengths are combined spatially by their different diffraction angle from the grating.

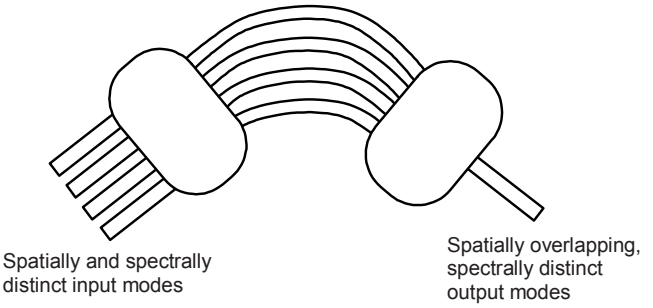


Figure 2.15. Integrated-optics implementation of WDM wavelength channels combiner. The light from the input is evenly distributed over the array of waveguides in the center of the device. The different propagation delay through the center section allows the different wavelength channels to preferentially couple to one of the spatially distinct output waveguides. This device is called an Array Waveguide Grating (AWG).

Figures 2.14 and 2.15 show two different wavelength-channel combiners for WDM communication systems. Conceptually these two solutions are very similar. The main difference is that the grating-based combiner of Fig. 2.14 relies on free-space propagation of the optical beams, while the device of Fig. 2.15 is implemented in a waveguide environment.

2.8 Summary of Fields and Waves

In this chapter we have introduced the important concept of an electromagnetic plane wave, and we have shown how it can be presented in the compact and convenient phasor notation. We will use the plane wave as a simple model for electromagnetic wave propagation when analyzing and designing a wide range of complex optical devices in this book.

In the last part of the chapter, we derived the Poynting Theorem and showed how it electromagnetic wave relate to energy flow in optical devices. Energy methods are not used to for detailed analysis or to construct detailed models for optical devices, but they provide valuable insight into the fundamentals and limitations of optics. As such they are invaluable for clarifying limitations, building design intuition, and generalizing results from specific analyses to wider classes of problems.

In analyzing and designing optics, w often find it useful to ask the question: Where does the optical energy go? The answer helps often illuminates issues that are obscured by detailed and complex field calculations. We will therefore use the energy methods derived in this Chapter extensively throughout this book.

The most important theoretical concept that we have introduced in this chapter are summarized in the following:

Boundary conditions (valid in source-free media (ρ=0, J=0)):

$$E_t \text{ is continuous: } \vec{S}\times\left(\vec{E}_2-\vec{E}_1\right)=0 \tag{2.78}$$

$$H_t \text{ is continuous: } \vec{S}\times\left(\vec{H}_2-\vec{H}_1\right)=0 \tag{2.79}$$

$$D_n \text{ is continuous: } \vec{S}\cdot\left(\vec{D}_2-\vec{D}_1\right)=0 \tag{2.80}$$

$$B_n \text{ is continuous: } \vec{S}\cdot\left(\vec{B}_2-\vec{B}_1\right)=0 \tag{2.81}$$

Wave equations:

Wave equation for the electric field:

$$\nabla^2 \vec{E} - \mu\varepsilon \frac{\partial^2 \vec{E}}{\partial t^2} = 0 \tag{2.82}$$

Wave equation for the magnetic field:

$$\nabla^2 \vec{H} - \mu\varepsilon \frac{\partial^2 \vec{H}}{\partial t^2} = 0 \tag{2.83}$$

Plane Waves:

Plane waves in phasor notation:

$$\begin{aligned} \vec{E}(z,t) &= \mathrm{Re}\left[\vec{E}(z)e^{j\omega t}\right], \\ \vec{E}(z) &= \vec{x} \cdot E_0 e^{-jkz} \end{aligned} \tag{2.84}$$

The corresponding magnetic field is found from Faraday's law:

$$\vec{H}(z) = -\frac{\nabla \times \vec{E}}{j\omega\mu_0} = \vec{y}\frac{kE_0}{\omega\mu_0}e^{-jkz} \Rightarrow \vec{H}(z) = \vec{z} \times \vec{E}/\eta_0 \tag{2.85}$$

where, $\eta_0 = \sqrt{\mu_0/\varepsilon_0} \approx 377\Omega$, is the wave impedance in free space.

The complex Poynting theorem:

$$\begin{aligned} &-\int_{surface} \left(\vec{E} \times \vec{H}^*\right) \cdot d\vec{S} = \\ &\int_{volume} \left[j\omega \left(\frac{\varepsilon_0}{2}\vec{E} \cdot \vec{E}^* + \frac{\mu_0}{2}\vec{H} \cdot \vec{H}^* + \vec{E}\vec{P}^* + \mu_0 \vec{H}\vec{M}^* \right) + \vec{E} \cdot \vec{J}^* \right] dv \end{aligned} \tag{2.86}$$

Energy conservation as expressed by the Poynting Theorem leads to the following conclusions:

1. Loss-less optical devices must have an equal number of input and output modes (i.e. the number of modes cannot be reduced by passing through a loss-less optical device).
2. Collimated beams are impossible, i.e. diffraction is an inevitable consequence of energy conservation.
3. Two-input, two-output optical devices divides optical power from each input symmetrically, and that the relative phases on the two outputs are

different by π, i.e. if the two parts of the output are in-phase on Output 1, then they must be exactly out of phase on Output 2.

4. The reflectivities from opposite sides of an interface when referred to the same reference plane are of opposite signs. This is important in all interferometric devices that include dielectric interfaces.
5. Fan-in and fan-out losses are equal in loss-less optical devices.
6. If the inputs to an optical device are distinct, but the outputs consists of non-zero responses to more than one input, then the relative phases of the different parts of the outputs are determined by the requirement that the total cross-term energy must equal zero.
7. Channels that are distinct in *n* dimensions (space, polarization, spectra) can be combined in *n-1* of these.

Further Reading

H.A. Haus, "Waves and Fields in Optoelectronics, Prentice Hall, 1984.

A. Yariv, P. Yeh, "Photonics: Optical Electronics in Modern Communications", 6th edition, Oxford University Press, 2007.

Exercises

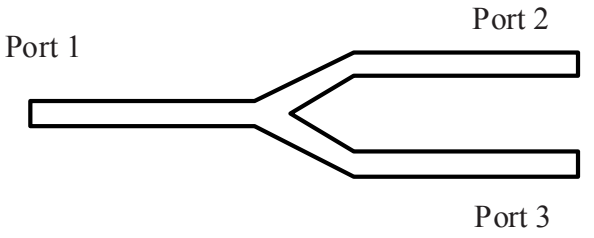
Problem 2.1 – Optical Black Box

A much used principle for determining if a linear, loss-less "optical black box" with n inputs and n outputs is realizable, is that it must be possible to deduce the inputs without ambiguity if the outputs are known (amplitude and phase). Explain how this principle can be deduced from energy conservation.

Problem 2.2 – Realizable Y-junctions

Which of the following optical devices are theoretically possible to implement, and which ones are not. Explain your reasoning.

a) A symmetric Y-coupler as shown in the figure in which the incident light (Port 1) in the single mode input is coupled equally into the two single mode output channels (Ports 2 and 3) with 45% of the total power in each.



b) A symmetric Y-coupler in which 75% of the incident light in the upper the single mode branch (Port 2) is coupled into the single mode output channel.

c) A symmetric Y-coupler in which incident light in the single mode branches (Ports 2 and 3) is coupled into the single mode output channel (Port 1) with 75% efficiency.

Problem 2.3 – Reflections from an interface

Consider a plane wave incident on an air-dielectric interface at normal incidence. Assume that there are no losses at the interface, i.e. that all optical power is either transmitted on reflected at the interface.

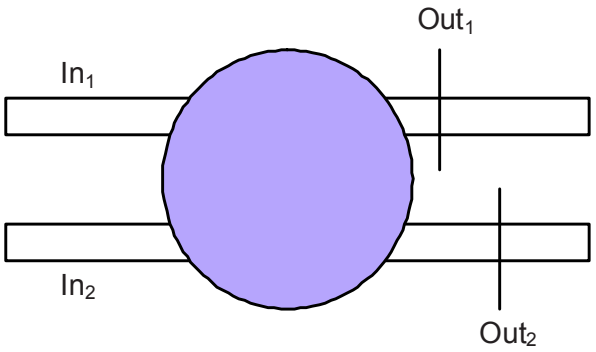
a) Explain how energy conservation demands that the reflections are the same for plane waves incident from the air side and from the dielectric side of the interface.

Consider again plane waves incident on an air-dielectric interface, but now at normal incidence.

b) Is it possible to create an interface that will give 100% reflection for plane waves incident from the air side and 100% transmission for plane waves of the same polarization incident from the dielectric side of the interface? (Explain your answer.)

Problem 2.4 – Waveguide coupling

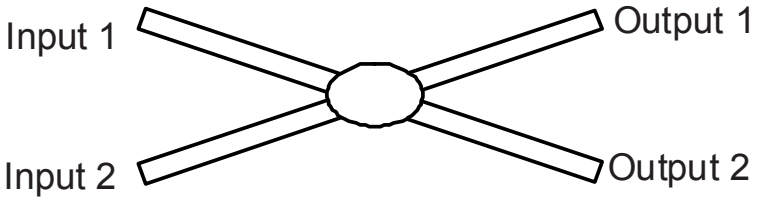
Consider the loss-less, single-mode waveguide device below. With only In_1 present, the outputs are E_{out1} and $E_{out2}=K_{12}E_{out1}$, where K_{12} a complex number. Similarly, with only In_2 present, the outputs are E_{out2} and $E_{out1}=K_{21}E_{out2}$. Express K_{21} in terms of K_{12}.



Problem 2.5 – Crossing waveguides

Consider the symmetric single-mode waveguide crossing in the figure below. The crossing is well designed, so that all the power of the inputs stays in one or the other of the outputs. If there is only power in *Input 1*, then 75 % goes to *Output 1* and 25 % to *Output 2*, and by symmetry, if there is only power in *Input 2*, then 25 % goes to *Output 1* and 75 % to *Output 2*.

What range of values can the ratio of power in Output 1 to power in Output 2 take when the power in the two inputs are the same? The light is at the same frequency, but not necessarily in phase.



Problem 2.6 – Loss in Array Waveguide Gratings

What is the average loss (over the spectrum) of a single-input, single-output Array Waveguide Grating (Fig. 2.15).

3: Plane Waves at Interfaces

3.1 Introduction to Plane Waves

This chapter is focused on the reflection and transmission of plane waves at interfaces between two different optical materials. In Chapter 2 we used energy-conservation to look at a particular part of this problem; the phase relationship between waves reflected from opposite sides of a dielectric interface. Here we will study waves at interfaces in more detail.

The first thing we will find is that consideration of continuity of planes waves at an interface allows us to derive the law of reflection and Snell's law of refraction! These two laws form the foundation for all of Geometrical Optics; a tremendously successful model that even today are used in analysis of the most complex imaging and lens design applications.

Analysis of planes waves at interfaces also gives us the Fresnel-reflection formulae. These expressions for reflection and transmission are very useful in their own right, and they lead to the important concepts of Total-Internal-Reflection that is the basis of traditional wave guiding, and to evanescent fields that are of particular importance in miniaturized and integrated optics.

In the last part of the chapter, we will extend the treatment to multiple interfaces so that we can calculate reflection and transmission through multilayered structures. This will give us the tools to analyze important optical components like Anti-Reflection (AR) coatings, and Bragg reflectors, as well as leaky waveguides and surface plasmons.

3.2 Plane Waves at a Dielectric Interface - Fresnel Reflections

3.2.1 Laws of Reflection and Refraction (Geometrical Optics)

The plane wave solutions combined with the boundary conditions for electromagnetic fields that we derived in Chapter 2 allow us to find formulas for reflections and transmissions of plane waves at a dielectric interface. Consider a monochromatic plane wave of the form

$$\vec{E}_i = \vec{E}_{0i} \cdot \exp\left[j\left(\omega_i t - \vec{k}_i \cdot \vec{r}\right)\right] \tag{3.1}$$

or

$$\vec{E}_i = \vec{E}_{0i} \cdot \cos\left(\omega_i t - \vec{k}_i \cdot \vec{r}\right) \tag{3.2}$$

incident on a dielectric interface described by the surface normal $\vec{S}$. The corresponding reflected and transmitted fields are

$$\begin{aligned} \vec{E}_r &= \vec{E}_{0r} \cdot \cos\left(\omega_r t - \vec{k}_r \cdot \vec{r} + \alpha_r\right) \\ \vec{E}_t &= \vec{E}_{0t} \cdot \cos\left(\omega_t t - \vec{k}_t \cdot \vec{r} + \alpha_t\right) \end{aligned} \tag{3.3}$$

where α_r and α_t are phase constants that allow for the fact that there in general is a phase shift associated with reflection and transmission.

Continuity of the tangential component of the electric field mandates

$$\begin{aligned} &\vec{S} \times \vec{E}_i + \vec{S} \times \vec{E}_r = \vec{S} \times \vec{E}_t \Rightarrow \\ &\vec{S} \times E_{0i} \cdot \cos\left(\omega_i t - \vec{k}_i \cdot \vec{r}\right) + \vec{S} \times E_{0r} \cdot \cos\left(\omega_r t - \vec{k}_r \cdot \vec{r} + \alpha_r\right) \\ &= \vec{S} \times E_{0t} \cdot \cos\left(\omega_t t - \vec{k}_t \cdot \vec{r} + \alpha_t\right) \end{aligned} \tag{3.4}$$

This equation is valid for all times at all points on the dielectric interface, so the arguments of the harmonic function must all be equal:

$$\begin{aligned} &\omega_t t - \vec{k}_i \cdot \vec{r}\Big|_{interface} = \omega_r t - \vec{k}_r \cdot \vec{r} + \alpha_r\Big|_{interface} \\ &= \omega_t t - \vec{k}_t \cdot \vec{r} + \alpha_t\Big|_{interface} \end{aligned} \tag{3.5}$$

The three waves are undergoing forced vibrations at the frequency of the incident wave, so $\omega_i = \omega_r = \omega_t$. We then have at the interface

$$\vec{k}_i \cdot \vec{r} = \vec{k}_r \cdot \vec{r} + \alpha_r = \vec{k}_t \cdot \vec{r} + \alpha_t \tag{3.6}$$

Consider now the first of these two equations.

$$\vec{k}_i \cdot \vec{r} = \vec{k}_r \cdot \vec{r} + \alpha_r \Rightarrow \left(\vec{k}_i - \vec{k}_r\right) \cdot \vec{r} = \alpha_r \tag{3.7}$$

The equality is valid when the position vector, $\vec{r}$, is in a plane parallel to the interface (if α_r is zero, then $\vec{r}$ spans the interface itself). The equation can therefore only hold if the difference of the *k*-vectors is perpendicular to the plane. We conclude that $\vec{k}_r$ is in the incident plane (the plane defined by the incident wave vector and the surface normal). Considering the fact that the magnitude of the *k*-vectors is the same (because they are in the same medium and have the same frequency), we can further conclude that

$$k_i \cdot \sin\theta_i = k_r \cdot \sin\theta_r \Rightarrow \sin\theta_i = \sin\theta_r \tag{3.8}$$

This is the law of reflection, which was well-known long before Maxwell wrote his equations and made possible this derivation.

Similarly we have

$$\vec{k}_i \cdot \vec{r} = \vec{k}_t \cdot \vec{r} + \alpha_t \Rightarrow \left(\vec{k}_i - \vec{k}_t\right) \cdot \vec{r} = \alpha_t \tag{3.9}$$

We see that $\vec{k}_t$ is also in the incident plane, and

$$k_i \cdot \sin\theta_i = k_t \cdot \sin\theta_t \Rightarrow n_i \cdot \sin\theta_i = n_t \cdot \sin\theta_r \tag{3.10}$$

This is Snell's law of refraction!

The laws of reflection and refraction form the basis of geometrical optics, also called ray optics. This very successful model relies on the concept of an optical ray; an optical beam of ideally zero cross sectional area that propagates, reflects, and refracts like a plane wave. We know that energy conservation do not allow the existence of collimated beams of finite, and much less zero, cross sections. The major postulate of geometrical optics is therefore in violation of fundamental physics.

Nevertheless, we will find that many practical results of geometrical optics are consistent with wave-optics. This is particularly true in situation where the aperture of the optics is very large compared to the wavelength of light. Even for miniaturized optics, which is the main focus of this book, geometrical optics is often a useful tool due to its simplicity.

We will continue to point out similarities and differences between wave optics and geometrical optics, and use geometric optics whenever possible, because it is a great tool for developing intuition. It just has to be applied correctly. Some of the

concepts and techniques of Geometrical Optics that we will use in this book are summarized in Appendix 1.

3.2.2 Fresnel Equations

Now we will extend our use of the boundary conditions of the optical fields to not only find the directions of the reflected and transmitted waves, but also their magnitudes. Consider a linearly polarized, monochromatic plane wave incident on an interface as shown in Fig. 3.1. First we study the situation where the E-field is perpendicular to the plane of incidence, which is defined by the wave vector and the surface normal. This transversal-electric field (TE) incident plane wave is also called an s-polarized wave. We will use both expressions interchangeably.

Continuity of the tangential component of the electric field can be expressed

$$\vec{E}_{0i} + \vec{E}_{0r} = \vec{E}_{0t} \tag{3.11}$$

We must also have continuity of the tangential component of the magnetic field at the interface, so

$$-H_{0i}\cos\theta_i + H_{0r}\cos\theta_r = -H_{0t}\cos\theta_t \tag{3.12}$$

Here we have explicitly taken into account the fact that if the *E*-fields of the incident and reflected fields are in phase at the interface, then their *H*-fields are necessarily exactly in opposite phase. This follows from the fact that the energy flows, which are in opposite directions for the incident and reflected waves, are given by the vector product of the *E* and *H*-field.

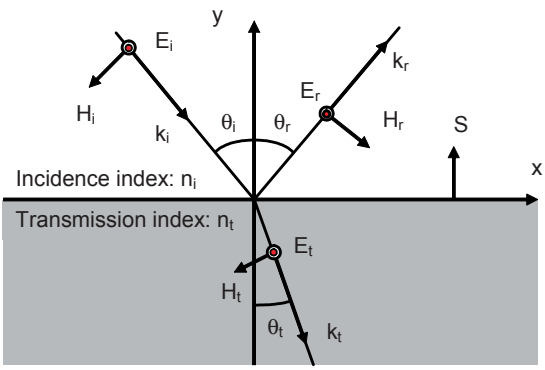


Figure 3.1. Schematic of the incident, reflected, and transmitted components of a Transversal-Electric field (TE) plane wave incident on a dielectric interface.

Into this equation, we substitute the identity

$$H_x = \frac{E_x}{\eta_x} = \frac{n_x\sqrt{\varepsilon_0}}{\sqrt{\mu_x}} E_x \tag{3.13}$$

and we use the law of reflection to find

$$\frac{n_i\sqrt{\varepsilon_0}}{\sqrt{\mu_i}}\left(E_{0i} - E_{0r}\right)\cos\theta_i = \frac{n_t\sqrt{\varepsilon_0}}{\sqrt{\mu_t}} E_{0t}\cos\theta_t \tag{3.14}$$

Combined with the continuity of the electric field at the interface, this results in

$$\frac{n_i\sqrt{\varepsilon_0}}{\sqrt{\mu_i}}\left(2E_{0i} - E_{0t}\right)\cos\theta_i = \frac{n_t\sqrt{\varepsilon_0}}{\sqrt{\mu_t}} E_{0t}\cos\theta_t \Rightarrow$$

$$\left.\frac{E_{0t}}{E_{0i}}\right|_{TE} = \frac{2\frac{n_i}{\sqrt{\mu_i}}\cos\theta_i}{\frac{n_i}{\sqrt{\mu_i}}\cos\theta_i + \frac{n_t}{\sqrt{\mu_t}}\cos\theta_t} \tag{3.15}$$

and

$$\left.\frac{E_{0r}}{E_{0i}}\right|_{TE} = \frac{\frac{n_i}{\sqrt{\mu_i}}\cos\theta_i - \frac{n_t}{\sqrt{\mu_t}}\cos\theta_t}{\frac{n_i}{\sqrt{\mu_i}}\cos\theta_i + \frac{n_t}{\sqrt{\mu_t}}\cos\theta_t} \tag{3.16}$$

In practice we are most often using materials in which $\mu_x \approx \mu_0$, so the Fresnel Equations for *s*-polarized waves simplify to

$$t_{TE} = \left.\frac{E_{0t}}{E_{0i}}\right|_{TE} = \frac{2n_i\cos\theta_i}{n_i\cos\theta_i + n_t\cos\theta_t} \tag{3.17}$$

and

$$r_{TE} = \left.\frac{E_{0r}}{E_{0i}}\right|_{TE} = \frac{n_i\cos\theta_i - n_t\cos\theta_t}{n_i\cos\theta_i + n_t\cos\theta_t} \tag{3.18}$$

The derivations of the Fresnel Equations for waves with Tranversal-Magnetic (TM), or *p*-polarized fields, (*E*-field polarized in the plane of incidence) are very similar. For a *p*-polarized, monochromatic plane wave incident on an interface as shown in Fig. 3.1, continuity of the tangential *E*-field requires

$$E_{0i} \cdot \cos\theta_i + E_{0r} \cos\theta_r = E_{0t} \cos\theta_t \Rightarrow \left(E_{0i} + E_{0r}\right) \cdot \cos\theta_i = \vec{E}_{0t} \cos\theta_t \tag{3.19}$$

Similarly, continuity of the magnetic field at the interface allow us to write

$$H_{0i} - H_{0r} = H_{0t} \tag{3.20}$$

The reason for the minus sign in the equation is that we are again explicitly taking into account the fact that the reflected wave, which is traveling away from the interface, must have its *H*-field reversed. Note that this choice is different from what is done in many text books, which write this equation as $H_i + H_r = H_t$. Both methods are equally valid, provided that it is applied consistently and that the results of the calculations are correctly interpreted. The advantage of our approach is that it implies a definition of phase shift that yields the same results for *p*-polarized and *s*-polarized waves at normal incidence, as our intuition tells us that it should be.

Again we use the relationship $H_x = \dfrac{n_x\sqrt{\varepsilon_0}}{\sqrt{\mu_x}} E_x$ to write =>

$$\frac{n_i}{\sqrt{\mu_i}} E_{0i} - \frac{n_i}{\sqrt{\mu_i}} E_{0r} = \frac{n_t}{\sqrt{\mu_t}} E_{0t}$$

Substituting into the equation of continuity of the electric field at the interface we find

$$\frac{n_t}{\sqrt{\mu_t}}\left(E_{0i} + E_{0r}\right) \cdot \cos\theta_i = \frac{n_i}{\sqrt{\mu_i}}\left(E_{0i} - E_{0r}\right) \cdot \cos\theta_t \Rightarrow \tag{3.21}$$

$$\left.\frac{E_{0r}}{E_{0i}}\right|_{TM} = \frac{\dfrac{n_i}{\sqrt{\mu_i}}\cos\theta_t - \dfrac{n_t}{\sqrt{\mu_t}}\cos\theta_i}{\dfrac{n_t}{\sqrt{\mu_t}}\cos\theta_i + \dfrac{n_i}{\sqrt{\mu_i}}\cos\theta_t} \approx \frac{n_i\cos\theta_t - n_t\cos\theta_i}{n_t\cos\theta_i + n_i\cos\theta_t} \tag{3.22}$$

Similarly we find the following expression for the transmitted field

$$\left.\frac{E_{0t}}{E_{0i}}\right|_{TM} = \frac{\dfrac{n_i}{\sqrt{\mu_i}} 2\cos\theta_i}{\dfrac{n_t}{\sqrt{\mu_t}}\cos\theta_i + \dfrac{n_i}{\sqrt{\mu_i}}\cos\theta_t} \approx \frac{2n_i\cos\theta_i}{n_t\cos\theta_i + n_i\cos\theta_t} \tag{3.23}$$

In summary, we have found the following complete set of Fresnel Equations for the reflection and transmission of plane waves through a dielectric interface:

$$\left.\frac{E_{0t}}{E_{0i}}\right|_{TE} = t_{TE} = \frac{2n_i \cos\theta_i}{n_i \cos\theta_i + n_t \cos\theta_t} \tag{3.24}$$

$$\left.\frac{E_{0r}}{E_{0i}}\right|_{TE} = r_{TE} = \frac{n_i \cos\theta_i - n_t \cos\theta_t}{n_i \cos\theta_i + n_t \cos\theta_t} \tag{3.25}$$

$$\left.\frac{E_{0t}}{E_{0i}}\right|_{TM} = t_{TM} = \frac{2n_i \cos\theta_i}{n_t \cos\theta_i + n_i \cos\theta_t} \tag{3.26}$$

$$\left.\frac{\vec{E}_{0r}}{E_{0i}}\right|_{TM} = r_{TM} = \frac{n_i \cos\theta_t - n_t \cos\theta_i}{n_t \cos\theta_i + n_i \cos\theta_t} \tag{3.27}$$

Often these Fresnel equations are presented in a slightly different form, in which the last equation is the negative of what we have derived here. This inversion is a consequence of the choice we made to explicitly take into account the phase reversal of the reflected magnetic field.

The advantage of our notation is that the reflection at normal incidence is the same for *s* and *p*-polarized waves. This can be verified by observing that if the incident angle is zero, then

$$r_{TM} = r_{TE} = \frac{n_i - n_t}{n_t + n_i} \tag{3.28}$$

This simple formula for reflections at normal incidence is useful to memorize, but keep in mind that this is the *field* reflections. The power or intensity reflections are the square of the field reflections as we will see.

3.2.3 Numerical Evaluation of the Fresnel Equations

To get a better understanding of the physics of planes waves at interfaces, we numerically evaluate the Fresnel Equations at an interface between two media of index 1.0 (air) and 1.5 (glass) respectively. Figure 3.2 shows graphs of reflection and transmission at the interface when the plane wave is incident on the low-index side of the interface. This situation of light impinging from the low-index side is called external reflection.

The results shown in Fig. 3.2 are not very surprising. The magnitudes of the reflection and transmission coefficients are all between zero and unity for the whole

range of input angles. The phase of transmission is zero, while the reflection for transversally polarized E-fields is π radians for all angles.

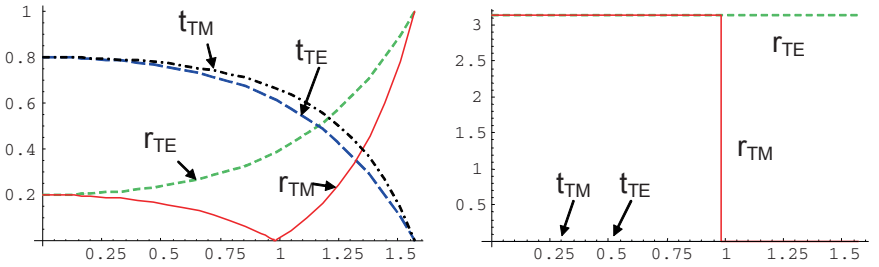


Figure 3.2. *Absolute values (left) and phase (right) of reflection and transmission when $n_i<n_t$ (external reflection) as a function of incident angle (in radians). The phases of the transmitted fields are zero for all angles. Notice the null in r_{TM} at the Brewster angle.*

The reflection for transversally polarized magnetic fields is more interesting with a null and a phase shift at the Brewster angle. This is a phenomenon that is useful in many optical devices, and we will look closer at it in Section 3.2.4. As pointed out above, with our definitions, we find that the magnitude and phase of the reflections are the same for TE and TM at normal incidence, as we would expect.

Figure 3.3 shows the numerical evaluation of reflection and transmission when the plane wave is incident on the high-index side of the interface. This situation is often referred to as internal reflection. The graphs show that beyond a critical angle, the reflections for both TE and TM waves go to unity. This phenomenon is called Total-Internal-Reflection (TIR). It is a consequence of the fact that beyond the critical angle, there simply are no plane wave solutions on the low index-side that can match the periodicity of the incident field on the high-index side. There are therefore no plane waves that can transition and travel away from the interface.

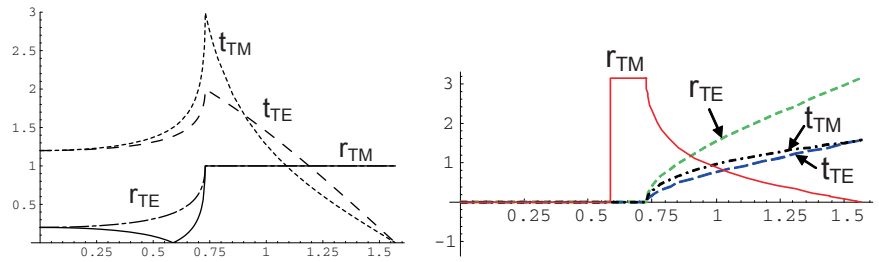


Figure 3.3. *Absolute values (left) and phase (right) of the reflection and transmission when $n_i>n_t$ (internal reflection). Again we notice the null in r_{TM} at the Brewster angle. Beyond the Total Internal Reflection (TIR) angle the transmitted energy is zero.*

Figure 3.3 shows that the transmission of the electric field is larger than unity for a wide range of input angles. For relatively small angles (below the critical TIR angle), this is simply due to the fact that matching the energy flows into and away from the interface requires higher fields on the low-index side.

At incident angles beyond TIR we see that we have transmitted fields even though the reflections go to unity! These transmitted fields are evanescent fields that are not carrying energy away from the interface. This is our first indication of non-traveling fields that we will study in more detail later.

3.2.3 Reflectance and Transmittance

To check that energy conservation is indeed satisfied by the Fresnel formulae we just derived, we'll calculate the energies of the reflected and transmitted fields. The time-averaged radiant flux, or intensity, (W/m^2) of the optical plane wave is given by

$$I = \frac{1}{2}\sqrt{\frac{\varepsilon}{\mu}} \cdot E_0^2 \cdot \frac{\vec{k}}{k} = \frac{n}{2}\sqrt{\frac{\varepsilon_0}{\mu}} \cdot E_0^2 \cdot \frac{\vec{k}}{k} \tag{3.29}$$

We define the reflectance as the ratio of the intensity reflected from an area, A, to the intensity incident on the same area

$$R = \frac{AI_r \cos\theta_r}{AI_i \cos\theta_i} = \frac{E_r^2}{E_i^2} = r^2 \tag{3.30}$$

Similarly we define the transmittance as the ratio of the intensity transmitted through an area, A, to the intensity incident on the same area

$$T = \frac{AI_t \cos\theta_t}{AI_i \cos\theta_i} = \frac{n_t \cos\theta_t}{n_i \cos\theta_i}\frac{E_t^2}{E_i^2} = \frac{n_t \cos\theta_t}{n_i \cos\theta_i} t^2 \tag{3.31}$$

These expressions for the reflectance and transmittance are plotted in Fig. 3.4. As before we are considering an air ($n_i=1$) to glass ($n_t=1.5$) interface. The graphs show that both the reflectance and transmittance are between 0 and 1 for all incident angles. By combining Eqs. 3.30 and 3.31 with the Fresnel formulae, we find that R+T=1, as we intuitively know that must be the case.

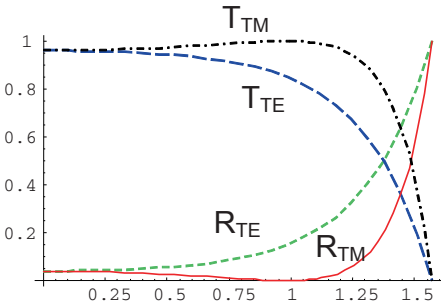


Figure 3.4. *Reflectance and transmittance from an interface between air (n_i=1) and glass (n_t =1.5). As expected for external reflection, the transmittance is finite for all values of the incident angle between 0 and $\pi/2$.*

3.2.4 Brewster Angle

Figures 3.2, 3.3, and 3.4 show that at a specific angle, called the Brewster angle, there is a null in the TM wave reflection, while TE reflections are finite. This is an interesting phenomenon that is useful for many optical devices, e.g. windows on gas-laser tubes that ideally should be completely transparent for one polarization, while introducing losses for the other.

By inspection of the formula for r_{TM}, we find that the null occurs at the angle given by

$$n_i \cos\theta_{tB} = n_t \cos\theta_{iB} \tag{3.32}$$

Combining this with Snell's law

$$n_i \sin\theta_i = n_t \sin\theta_t \tag{3.33}$$

we find

$$\begin{aligned} &\sin\theta_{tB} \cdot \cos\theta_{tB} = \sin\theta_{iB} \cdot \cos\theta_{iB} \Rightarrow \\ &\sin(2\theta_{tB}) = \sin(2\theta_{iB}) \end{aligned} \tag{3.34}$$

This equation has the trivial solution θ_{iB}=θ_{tB} (which means that n_i=n_t, i.e. there is no interface), but also the more interesting solution

$$\theta_{tB} + \theta_{iB} = \frac{\pi}{2} \tag{3.35}$$

Substituting back into the first equation we find the following expression for the Brewster angle

$$\tan\theta_{iB} = \frac{n_t}{n_i} \tag{3.36}$$

We see that at the Brewster angle, the transmitted and reflected waves are perpendicular.

It is interesting to note that this fact (that when the transmitted and reflected wave are perpendicular, the reflection goes to zero), which we found by considering the continuity of the waves at the interface, is exactly what we expect if we consider the dipoles that drive the transmitted and reflected fields. Because the reflected wave is perpendicular to the dipoles that set up the transmitted field, these dipoles cannot deliver power in the direction of the reflected wave, so the reflection goes to zero. This is illustrated in the figure below.

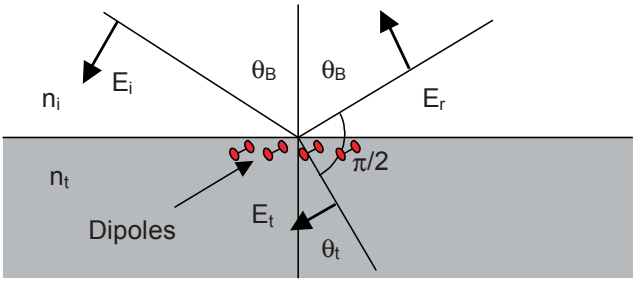


Figure 3.5. *A TM plane wave incident on a dielectric interface drives dipoles in the dielectric that set up the transmitted field. If the reflected wave is parallel to the dipole axis (perpendicular to the transmitted field), the reflection disappears. This condition is only possible for TM polarized light. TE polarized light sets up dipoles perpendicular to the plane of incidence, so the dipole axis cannot coincide with the reflected wave vector.*

3.3 Wave description of Total Internal Reflection (TIR)

As pointed out in the previous discussion, we see from Fig. 3.3 that when plane waves are incident on an interface between one material and another with a lower index of refraction (internal reflection), the reflection becomes unity for all angles beyond a critical value, θ_{cr}. The value of the critical angle can be found by setting the transmitted angle equal to $\pi/2$ (the maximum value it can have and still be a transmitted wave) in Snell's law. The results is

$$n_i \cdot \sin\theta_{cr} = n_t \cdot \sin\theta_t\big|_{\theta_t=\pi/2} \Rightarrow \theta_{cr} = \sin^{-1}\left(\frac{n_t}{n_i}\right) \tag{3.37}$$

where $n_t<n_i$.

Beyond the critical angle the propagation vector of the transmitted wave is imaginary in the low-index material. This exponentially decaying wave, which is called an evanescent field, does not carry power into the low-index material (the electric and magnetic fields are 90 degrees out of phase so the time averaged Poynting vector is zero). The evanescent field does, however, play a very significant role in the operation of directional couplers, ring filters, surface-plasmon sensors, and some types of fiber optic switches.

3.3.1 Evanescent Fields

To develop a quantitative understanding of Total Internal Reflection, including the evanescent field, consider a dielectric interface with incident and transmitted TE fields of the form

$$\begin{aligned} E_i(x,y) &= \vec{z}\cdot E_{0i}\cdot\exp\left[-jk_0 n_i\left(x\cdot\sin\theta_i - y\cdot\cos\theta_i\right)\right] \\ E_t(x,y) &= \vec{z}\cdot t_{TE}\cdot E_{0i}\cdot\exp\left[-jk_0 n_t\left(x\cdot\sin\theta_t - y\cdot\cos\theta_t\right)\right] \end{aligned} \tag{3.38}$$

Here we have dropped the time dependence for convenience. The incident and transmitted angles are related by

$$\sin\theta_t = \frac{n_i}{n_t}\sin\theta_i \Rightarrow \cos\theta_t = \sqrt{1-\left(\frac{n_i}{n_t}\sin\theta_i\right)^2} \tag{3.39}$$

Notice that we use the positive square root. There is no ambiguity here, because if we took the negative root we would have a wave propagating in a different direction (with a positive y-component of the propagation vector). Using this expression, the transmitted field is

$$\begin{aligned} &E_t(x,y) = \vec{z}\cdot t_{TE}\cdot E_{0i}\cdot \\ &\exp\left[-jk_0 n_t\left(x\cdot\frac{n_i}{n_t}\cdot\sin\theta_i - y\cdot\sqrt{1-\frac{n_i^2}{n_t^2}\sin^2\theta_i}\right)\right] \end{aligned} \tag{340}$$

At the critical angle

$$\sin\theta_{cr} = \frac{n_t}{n_i} \tag{3.41}$$

the transmitted angle is 90 degrees and the transmitted field is a plane wave along the interface

$$E_t(x,y) = \vec{z} \cdot t_{TE} \cdot E_{0i} \cdot \exp[-jk_0 n_t \cdot x] \tag{3.42}$$

Beyond the critical angle our expression for cosine of the transmitted angle becomes imaginary. Now there is an ambiguity in the choice of sign of the root of the expression for cosine of the transmitted angle. To see the implications of this choice, we use the general expression

$$\sin\theta_i > \frac{n_t}{n_i} \Rightarrow \cos\theta_t = \pm j \cdot \sqrt{\frac{n_i^2}{n_t^2}\sin^2\theta_i - 1} \tag{3.43}$$

The transmitted field is then

$$\begin{aligned} &E_t(x,y) = \vec{z} \cdot t_{TE} \cdot E_{0i} \cdot \\ &\exp\left[-jk_0 n_t \left(x \cdot \frac{n_i}{n_t} \cdot \sin\theta_i - y \cdot \left\{ \pm j \sqrt{\frac{n_i^2}{n_t^2}\sin^2\theta_i - 1} \right\} \right)\right] \\ &\Rightarrow E_t(x,y) = \vec{z} \cdot t_{TE} \cdot E_{0i} \cdot \\ &\exp\left[-jk_0 n_i x \cdot \sin\theta_i - k_0 n_t y \cdot \left\{ \pm \sqrt{\frac{n_i^2}{n_t^2}\sin^2\theta_i - 1} \right\}\right] \end{aligned} \tag{3.44}$$

To obtain a solution that is decaying exponentially in the negative y-direction, we must choose the negative root in this expression.

Keeping in mind the choice of the negative root that we just made, let's consider the phase shift of the reflected light under total internal reflection. Substituting the above expressions for the sine and cosine of the transmitted angle we find

$$\left.\frac{E_{0r}}{E_{0i}}\right|_{TE} = \frac{n_i \cos\theta_i + j \cdot n_t \sqrt{\frac{n_i^2}{n_t^2}\sin^2\theta_i - 1}}{n_i \cos\theta_i - j \cdot n_t \sqrt{\frac{n_i^2}{n_t^2}\sin^2\theta_i - 1}} \tag{3.45}$$

The phase shift is

$$\Phi_{TE} = 2 \cdot \tan^{-1} \frac{\sqrt{n_i^2 \sin^2\theta_i - n_t^2}}{n_i \cos\theta_i} \tag{3.46}$$

We see that the phase shift goes from zero at the critical angle to $+\pi$ at gracing incidence, as shown in the graph in Fig. 3.3. If we plot the phase based on the Fresnel Equations and simply choose the positive root of the equation for cosine of the

transmitted angle without concern for the unphysical nature of an exponentially increasing wave, then we get the exact opposite, i.e. the calculated phase shift then goes from zero at the critical angle to $-\pi$ at gracing incidence. This illustrates that we must be careful when applying the Fresnel equations to Total Internal Reflection!

3.3.2 Goos-Hänchen Shift

The phase shift associated with TIR appears to correspond to a an offset between the incident and reflecting planes as can be seen by considering the equation we used to derive the law of reflections

$$\vec{k}_i \cdot \vec{r}_i = \vec{k}_r \cdot \vec{r}_r + \alpha_r \Rightarrow \left(\vec{k}_i - \vec{k}_r\right) \cdot \vec{r}_i = \left(\vec{r}_r - \vec{r}_i\right) \cdot \vec{k}_r - \alpha_r \tag{3.47}$$

The offset, that is required to explain the phase shift of TIR, implies a shift of the beam along the interface as shown in Fig. 3.6. A beam undergoing TIR therefore appears to be reflected not from the interface between the low and high index material, but from within the high index material. This shift has practical consequences in some integrated optics devices.

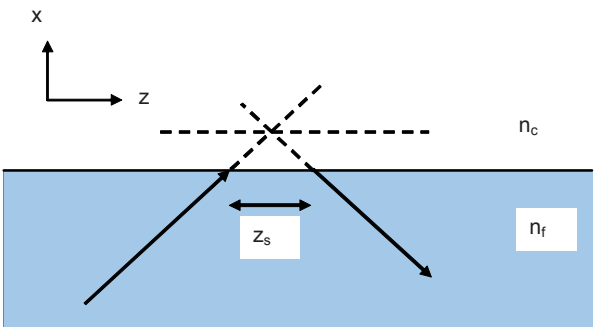


Figure 3.6. The phase shift associated with TIR leads to a lateral shift of a reflected beam of finite cross section. This shift is called the Goos-Hänchen shift.

3.3.3 Optical Devices Based on TIR

Total Internal Reflection (TIR) is the basis of a variety of optical devices. Many types of prisms that perform special function are based on TIR, and we will later use TIR for a first order explanation of optical waveguides. The evanescent fields that are created as a consequence of TIR are used to advantage in numerous devices, e.g. TIR spectrometers that allows spectra to be obtained in very small volumes.

As an example of an optical MEMS device that depends on TIR, let's briefly consider the Champagne fiber switch developed by Agilent. The basic principle is shown in Fig. 3.7. The object is to switch light in one input waveguide between two output waveguides.

The light is typically not guided in the switching region (or at least the guides are multimode in this region). In the off state, the light passes right through the waveguide crossing with little loss. The light path has two discontinuities in this state; the light passes from the silica waveguide into an index match liquid and back into the silica.

To switch the light the liquid in the switching cell is evaporated. This changes the refractive index in the cell from about 1.5 (matched to the glass) to close to unity, i.e. a very large index shift. By comparison, the index shift in most electrooptic materials never exceeds 10^{-4}.

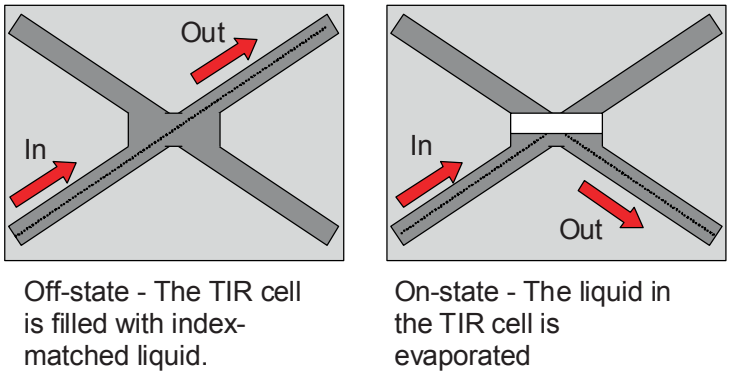


Figure 3.7. *Fiber optic switch based on TIR. In the off-state the TIR cell is filled with index-matched liquid, and the light is transmitted through the interface with very little reflection. In the on-state, the liquid in the TIR cell is evaporated by resistive heaters. This leads to TIR at the glass-gas interface, and virtually all the light is reflected.*

Notice the role of the Goos-Hänchen shift in the design of the champagne switch. The TIR interface is not placed at the intersection of the optical axes of the crossing waveguides, to account for the lateral shift caused by TIR.

The most significant advantage of the Champagne switch is its simplicity and compactness, which allow relatively large numbers (~ 1000) of switches to be integrated into moderately large switching matrices. The drawbacks are speed, and the fact that each individual switching element is only capable of 1 by 2 switching, which means that N^2 elements are needed for a N by N switch.

3.4 Multilayer Stacks

Now we will expand on our plane-waves-at-an-interface model to find the reflection and transmission of layered materials with arbitrary layer thicknesses and parallel interfaces. We will use an approach similar to the one we used to derive the Fresnel equations. Our starting point is the reflections from a single dielectric film, i.e. two interfaces with a constant separation. The incident, transmitted and reflected TM fields for a dielectric film are shown in Fig. 3.8.

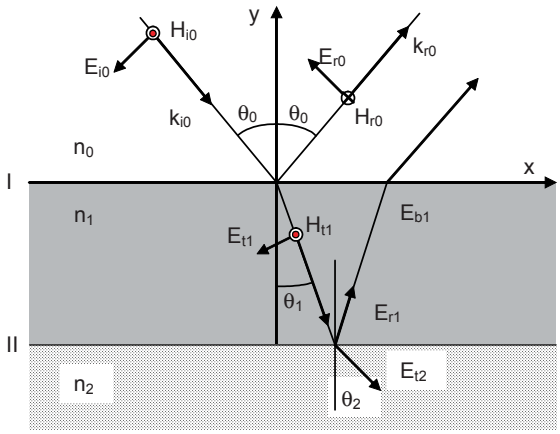


Figure 3.8. Schematic of the incident, reflected, transmitted and back reflected components of a Transversal-Electric field (TE) plane wave incident on a dielectric film.

The boundary conditions at interface I (0-1) are

$$E_I = E_{i0}\cos\theta_0 + E_{r0}\cos\theta_0 = E_{t1}\cos\theta_1 + E_{b1}\cos\theta_1 \tag{3.48}$$

$$E_I = \left(E_{i0} + E_{r0}\right)\cos\theta_0 = \left(E_{t1} + E_{b1}\right)\cos\theta_1 \tag{3.49}$$

and

$$H_I = H_{i0} - H_{r0} = H_{t1} - H_{b1} \tag{3.50}$$

We will continue to restrict our investigations to non-magnetic materials, so the magnetic and electric fields are related by

$$\vec{H} = \sqrt{\frac{\varepsilon_0}{\mu_0}} \cdot n \cdot \frac{\vec{k}}{|k|} \times \vec{E} \tag{3.51}$$

We can rewrite the second boundary condition as

$$H_I = \sqrt{\frac{\varepsilon_0}{\mu_0}} \cdot n_0 \cdot (E_{i0} - E_{r0}) = \sqrt{\frac{\varepsilon_0}{\mu_0}} \cdot n_1 \cdot (E_{t1} - E_{b1}) \tag{3.52}$$

Similarly we have at the second boundary (1-2):

$$E_{II} = (E_{i1} + E_{r1})\cos\theta_1 = E_{t2}\cos\theta_2 \tag{3.53}$$

$$H_{II} = H_{i1} - H_{r1} = H_{t2} \Rightarrow$$
$$H_{II} = \sqrt{\frac{\varepsilon_0}{\mu_0}} \cdot n_1 \cdot (E_{i1} - E_{r1}) = \sqrt{\frac{\varepsilon_0}{\mu_0}} \cdot n_2 \cdot E_{t2} \tag{3.54}$$

We now need to find the phase shift of a beam crossing a dielectric film of a given thickness, d, at a given angle, θ. From Fig. 3.9 we see that the phase shift is

$$\frac{2\pi}{\lambda} d \cdot \cos\theta = \frac{2\pi}{\lambda_0} nd \cdot \cos\theta = k_0 nd \cdot \cos\theta \tag{3.55}$$

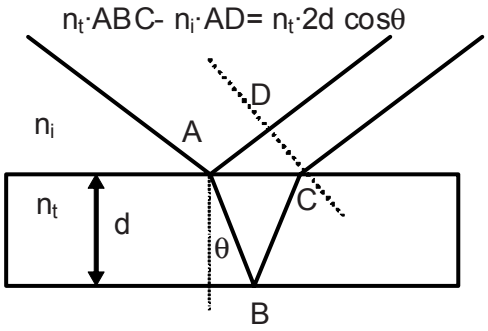


Figure 3.9. Geometry of plane wave reflected from a thin film. The phase shift of the light reflected from the back of the film is equal to $2d\cos\theta$.

Phase delay through tilted etalon

The result of the calculation in Fig. 3.9 is somewhat counterintuitive. One expects that rotating the plate so that the beam goes a longer way would lead to a larger phase delay, i.e. $\frac{2\pi}{\lambda} \cdot \frac{d}{\cos\theta}$, but instead we find the exact opposite ($\frac{2\pi}{\lambda} \cdot d\cos\theta$) when we also carefully consider the lateral shift of the beam. This counter intuitive result has led to many embarrassing errors in optical design and analysis. Make sure you don't become the next victim!

We then have the following relations between the waves on either side of the dielectric film

$$E_{b1} = E_{r1} \cdot e^{-jk_0 \cdot nd\cos\theta_1} \tag{3.56}$$

and

$$E_{i1} = E_{t1} \cdot e^{-jk_0 \cdot nd\cos\theta_1} \tag{3.57}$$

Using these expressions we can write the fields at the interfaces in the following way

$$E_I = (E_{t1} + E_{b1})\cos\theta_1 \tag{3.58}$$

$$H_I = \sqrt{\frac{\varepsilon_0}{\mu_0}} \cdot n_1 \cdot (E_{t1} - E_{b1}) \tag{3.59}$$

$$E_{II} = (E_{i1} + E_{r1})\cos\theta_1 = \left(E_{t1} \cdot e^{-jk_0 \cdot nd\cos\theta_1} + E_{b1} \cdot e^{jk_0 \cdot nd\cos\theta_1}\right)\cos\theta_1 \tag{3.60}$$

$$H_{II} = \sqrt{\frac{\varepsilon_0}{\mu_0}} \cdot n_1 \cdot (E_{i1} - E_{r1}) = \sqrt{\frac{\varepsilon_0}{\mu_0}} \cdot n_1 \cdot \left(E_{t1} \cdot e^{-jk_0 \cdot nd\cos\theta_1} - E_{b1} \cdot e^{jk_0 \cdot nd\cos\theta_1}\right) \tag{3.61}$$

We can now solve the last two equations for E_{t1} and E_{b1} and substitute into the two first equations for the fields at interface I to find relationships between the fields at interfaces I and II. In matrix form these relationships can be expressed as

$$\begin{bmatrix} E_I \\ H_I \end{bmatrix} = \begin{bmatrix} \cos(k_0 \cdot nd\cos\theta_1) & j\sin(k_0 \cdot nd\cos\theta_1)/\Gamma_{1TM} \\ j\sin(k_0 \cdot nd\cos\theta_1)\cdot\Gamma_{1TM} & \cos(k_0 \cdot nd\cos\theta_1) \end{bmatrix} = \begin{bmatrix} E_{II} \\ H_{II} \end{bmatrix} \tag{3.62}$$

where $\Gamma_{1TM} = \sqrt{\frac{\varepsilon_0}{\mu_0}} n_1 \Big/ \cos\theta_1$. For TE (s-polarized) waves we get the same result, except we now replace Γ_{1TM} with $\Gamma_{1TE} = \sqrt{\frac{\varepsilon_0}{\mu_0}} n_1 \cdot \cos\theta_1$.

This matrix formalism is conveniently extended to multi-layer systems. We use the label M_1 for the matrix connecting the fields on either side of film 1. We can then relate the fields on either side of a stack of films by simply multiplying the matrices of all the films in the stack. The fields at the front surface of a stack of p layers is then related to the fields at the bottom of film p in the following way

$$\begin{bmatrix} E_I \\ H_I \end{bmatrix} = M_1 M_2 M_3 .....M_p \begin{bmatrix} E_{p+1} \\ H_{p+1} \end{bmatrix} = M \begin{bmatrix} E_{p+1} \\ H_{p+1} \end{bmatrix} \tag{3.63}$$

To see how this simple formalism can be used to find the reflection and transmission from stacks of dielectric films, we rewrite the last equation by using the boundary conditions on the first and last interface. For TM (p-polarized) waves we find

$$\begin{bmatrix} E_{i0} + E_{r0} \\ \Gamma_{0TM} \cdot (E_{i0} - E_{r0}) \end{bmatrix} = M \begin{bmatrix} E_{tp+1} \\ \Gamma_{p+1TM} \cdot E_{tp+1} \end{bmatrix} \tag{3.64}$$

where we as before have used the definition $\Gamma_{kTM} = \sqrt{\frac{\varepsilon_0}{\mu_0}} n_k \Big/ \cos\theta_k$. Solving these two equations in three unknowns we find the reflection and transmission ratios for the electric field

$$\left.\frac{E_{r0}}{E_{i0}}\right|_{TM} = r_{\|} = \frac{\Gamma_{0TM} m_{11} + \Gamma_{0TM}\Gamma_{p+1TM} m_{12} - m_{21} - \Gamma_{p+1TM} m_{22}}{\Gamma_{0TM} m_{11} + \Gamma_{0TM}\Gamma_{p+1TM} m_{12} + m_{21} + \Gamma_{p+1TM} m_{22}} \tag{3.65}$$

$$\left.\frac{E_{tp+1}}{E_{i0}}\right|_{TM} = t_{\|} = \frac{2\Gamma_{TM0}}{\Gamma_{0TM} m_{11} + \Gamma_{0TM}\Gamma_{p+1TM} m_{12} + m_{21} + \Gamma_{p+1TM} m_{22}} \tag{3.66}$$

We can apply the same procedure to TE waves to get similar expressions. We just have to make the substitution

$$\Gamma_{k,TM} = \sqrt{\frac{\varepsilon_0}{\mu_0}} n_k \Big/ \cos\theta_k \rightarrow \Gamma_{k,TE} = \sqrt{\frac{\varepsilon_0}{\mu_0}} n_k \cdot \cos\theta_k \tag{3.67}$$

3.5 Applications of Layered Structures

The formalism we have developed can be used to calculate the reflection and transmission through any layered optical structure with parallel interfaces. These types of stacks of multiple layers, or thin films, of different materials are technologically very important with numerous applications throughout optics. In the following we will show some important examples, and discuss their use in micro- and nano-optics.

3.5.1 Anti-Reflection Coatings

First we consider the reflection from a single film at normal incidence. Just as for Fresnel reflections, the reflections at normal incidence are the same for TE and TM waves, and it can be written

$$r = \frac{n_1(n_0 - n_2)\cdot\cos(k_0\cdot nd) + \left(n_0\cdot n_2 - n_1^2\right)\cdot j\sin(k_0\cdot nd)}{n_1(n_0 + n_2)\cdot\cos(k_0\cdot nd) + \left(n_0\cdot n_2 + n_1^2\right)\cdot j\sin(k_0\cdot nd)} \tag{3.68}$$

To find the reflectance we multiply r with its conjugate to obtain

$$R = rr^* = \frac{n_1^2(n_0 - n_2)^2\cdot\cos^2(k_0\cdot nd) + \left(n_0\cdot n_2 - n_1^2\right)^2\cdot\sin^2(k_0\cdot nd)}{n_1^2(n_0 + n_2)^2\cdot\cos^2(k_0\cdot nd) + \left(n_0\cdot n_2 + n_1^2\right)^2\cdot\sin^2(k_0\cdot nd)} \tag{3.69}$$

For the important case that the film thickness is a quarter of a wavelength, i.e. $d = \frac{\lambda_0}{4\pi\cdot n} \Leftrightarrow k_0\cdot nd = \pi/2$, we find

$$R = \frac{\left(n_0\cdot n_2 - n_1^2\right)^2}{\left(n_0\cdot n_2 + n_1^2\right)^2} \tag{3.70}$$

If the index of the film is the geometrical mean of the index of the incident medium and the substrate, then the reflectance goes to zero! In other words, a film of the correct thickness $\left(d = \frac{\lambda_0}{4\pi\cdot n}\right)$ and correct index $n_1 = \sqrt{n_0\cdot n_2}$ removes reflections between an incidence medium and a substrate. This is called a quarter-wave Anti-Reflection (AR) coating.

Better AR coating consisting of multiple layers can be designed, but the single quarter-wavelength AR coating is still very much used, particularly in applications that cannot use traditional thin-film deposition techniques. For example, multilayer thin film deposition is difficult in typical optical MEMS, because the films are challenging to pattern, and because the stacks typically have more built-in mechanical stress than can be tolerated by the MEMS structures.

A simple, yet effective, technique for avoiding reflections from an air-silicon interface is to apply a single silicon-nitride quarter-wave film. The index of silicon in the optical-communication wavelength bands (1,200 nm to 1,600 nm) is about 3.5, and the index of silicon nitride is about 2.0 in the same wavelength range (it can be adjusted to be even closer to the ideal value $n_{SiN} = \sqrt{1\cdot 3.5} = 1.87$). A single quarter-wave silicon-nitride film will then reduce the reflection at the an air-

silicon interface from $R_{Si} = \frac{(n_{air} - n_{Si})^2}{(n_{air} + n_{Si})^2} = \frac{(1-3.5)^2}{(1+3.5)^2} \approx 0.31$ for bare silicon to as little as $R_{AR} = \frac{\left(n_{air} \cdot n_{Si} - n_{SiN}^2\right)^2}{\left(n_{air} \cdot n_{Si} + n_{SiN}^2\right)^2} = \frac{\left(1 \cdot 3.5 - 2^2\right)^2}{\left(1 \cdot 3.5 + 2^2\right)^2} \approx 0.0044$ with the AR coating .

3.5.2 Bragg reflectors

The power of our formalism is that we can calculate reflections from any combination of parallel interfaces. Let's now use this capability to look at the reflections from a stack of films pairs of alternating high and low index as in Fig. 3.10 that shows a reflector consisting of silicon films separated by air gaps. Each silicon film and each air gap are adjusted to have an optical thickness of one quarter wave, i.e. their physical thickness is one quarter of the vacuum wavelength divided by the index of refraction of the layer. Such a mirror can be thought of as one-dimensional Photonic Crystals and are increasingly used in optical MEMS implementations.

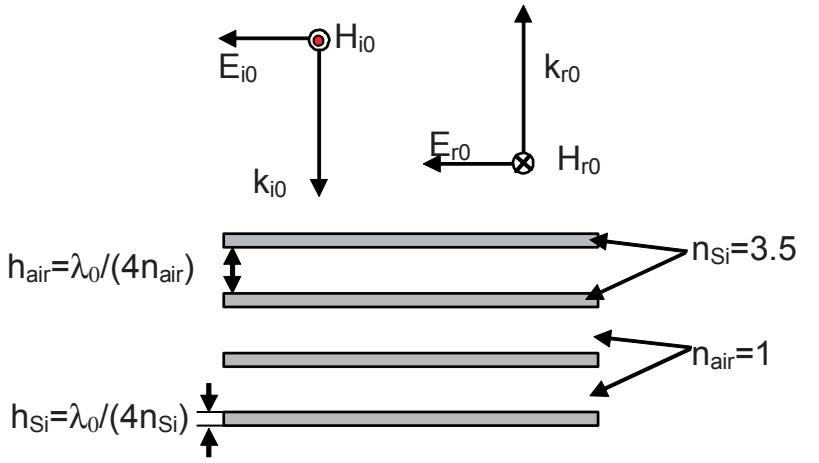


Figure 3.10. Silicon reflector consisting of alternating silicon layers and air gaps. The layers and gaps are each nominally one quarter-wave thick at the center wavelength of interest. .

The reflectance for Si-air multi-layer mirrors are shown in Figure 3.11. We have chosen to plot the reflectance as a function of the wave vector, $k=2\pi/\lambda$, instead of as a function of wavelength, because it results in nice symmetric graphs as evident in Fig. 3.11.

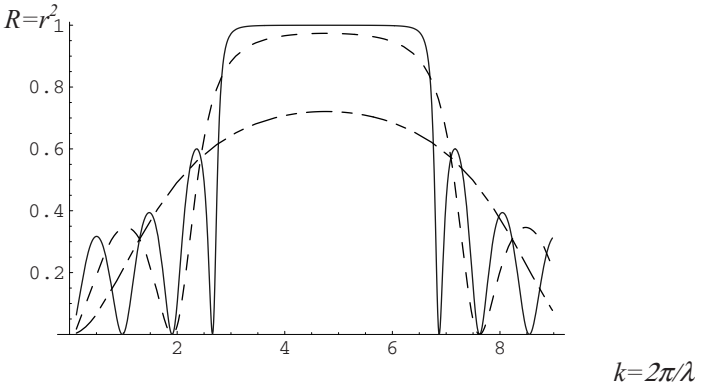


Figure 3.11. Reflectance from a silicon-air Bragg mirror with 1 (dot-dashed), 2 (dashed), and 3 (solid) layer pairs. The maximum reflectance values at k=4.7636=2π/1.319 are 0.7209, 0.9737, and 0.9998 respectively.

Three different mirror configurations are shown; one layer pair, two pairs, and four pairs. The refractive index of Silicon is assumed to be 3.5, and the thickness of each layer is chosen to be a quarter of a wavelength at 1.319 nm wavelength, which is the absorption minimum of Silicon.

Bragg mirrors consisting of only a few layers have quite high reflectivity. The maximum reflectances are larger than 0.9998 for four film pairs, 0.9737 for two pairs, and 0.7209 for a single pair. We observe that the side lobes of the reflectance spectra vary more rapidly with wave vector variations as the total mirror thickness increases. This inverse relationship between physical thickness and periodicity in wave-vector or wave-length space is something we will see many examples of and come to expect as a general rule.

Viewed as a filter, the four-layer-pair Bragg mirror has a flat pass-band and sharp transition bands between the pass band and the rejection bands. These desirable characteristics are somewhat offset by the side lobes that in this example are unacceptably large for most filter applications. These side lobes can, however, be reduced by more sophisticated layer design that creates a softer transition into the high-reflectivity region of the filter. Such "edge-softening" or "appodization" techniques are used in many optical devices to avoid just the type of "ringing" represented by the side lobes of the reflectance spectra of the high-reflectivity mirrors of Fig. 3.11.

3.5.3 Photon Tunneling

The reflections from a thin, low-index film provide a vivid example of the importance of evanescent fields. Consider the structure shown in Fig. 3.12. Here we

have plane wave incident on a thin glass film between two thicker layers of silicon.

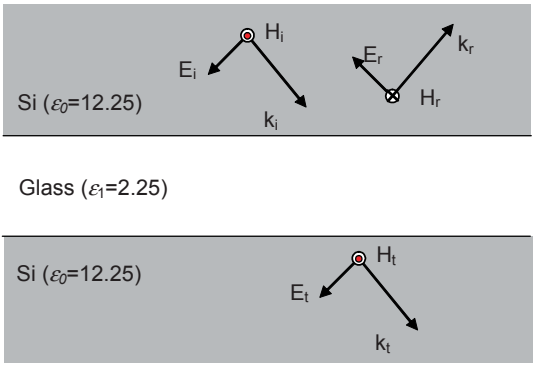


Figure 3.12. Structure that demonstrates photon tunneling. If the glass film is thick, then total internal reflection will guarantee that the reflection is unity beyond the TIR critical angle. In a film of finite thickness, the evanescent fields enable transmission of energy across the film and the reflection is reduced.

We expect based on our earlier observation of Total Internal Reflections that beyond the critical TIR angle all the energy of the incident wave will be reflected from the silicon-glass interface. Complete reflection beyond the TIR critical angle would indeed be the result if the film was infinitely thick, but with a thin film, we observe that there is transmission at all incident angle up to $\pi/2$! This is shown in Fig. 3.13 that shows reflection as a function of incident angle at 1.319 um wavelength for glass-film thicknesses of 2 um, 0.2 um, and 0.05 um.

The reflection from the 2 um film is roughly what we expect, even though the transition to unity reflection at the TIR critical angle is not as sharp as it would be for an infinitely thick film. The thinner films, however, clearly show that there is substantial transmission through the thin glass film beyond the critical TIR angle.

The finite thickness of the film also gives rise to interference fringes that can be observed for angles less than the TIR critical angle. When the film thickness increases, the interference fringes will as expected be more densely placed in angle space. Any finite film would show interference fringes, but an infinitely thick film would not, because there would be no second glass-silicon interface to provide interfering waves.

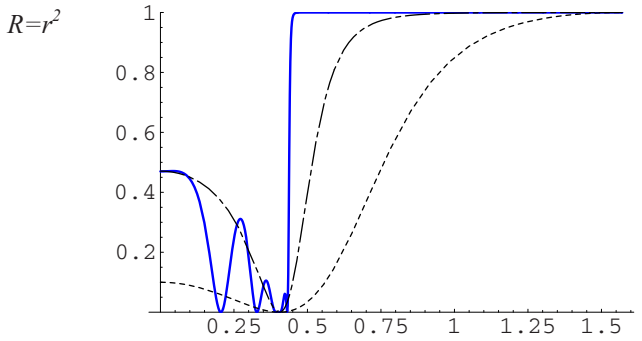


Figure 3.13. Reflection vs. incident angle of 1.319 um wavelength plane waves from a silicon-glass-silicon tunneling structure. The thicknesses of the glass layers are 2 um (solid), 0.2 um (dot-dashed) and 0.05 um (dashed). As the glass layer thickness is reduced, the evanescent fields in the glass facilitate photon tunneling that reduces the reflectivity to less than unity even beyond the TIR critical angle.

The explanation of the reduced reflectivity of the thin glass films is that the evanescent fields allow energy to flow across the film and to propagate away from the second interface as a plane wave in the silicon beyond the film. This type of transmission by evanescent waves over short distances, called tunneling, is utilized in many important optical devices, including fiber-optic directional couplers, Near-field Scanning Optical Microscopes (NSOMs), and photon tunneling sensors that we will discuss later.

3.5.4 Surface Plasmons

So far we have used the Fresnel reflections and their extensions to multi-layer structures with real values for the dielectric constants. There is, however, nothing in the formalism that forbids complex dielectric constants. In fact, the formulae that we have derived for reflectivity and transmission are all valid for interfaces and layered structures involving absorptive materials that must be described by complex dielectric constants.

As an example of a complex dielectric constant, consider the structure shown in Fig. 3.14. It is similar to the tunneling structure of Fig. 3.12, except that the bottom Si substrate is replaced by a gold layer. Other metals, e.g. silver or aluminum could be used instead, but gold is most often the choice in practical applications of surface plasmons because it has better optical properties in the near-infrared than aluminum and it is technologically superior to silver from a manufacturing point of view. Gold also has the advantage of being the substrate-material of choice for the preparation of a vast array of biological thin films.

The dielectric constant of gold at near-infrared wavelengths is negative and it has an imaginary part that signifies the fact the gold absorbs electromagnetic energy at these wavelengths. Specifically we have that gold at 1.319 um wavelength has a dielectric constant of -70.72-j7.06 [1]. The negative imaginary constant is due to our choice of signs in the description of plane waves. We chose to write plane waves in the form $\vec{E} = \vec{x} \cdot E_0 \cdot \cos(\omega t - kz)$. We could equally well have changed the signs of the argument in the co-sine function. The consequence would have been that we would have to also change the sign of the imaginary part of the dielectric constant of gold. This subtlety is important to keep in mind when looking up values for the dielectric constant of materials.

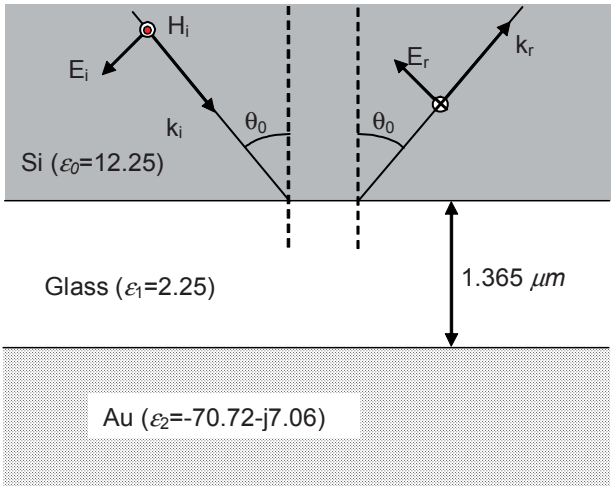


Figure 3.14. Schematic of the incident, reflected, transmitted and back reflected components of a Transversal-Magnetic field (TM) plane wave incident on a dielectric film bounded by a gold layer.

Figure 3.15 shows the reflectance from the Si-Glass-Au interface stack when the glass-film thickness is set to 1.365 um. The left-hand graph in the figure shows the reflectance at all angles of incidence from zero to $\pi/2$. It reveals that there is a very sharp reflectance minimum at one specific incident angle. The close up of the reflectance minimum in the right-hand graph shows that it occurs at an incident angle of ~ 0.45, which is beyond the critical TIR angle for the Si-Glass interface!

The explanation of the sharp and deep reflectance minimum is that the incident field tunnels through the glass film and couples to a Surface-Plasmon on the Glass-Au interface. This phenomenon is called Attenuated Total Internal Refection (ATIR). Coupling of the incident field to the surface Plasmon can only happen at the incident angle that allows the incident field and the surface plasmon to match wave vectors along the interface (phase match). Notice that this requires

that the electric field is in the plane of incidence, so, unlike photon tunneling that works for both TE and TM waves, only TM waves can couple to surface plasmons. To get a reflectance minimum that approaches zero as in Fig. 3.15, the glass film thickness must be carefully chosen such that the coupling of the incident field to the surface plasmon is exactly equal to the absorption in the gold film (impedance match).

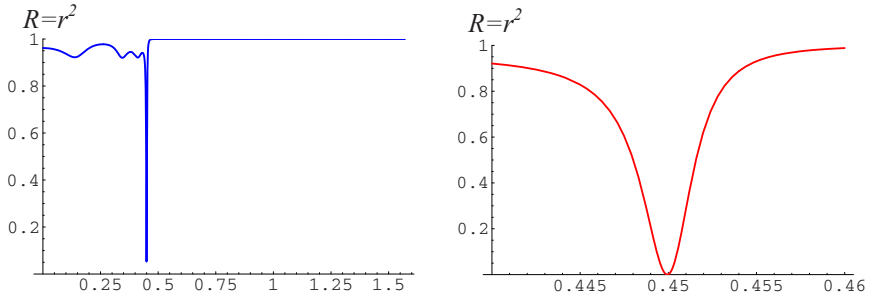


Figure 3.15. Reflectance vs. incident angle of 1.319 um light from a Si-Glass-Au structure, in which the thickness is optimized for coupling the incident plane wave to the Surface Plasmon on the Glass-Au interface.

3.6 Summary of Plane Waves

This chapter is in its entirety devoted to the study of how plane waves behave when they are incident on an interface or set of interfaces. Plane waves are mathematical structures that cannot be physically implemented with perfect accuracy, but are still useful models for free-space propagation of optical fields. Using the simple concept of phase continuity of planes waves at an interface, we were able to derive the Laws of Reflection and Refraction that form the basis of geometrical optics. The same principle of phase continuity also gave us the Fresnel-reflection formulas that led to descriptions of several very important optical phenomena, including Brewster angle, evanescent fields, Total-Internal-Reflection, and Goos-Hänchen shifts. Finally, we extended the Fresnel formulas to multilayer stacks and used them to demonstrate the existence of Anti-Reflection coatings, Bragg mirrors, photon tunneling, and surface plasmons.

The most important mathematical formulas that we derived in this chapter are summarized here.

$$\text{Law of reflection:} \quad \sin\theta_i = \sin\theta_r \tag{3.71}$$

$$\text{Snell's law of refraction:} \quad n_i \cdot \sin\theta_i = n_t \cdot \sin\theta_r \tag{3.72}$$

Fresnel Equations

Transmission and reflection for s-polarized waves (E-field polarized perpendicular to the plane of incidence)::

$$t_{TE} = \left.\frac{E_{0t}}{E_{0i}}\right|_{TE} = \frac{2n_i \cos\theta_i}{n_i \cos\theta_i + n_t \cos\theta_t} \tag{3.73}$$

$$r_{TE} = \left.\frac{E_{0r}}{E_{0i}}\right|_{TE} = \frac{n_i \cos\theta_i - n_t \cos\theta_t}{n_i \cos\theta_i + n_t \cos\theta_t} \tag{3.74}$$

Fresnel Equations for p-polarized waves (E-field polarized in the plane of incidence):

$$t_{TM} = \left.\frac{E_{0t}}{E_{0i}}\right|_{TM} = \frac{2n_i \cos\theta_i}{n_i \cos\theta_t + n_t \cos\theta_i} \tag{3.75}$$

$$r_{TM} = \left.\frac{E_{0r}}{E_{0i}}\right|_{TM} = \frac{n_i \cos\theta_t - n_t \cos\theta_i}{n_i \cos\theta_t + n_t \cos\theta_i} \tag{3.76}$$

Brewster angle: $\tan\theta_{iB} = \dfrac{n_t}{n_i}$

Critical Angle for Total Internal Reflection (TIR): $\theta_{cr} = \sin^{-1}\left(\dfrac{n_t}{n_i}\right)$ where $n_t < n_i$

The phase shift associated with TIR appears to correspond to an offset between the incident and reflecting planes.

Multilayer Stacks

The following matrix formulation can be used to calculate reflection and transmission from any combination of parallel interfaces.

$$\begin{bmatrix} E_I \\ H_I \end{bmatrix} = \begin{bmatrix} \cos(k_0 \cdot nd\cos\theta_1) & j\sin(k_0 \cdot nd\cos\theta_1)/\Gamma_1 \\ j\sin(k_0 \cdot nd\cos\theta_1)\cdot\Gamma_1 & \cos(k_0 \cdot nd\cos\theta_1) \end{bmatrix} = \begin{bmatrix} E_{II} \\ H_{II} \end{bmatrix} \tag{3.77}$$

$$\begin{bmatrix} E_I \\ H_I \end{bmatrix} = M_1 M_2 M_3 \ldots\ldots M_p \begin{bmatrix} E_{p+1} \\ H_{p+1} \end{bmatrix} = M \begin{bmatrix} E_{p+1} \\ H_{p+1} \end{bmatrix} \tag{3.78}$$

$$\left.\frac{E_{r0}}{E_{i0}}\right|_{TM} = r_{TM} = \frac{\Gamma_0 m_{11} + \Gamma_0 \Gamma_{p+1} m_{12} - m_{21} - \Gamma_{p+1} m_{22}}{\Gamma_0 m_{11} + \Gamma_0 \Gamma_{p+1} m_{12} + m_{21} + \Gamma_{p+1} m_{22}} \tag{3.79}$$

$$\left.\frac{E_{tp+1}}{E_{i0}}\right|_{TM} = t_{TM} = \frac{2\Gamma_0}{\Gamma_0 m_{11} + \Gamma_0 \Gamma_{p+1} m_{12} + m_{21} + \Gamma_{p+1} m_{22}} \tag{3.80}$$

$$\Gamma_{k,TM} = \sqrt{\frac{\varepsilon_0}{\mu_0}} n_k \Big/ \cos\theta_k \tag{3.81}$$

$$\Gamma_{k,TE} = \sqrt{\frac{\varepsilon_0}{\mu_0}} n_k \cdot \cos\theta_k \tag{3.82}$$

This formalism can be used to calculate the reflection and transmission through many important optical devices including Anti-Reflection coatings, Bragg filters, photon tunneling structures, and Surface Plasmon couplers.

Further Reading

E. Hecht, "Optics (4th Edition)", Addison-Wesley, 2002.

Exercises

Problem 3.1 - Phase shift through a glass slide

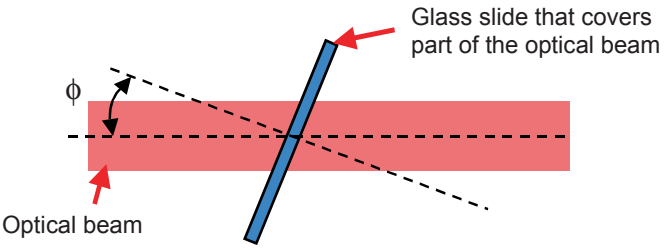
In the optics laboratory it is often useful to create a phase delay in an optical beam by letting it pass through a microscope cover slip (glass plate) as shown below. You can adjust the amount of phase delay the beam sees by changing the angle between the optical axis and the surface normal of the cover slip.

Prove that the phase shift of the part of the beam that passes through the glass relative to the other part of the beam is given by:

$$\theta = \frac{2\pi}{\lambda} h \cdot \left[n \cdot \cos(\phi_i) - \cos(\phi) \right]$$

where h is the thickness of the cover slip

n is the index of the glass
λ is the wavelength of the light
ϕ is the angle between the surface normal of the cover slip and the optical axis
ϕ_i is the angle of the optical beam inside the glass, i.e. $sin\phi = n \cdot sin\phi_i$.



Geometry of set-up for modulating the phase front of an optical beam.

Problem 3.2 – Waves at an interface

Consider the figure below that shows a TM-polarized plane wave incident on an interface between dielectrics with refractive indices of 1.5 and 1.4.

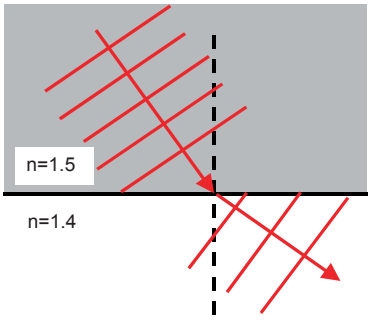


Figure 1: TM Plane wave incident on dielectric interface.

a) Graph the amplitude and phase of the reflection of the field as a function of incident angle for the TM wave. Calculate the Brewster angle and the critical angle for TIR, and indicate their positions in the graph.

The structure is changed so that there is only a thin film of the low-index material (Fig. 2).

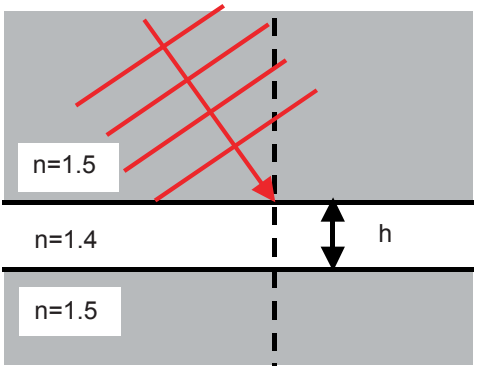
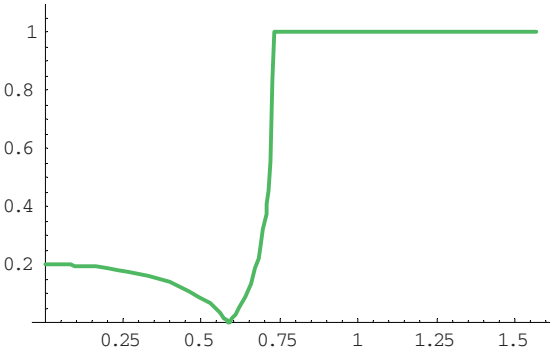


Figure 2 *TM plane wave incident on low-index film.*

b) Explain conceptually how the reflectivity changes as a function of the thickness, h, of the low-index film.

Problem 3.3 – Brewster angle

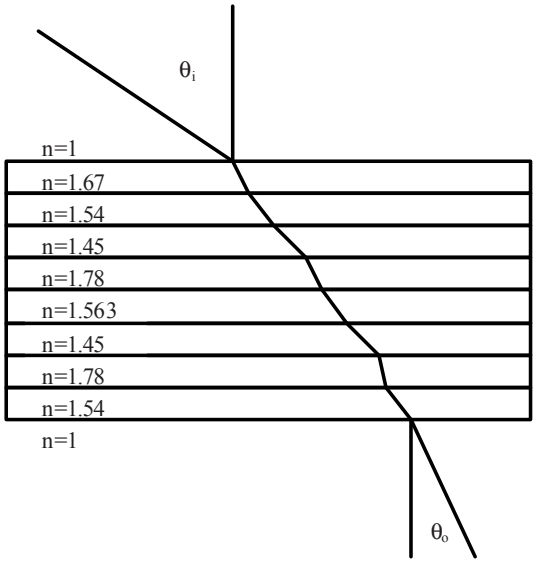
The figure shows a plot of the absolute value of the reflection of the electric field from an air-dielectric interface.



a) What is the polarization of the incident wave? Explain.
b) What is the index of the dielectric?

Problem 3.4 - Reflections from an interface

Calculate the output angle of the light after passing through the 8 layers of the structure below.

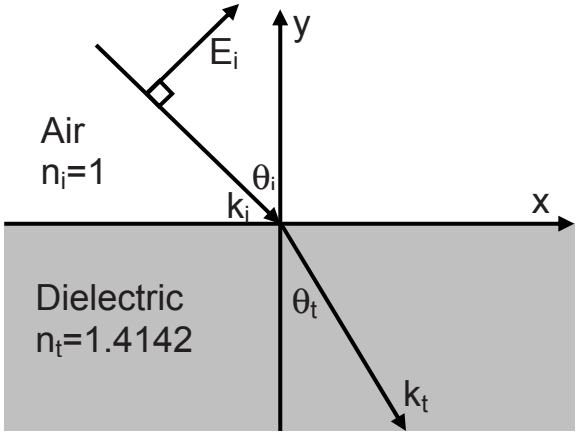


Problem 3.5 – Total Internal Reflection

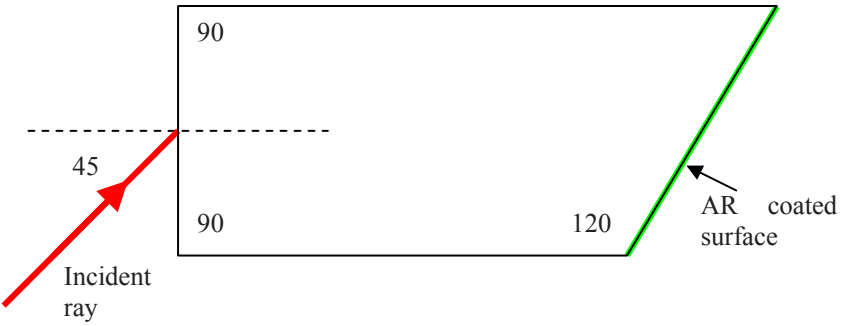
Explain the concept of total-internal-reflection using the wave description of light.

Problem 3.6 – Waves at an interface:

Consider the air-dielectric interface with an incident plane wave shown below. The incident angle is 45 degrees and the index of the dielectric is $n_t=2^{.5}\approx1.4142$.



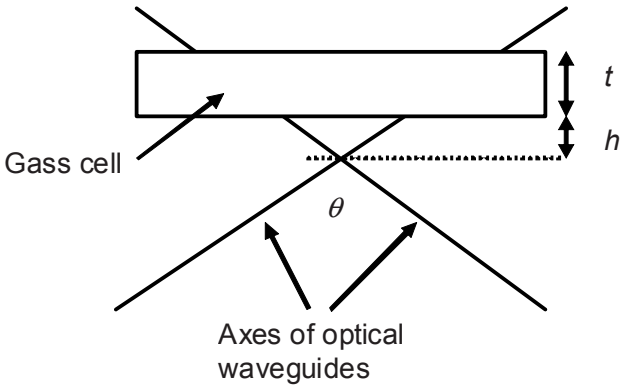
a) What is the direction of the propagation vector, $\vec{k}_t$, in the dielectric?
b) What is the direction of the magnetic field, $\vec{H}_t$, in the dielectric?
c) What is the direction of the electric field, $\vec{E}_t$, in the dielectric?
d) The same dielectric plate is used to make a prism as shown in Fig. 1.2. The incident angle is again 45 degrees. Using the ray model, draw all beams that are created by transmission and reflection of the incident ray. Note that the right surface of the prism has an anti-reflection (AR) coating.



Problem 3.7 - Champagne switch

Design a Champagne fiber optic switch based on TIR. Assume that the effective index of the waveguides is 1.5, that the index of the evaporated gas is 1, and that a cross-talk between channels of less than –40dB is required. Use the formulas for plane waves to calculate reflection and transmission.

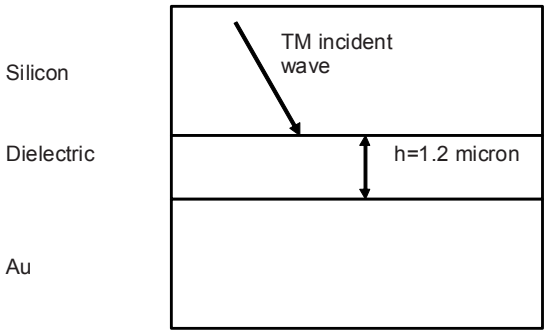
a) Choose values for the parameters θ, t, and h (defined in the figure) that will fulfill the requirements for the switch. Explain your choices and reasoning.
b) How closely must the index of the liquid match the index of the waveguides?



Problem 3.8 – Surface Plasmons

Calculate and plot the *TM* reflectance from the interface in the figure below as a function of incident angle at *1.319* micron wavelength. The relative dielectric constant are: *12.295, 2.66* and *–70.72-j7.06* for silicon, the dielectric and gold respectively at *1.319* micron wavelength.

What is the explanation for the sharp dip in reflectance at approximately 0.5 radians?



Problem 3.9 – Excitation of surface plasmons

In the preceding problem you calculated and plotted the *TM* reflectance from the interface in the figure above as a function of incident angle using the following parameters: λ*=1.319 micron*, $\varepsilon_{silicon}$*=12.295*, $\varepsilon_{dielectric}$*=2.66*, ε_{Au} *= –70.72-j7.06*, and found a sharp dip in reflectance at approximately 0.5 radians. Now assume that the silicon is a flat wafer or chip with parallel surfaces.

a) Can you design a film or stack of films that allow you to get the light into the silicon at the angle that allow you to observe the reflectance dip?
b) If you can, show an example of such a film or stack of films. If you cannot, design another optical system that will allow the light to get in at the right angle.

Problem 3.10 – Light Emitting Diodes

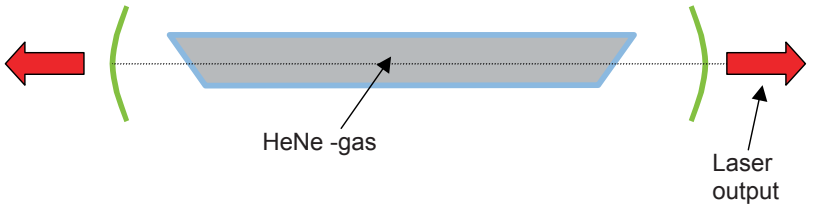
One of the problems with Light Emitting Diodes (LEDs) is that most of the light that is generated cannot escape from the semiconductor material. Assume that we can model an LED as a cube of semiconductor material with a refractive index of 3.5, and that the light is generated at the center of the cube.

a) Calculate approximately what percentage of the light that will be able to escape the LED.
b) c) Explain why laser light does not suffer from the same inability to escape the semiconductor material as light generated in an LED.

Problem 3.11 – Polarized laser light

The figure shows a typical He-Ne laser laser consisting of a gas-filled glass tube and external mirrors. Notice the angled facets of the gas tube.

What is the polarization of the laser light? Explain.



References

1. E.D. Palik, "Handbook of Optical Constants of Solids", Academic Press, Inc., Orlando, 1985.

4: Diffraction and Gaussian Beams

4.1 Introduction to Diffraction and Gaussian Beams

The plane waves we have studied in Chapters 2 and 3 are useful models that give insight into the operation and design of many optical devices. The mathematical description shows, however, that plane waves are not physically realizable, because if the wave has finite energy in any cross section, then the total energy is infinite. In practice we can create good approximations to plane waves, but we know from energy conservation arguments presented in Chapter 2 that any physically realizable beam cannot propagate in a homogeneous medium without diffraction.

Many important optical systems can be modeled without consideration of diffraction. This is particularly true for large-aperture systems like cameras and lithography tools. Microoptical devices, however, cannot be understood without wave diffraction, and in most cases, diffraction is the effect that limits miniaturization. In this Chapter we'll focus on Gaussian beam propagation as tools to model and develop intuitive understanding of diffraction phenomena.

There are many reasons to choose Gaussian beams as the starting point for the study of diffraction. First, Gaussian-beam theory provides a convenient mathematical formalism that allows closed-form solutions, or approximate solutions, to many diffraction problems. The simplicity of the Gaussian formalism has lead to its use in modeling of dispersion of (temporal) Gaussian pulses. The fundamental Gaussian is also a very good model for real modes on single-mode waveguides and for output modes of many important lasers. This is very useful for deriving closed-form expressions for mode-coupling phenomena.

A more fundamental reason for studying Gaussian beams is that they represent electromagnetic waves with minimum uncertainty, i.e. the product of the beam sizes in real space and wave-vector space is the theoretical minimum (in perfect analogy to Heisenberg's uncertainty principle in Quantum Mechanics). This means that the 0^{th}-order, or fundamental, Gaussian has less diffraction than any other optical fields of the same size. The fundamental Gaussian therefore establishes a lower limit for diffraction of real optical beams.

Finally, although the Gaussian beam is an approximate solution to the wave equation, we find that most major physical predictions come out correctly, so Gaussian beam theory establishes the correct physical understanding of diffraction.

4.2 Paraxial Wave Equation

Gaussian beams are paraxial approximations to the wave equation. This means that they are good approximations for light "beams", i.e. optical waves that have a well-defined direction of propagation, along which the variation of the beam cross section is relatively slow (a more precise formulation will follow).

To derive the Gaussian beam equations we solve the paraxial wave equation. Consider a monochromatic optical field in the following form

$$E(x,y,z,t) = E_0(x,y,z)e^{j\omega t} \tag{4.1}$$

The spatial part must be a solution to the equation

$$\left(\nabla^2 + k^2\right)E_0(x,y,z) = 0 \tag{4.2}$$

We now assume that the light is propagating in the z-direction with only a slow variation of its envelope. In other words, we express the field as the product of an envelope with a slow z-dependence and a phase term with a rapid z-variation

$$E_0(x,y,z) = u(x,y,z)\exp[-jkz] \tag{4.3}$$

When we insert this formulation of the optical field in the wave equation, we will assume that the term $\partial^2 u/\partial^2 z$ can be neglected. This is the paraxial approximation, resulting in the paraxial wave equation

$$\nabla_\perp^2 u - 2jk\frac{\partial u}{\partial z} = 0 \tag{4.4}$$

where

$$\nabla_\perp^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \tag{4.5}$$

The paraxial wave equation

$$\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}u - 2jk\frac{\partial u}{\partial z} = 0 \tag{4.6}$$

has the solution

$$u = \exp\left[-j\left(P + \frac{k}{2q}r^2\right)\right] = \exp\left[-j\left(P + \frac{k}{2q}\left(x^2 + y^2\right)\right)\right] \tag{4.7}$$

where *P(z)* is a complex phase shift and *q(z)* is a complex beam parameter associated with beam propagation.

4.2.1 The Fundamental Gaussian Profile

We now insert the above solution into the paraxial wave equation and compare terms in equal powers in *r* to get

$$\frac{\partial}{\partial z}P = -\frac{j}{q} \tag{4.8}$$

$$\frac{\partial}{\partial z}q = 1 \Rightarrow q_2 = q_1 + z \tag{4.9}$$

In the next section, we will use the first of these equations to find the phase of the Gaussian Beam. The second equation relates the complex beam parameter in one plane (1) to another (2) separated by *z*. We express this complex beam parameter in terms of two real beam parameters

$$\frac{1}{q} = \frac{1}{R} - j\frac{\lambda}{\pi \cdot \omega^2} \tag{4.10}$$

With this definition of the beam parameter, the envelope f the Gaussian beam can be written

$$u = \exp\left[-j\left(P + \frac{k}{2}r^2\left(\frac{1}{R} - j\frac{\lambda}{\pi \cdot \omega^2}\right)\right)\right] \tag{4.11}$$

This expression shows that *R* is the radius of curvature of the wavefront of the propagating beam, and ω is the beam radius (*1/e* for the field).

The field envelope is shown graphically in Fig. 4.1. It has a symmetric bell shape with rapidly decreasing field as a function of distance from the center. The fact that it has infinite extent, i.e. the field does not completely vanish for any values of *r*, means that Gaussian Beams, like plane waves, cannot be perfectly reproduced in practical experiments.

Unlike plane waves, however, Gaussian Beams can be truncated and still behave very much like perfect Gaussians. Later we will study truncated Gaussians and find that centered apertures with a diameter larger than 3ω have relatively minor effects on Gaussian Beam propagation.

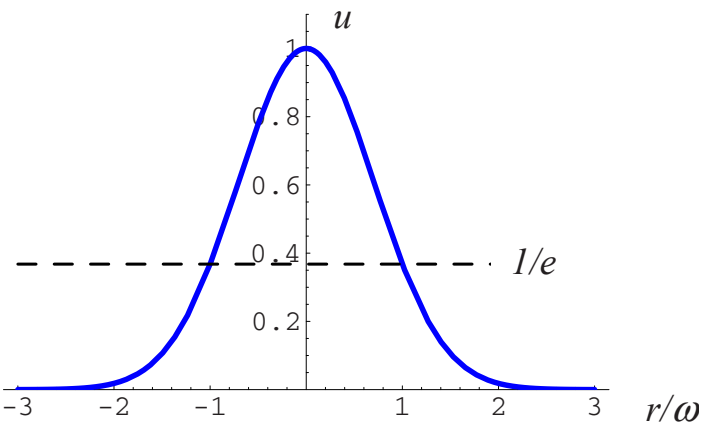


Figure 4.1 Fundamental Gaussian beam profile.

4.2.2 Beam Waist

The Gaussian beam narrows to a minimum radius, ω_0, called the beam waist. The expressions for the fundamental Gaussian beam become particularly simple when written in terms of ω_0 with z measured from the waist. At the waist, the beam parameter is purely imaginary

$$q_0 = j\frac{\pi \cdot \omega_0^2}{\lambda} \tag{4.12}$$

With the origin of the z-axis at the waist we have

$$q(z) = q_0 + z = j\frac{\pi \cdot \omega_0^2}{\lambda} + z \tag{4.13}$$

Using this expression with the definition of the complex beam parameter (Eq. 4.10), we find the beam radius and curvature as a function of z (where z is now the distance from the beam waist)

$$\omega^2(z) = \omega_0^2\left[1 + \left(\frac{\lambda \cdot z}{\pi \cdot \omega_0^2}\right)^2\right] \tag{4.14}$$

$$R(z) = z\left[1 + \left(\frac{\pi \cdot \omega_0^2}{\lambda \cdot z}\right)^2\right] \tag{4.15}$$

These expressions show how the fundamental Gaussian changes as a function of propagation distance, z.

The factor P in the expression for the Gaussian mode can now be calculated

$$\frac{\partial}{\partial z}P = -\frac{j}{q} = -\frac{j}{j\frac{\pi \cdot \omega_0^2}{\lambda} + z} \Rightarrow jP = \ln\left[\frac{\omega}{\omega_0}\right] - j\arctan\left(\frac{z\lambda}{\pi \cdot \omega_0^2}\right) \tag{4.16}$$

The real part of this expression is simply a renormalization accounting for the fact that the beam is diverging as a function of z, while the imaginary part shows that the Gaussian beam accumulates an extra phase shift (compared to a plane wave) when propagating.

Using these established relationships, the fundamental Gaussian beam can be expressed

$$u = \frac{\omega_0}{\omega}\exp\left[-j\left(\phi + \frac{k}{2}r^2\left(\frac{1}{R} - j\frac{\lambda}{\pi \cdot \omega^2}\right)\right)\right] \tag{4.17}$$

where

$$\phi = \arctan\left(\frac{\lambda \cdot z}{\pi \cdot \omega_0^2}\right) \tag{4.18}$$

This last term quantifies the "extra" phase that the fundamental Gaussian beam accumulate compared to a plane wave. The functional dependence on propagation distance shows that this extra phase shift, called Gouy phase, is concentrated around the focus of the beam, and that it only contributes a total phase of π radians. For most calculations we can therefore assume that a Gaussian beam accumulates phase in the same fashion a plane way.

Another important concept is the Rayleigh range, which is defined as the distance from the waist to the point where the beam radius has increased by $\sqrt{2}$.

$$\omega^2(z) = 2\omega_0^2 = \omega_0^2\left[1 + \left(\frac{\lambda \cdot z_R}{\pi \cdot \omega_0^2}\right)^2\right] \Rightarrow \frac{\lambda \cdot z_R}{\pi \cdot \omega_0^2} = 1 \Rightarrow z_R = \frac{\pi \cdot \omega_0^2}{\lambda} \tag{4.19}$$

The confocal parameter is defined as twice the Rayleigh length.

The Rayleigh length can be though of as the length of the "collimated" section of a Gaussian beam. This might seem like an arbitrary choice, but we will find that a remarkable number of widely different miniaturized optical systems are optimized when we chose to separate focusing elements by the Rayleigh length of the propagating Gaussian beam.

Finally it is useful to define a far-field diffraction angle for the Gaussian beam

$$\theta = \lim_{z \to \infty} \frac{\omega(z)}{z} = \frac{\lambda}{\pi \cdot \omega_0} \tag{4.20}$$

In this definition we have made the choice that angular extent is defined by the beam radius $\omega(z)$. This is of course completely arbitrary. When we discuss design of scanning Optical MEMS, we will introduce application-dependent criteria for Gaussian-beam widths that give far-field diffraction angles that are proportional to (same dependence on ω_0 and λ), but numerically different from the one defined here.

The parameters of Gaussian-beam propagation are illustrated in Fig. 4.2. The main concept is simply that any beam of any cross-sectional phase and amplitude distribution, has a converging regime, a focus, and a diverging regime. The fundamental Gaussian has the smallest possible waist-angle product of λ/π.

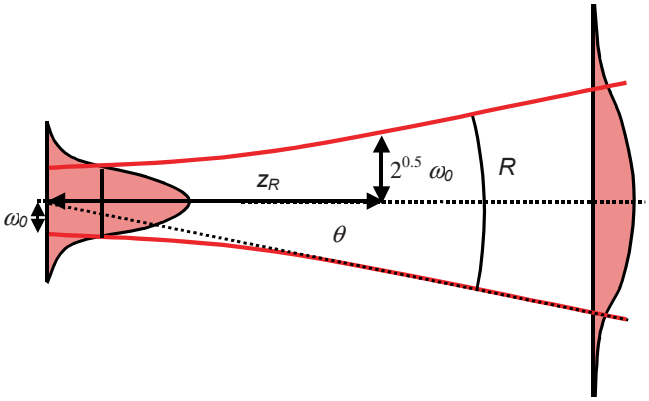


Fig. 4.2 *Illustration of the propagation of a fundamental Gaussian beam.*

4.2.3 Higher Order Gaussian Modes

The fundamental Gaussian is the most important solution to the paraxial wave equation, but it is by no means the only one. To find further solutions, we will try the following expressions

$$u = H_l\left(\frac{\sqrt{2}x}{\omega}\right)H_m\left(\frac{\sqrt{2}y}{\omega}\right)\exp\left[-j\left(P+\frac{k}{2q}\left(x^2+y^2\right)\right)\right] \tag{4.21}$$

Substituting this into the paraxial wave equation, we find the following equation for the functions $H_{l,m}$.

$$\frac{\partial^2}{\partial x^2}u+\frac{\partial^2}{\partial y^2}u-2jk\frac{\partial u}{\partial z}=0\Rightarrow\frac{\partial^2 H_l}{\partial x^2}-2x\frac{\partial H_l}{\partial z}+2\cdot l\cdot H_l=0 \tag{4.22}$$

This is the defining equation for the Hermite-Gaussian mode of order *m,l*. (If we wrote the paraxial equation in cylindrical coordinates, we would find the equation for the Laguerre-Gaussian functions). Some of the lower-order Hermite-Gaussians are summarized below

$$H_0=1 \tag{4.23}$$

$$H_1=2x \tag{4.24}$$

$$H_2=4x^2-2 \tag{4.25}$$

$$H_3=8x^3-12x \tag{4.26}$$

The slowly-varying envelope of a Hermite-Gaussian mode of order *m,l* can then be expressed similarly to the fundamental Gaussian

$$u(r,z)= H_l\left(\frac{\sqrt{2}x}{\omega}\right)H_m\left(\frac{\sqrt{2}y}{\omega}\right)\frac{\omega_0}{\omega(z)}\exp\left[-jkz+j\phi(z)-\frac{r^2}{\omega^2(z)}-jk\frac{r^2}{2R(z)}\right] \tag{4.27}$$

Here the beam radius and radius of curvature are defined exactly as for the fundamental, i.e. we have $\omega(z)=\omega_0\sqrt{1+\left(z/z_R\right)^2}$ and $R(z)=z+z_R^2/z$ as before. The Rayleigh length is also defined as for the fundamental; $z_R=\pi\omega_0^2/\lambda$.

The intensity profiles of some of the low-order Hermite-Gaussian are shown in Fig. 4.3.

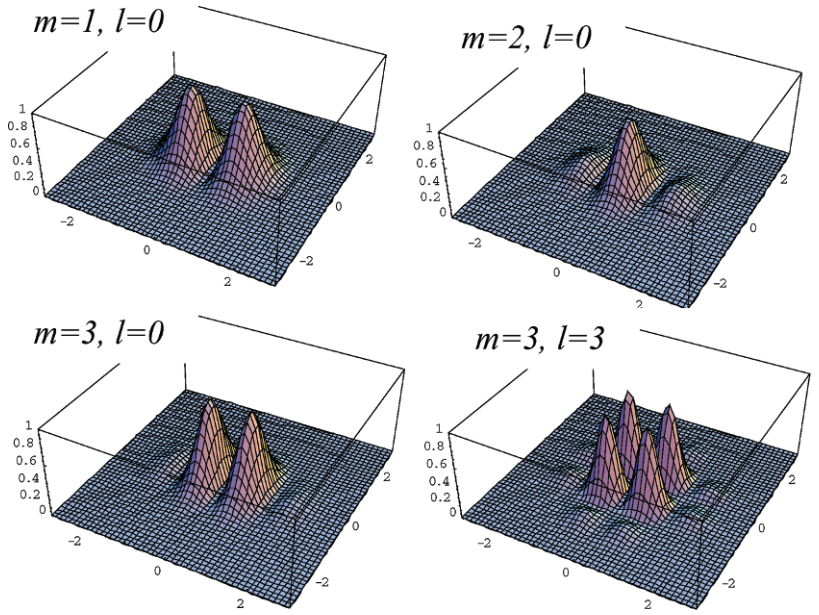


Figure 4.3 *Intensity profiles of Gaussian TEM_{10}, TEM_{20}, TEM_{30}, and TEM_{33} modes.*

The higher-order Hermite-Gaussians functions are different from the fundamental in that they accumulate phase differently through focus. The Gouy phase of an Hermite Gaussian function of order *m,l* can be written

$$\phi(z) = (m + l + 1)\tan^{-1}\left(\frac{z}{z_R}\right) \tag{4.28}$$

This is the only parameter that depends on the mode numbers, *m,l*. This through-focus phase shift is important in establishing the resonance frequency in optical resonators, but, as for the fundamental Gaussian, it can most often be ignored in system designs that support freely propagating beams.

We conclude that the propagation of the higher-order Hermite Gaussians is very similar to the propagation of the fundamental. Higher-order modes occupy larger areas for a given beam radius, because the basic exponential function is multiplied by a higher order polynomial, but all Hermite-Gaussian modes propagate according to the same simple rule ($q_2 = q_1 + z$), and all are described by the same two basic parameters; beam radius and radius of curvature. Only their transversal field distributions and their Gouy phase shifts are different. This is illustrated in Fig. 4.4 that shows the the propagation of the *1,0* Hermite-Gaussian mode.

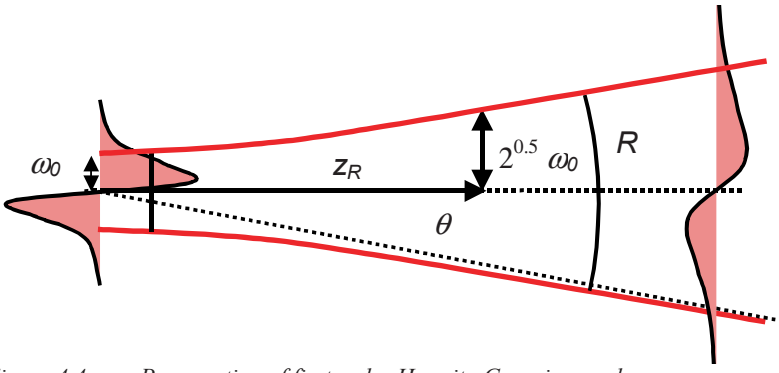


Figure 4.4 Propagation of first order Hermite-Gaussian mode.

The Hermite-Gaussian modes form a complete, orthogonal set of functions, so an arbitrary optical field can be expanded on the Hermite-Gaussians. A general diffraction problem, i.e. one where we want to compute the shape of an arbitrary field distribution as it propagates over a certain distance, can therefore be solved by writing an arbitrary field distribution as a sum of Gaussians, calculate the propagation of each of the individual Gaussian modes, and re-sum. The different Gouy phase shifts must be taken into account when propagating and summing sets of Hermite-Gaussian modes.

This method is conceptually appealing, but computationally inefficient due to the large number of Gaussian modes needed to represent complex beam profiles. The most common use of Gaussian is therefore (1) as a conceptual tool, and (2) for detailed calculations of propagation of simple beams (i.e. beams that are close approximations to Gaussians). We will use Gaussian beam propagation extensively for both these purposes.

4.3 Gaussian Beam Transformation in Lenses

To analyze optical systems, we need to understand how Gaussians are affected by lenses. Together with the simple law of Gaussian Beam propagation found above, this will give us the tools needed to model a large number of practical systems.

An ideal lens does not change the transverse distribution of an optical field, so a Gaussian Beam will remain in the same order after passing through a lens. The radius of curvature, $R(z)$, and the beam radius, $\omega(z)$, will in general change when passing through a thick lens consisting of two diffracting surfaces separated by a significant propagation distance .

In a thin lens, which we can think of as two diffracting surfaces in the same plane, only $R(z)$ changes. The beam radius does not change, because there is no propaga-

tion distance over which a change can take place. When considering Gaussian beam propagation through systems of lenses, we therefore typically treat thick lenses as a combination of thin lenses separated by regions of unguided propagation. To see how this is done, consider propagation of a fundamental Gaussian beam through a thin lens as shown in Fig. 4.5.

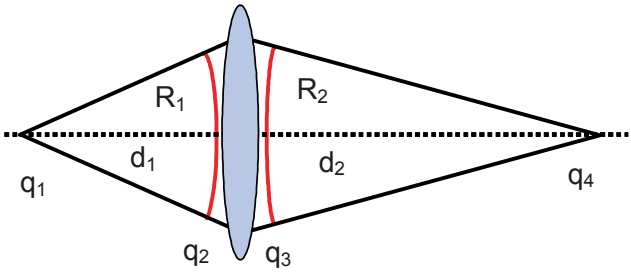


Figure 4.5 Propagation of a Gaussian beam through a thin lens.

The beam propagates from z_1, where it is characterized by the beam parameter q_1, to the lens at position z_2, according to the propagation law

$$q_1 = q_2 + z \tag{4.29}$$

To understand how a thin lens changes the radius of curvature of an optical, consider the geometrical-optics lens law

$$\frac{1}{d_1} + \frac{1}{d_2} = \frac{1}{f} \tag{4.30}$$

where f is the focal length and $d_{1,2}$ are defined in Fig. 4.5.

Comparing Gaussian and geometrical optics, we have that $d_1 = R_1$, and $d_2 = -R_2$, so

$$\frac{1}{R_1} - \frac{1}{R_2} = \frac{1}{f} \tag{4.31}$$

This expression can be considered the definition of a thin lens. We know that a thin lens does not change the beam radius, so we can rewrite this equation in terms of the Gaussian beam parameter

$$\frac{1}{q_1} - \frac{1}{q_2} = \frac{1}{f} \tag{4.32}$$

Repeated applications of this lens law and the propagation law for Gaussian Beams allow us to find the transformation of the beam through any system of optical elements.

Referring back to Fig. 4.5, we have for the beam parameter at the left side of the lens

$$q_2 = q_1 + d_1 \tag{4.33}$$

The lens changes this into

$$\frac{1}{q_3} = \frac{1}{q_2} - \frac{1}{f} \Rightarrow q_3 = \frac{f q_2}{f - q_2} \tag{4.34}$$

Finally we have for the beam waist at the distance d_2 from the lens

$$\begin{aligned} & q_4 = q_3 + d_2 \\ & \Rightarrow q_4 = \frac{f(q_1 + d_1)}{f - q_1 - d_1} + d_2 = \frac{\left(1 - \frac{d_2}{f}\right) q_1 + d_1 + d_2 - \frac{d_1 d_2}{f}}{1 - \frac{q_1}{f} - \frac{d_1}{f}} \end{aligned} \tag{4.35}$$

4.3.1 Focusing and Collimation of Gaussian Beams

We will now use the expressions we have derived for the beam parameter in a lens system to investigate focusing of Gaussian beams. We assume that the incoming wavefront on the lens is flat ($R_1=\infty$), i.e. we are considering the special case of $d_1=0$ (i.e. position 1 and 2 are the same), as shown in Fig. 4.6.

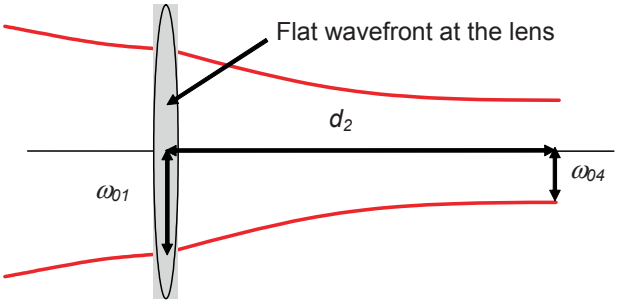


Figure 4.6 *Illustration of focusing of a Gaussian beam by a lens. The beam has a waist (flat wave front) at the lens. If the curvature is finite at the lens, it can be accounted for by defining an effective focal length of the lens (see text).*

The expression we just found for the beam parameter to the right of the lens then simplifies to

$$d_1 = 0 \Rightarrow q_4 = \frac{\left(1 - \frac{d_2}{f}\right) q_1 + d_2}{1 - \frac{q_1}{f}} = \frac{Aq_1 + B}{Cq_1 + D} \tag{4.36}$$

In this expression we used the ABCD matrix notation (se Appendix A) for a lens

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} = \begin{vmatrix} 1 - d_2/f & d_2 \\ -1/f & 1 \end{vmatrix} \tag{4.37}$$

The assumption that the input beam has its waist at the lens is not as restrictive as it may seem, because if the wavefront has finite curvature at the lens, we may simply incorporate that curvature into an effective focal length, f_{eff}, given by

$$\frac{1}{R_1} - \frac{1}{R_2} = \frac{1}{f} \Rightarrow \frac{1}{f} - \frac{1}{R_1} = -\frac{1}{R_2} \Rightarrow \frac{1}{f_{eff}} = \frac{1}{f} - \frac{1}{R_1} \tag{4,38}$$

With a flat wavefront at the lens, the equations for the beam parameter at the waist at the left simplifies to

$$\frac{1}{q_1} = -\frac{j\lambda}{\pi \cdot \omega_{01}^2} = -\frac{j}{z_{R1}} \tag{4.39}$$

so that we can write

$$\frac{1}{q_4} = \frac{-1/f - \frac{j}{z_{R1}}}{1 - d_2/f - \frac{jd_2}{z_{R1}}} = \frac{-\frac{1}{f} + \frac{d_2}{f^2} + \frac{d_2}{z_{R1}^2} - \frac{j}{z_{R1}}}{(1 - d_2/f)^2 + \frac{{d_2}^2}{z_{R1}^2}} \tag{4.40}$$

At the waist, the complex beam parameter is purely imaginary, so we find d_2 by requiring the real part to be zero

$$-\frac{1}{f} + \frac{d_2}{f^2} + \frac{d_2}{z_{R1}^2} = 0 \Rightarrow d_2 = \frac{1/f}{1/f^2 + 1/z_{R1}^2} = \frac{f}{1 + f^2/z_{R1}^2} \tag{4.41}$$

We see that the distance, d_2, to the waist is always less than f, and it has a maximum value given by:

$$\frac{d}{d\left(f/z_{R1}\right)}\frac{d_2}{z_{R1}}=\frac{1-f^2/z_{R1}^2}{\left(1+f^2/z_{R1}^2\right)^2}=0\Rightarrow z_{R1}=f\Rightarrow$$
$$\left(\frac{d_2}{z_{R1}}\right)_{\max}=\frac{1}{2} \tag{4.42}$$

The beam waist can never be further away from the lens than half the Rayleigh length of the beam at the lens. This occurs when the focal length equals the Rayleigh length.

The Gaussian beam parameter is purely imaginary at the focus. Using this fact, we find that the beam waist at the focus is given by

$$\frac{\omega_{04}}{\omega_{01}}=\frac{f/z_{R1}}{\sqrt{1+f^2/z_{R1}^2}} \tag{4.43}$$

From this equation we see that the beam radius at the focus increases asymptotically towards the beam radius at the lens with increasing focal length. For very short focal lengths, the beam radius becomes

$$f<<z_{R1}\Rightarrow\omega_{04}\approx\frac{f\lambda}{\pi\cdot\omega_{01}} \tag{4.44}$$

and

$$d_2=\frac{f}{1+f^2/z_{R1}^2}\approx f \tag{4.45}$$

In this limit the paraxial approximation breaks down, because $f<<z_{R1}$ implies that the optical field is converging at a large angle. We will see, however, that Gaussian Beam theory correctly predicts beam behavior even in this limit.

The formulas we have found for the distance-to-focus and beam radius are plotted in Fig. 4.7. We see that focusing and collimation of Gaussian beams can be summarized as follows:

$$f<<z_R \Rightarrow\ d_2=f\ \text{ and }\ \omega_{04}=\frac{f}{z_{R1}}\omega_{01} \tag{4.46}$$

$$f=z_R \Rightarrow\ d_{2\max}=\frac{f}{2}=\frac{z_R}{2}\ \text{ and }\ \omega_{04}=\frac{1}{\sqrt{2}}\omega_{01} \tag{4.47}$$

$$f>>z_R \Rightarrow\ d_2 = \frac{z_R^2}{f} \text{ and } \omega_{04} = \omega_{01} \tag{4.48}$$

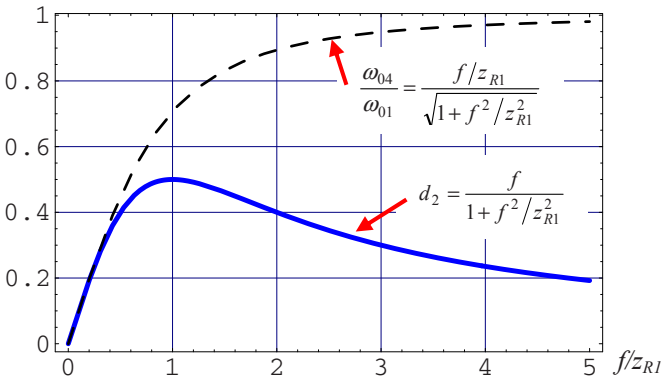


Figure 4.7 *Plots of distance-to-focus (solid) and beam radius (dashed) of a Gaussian beam focused by a lens. The parameters are normalized to the Rayleigh length and plotted as functions of the focal length also normalized to the Rayleigh length.*

The third of these regimes, described by Eq. 4.48, is not particularly interesting. It simply says that if the focal length of the lens is long compared to the Rayleigh length, then the lens doesn't do much; the beam-waist radius is unchanged and appears very close to the lens. This is what we would expect from a weak lens. The two other cases, on the other hand, have important consequences for miniaturization of optics, so we'll study each in some detail.

4.4 Resolution of a Lens

The resolution, or point-spread-function, of a lens is the size of the image that lens creates when the object is a mathematical point. In general, it will depend on the imaging condition. Most often we consider the imaging of a point at infinity so that the incident light on the lens has a flat wavefront and the image is formed in the focal plane as indicated in Fig. 4.8. The figure explicitly shows the finite aperture of the lens, and defines the two most important characteristic lens parameters; the focal length, f, and the lens-aperture diameter, D. The ratio f/D is called the f-number of the lens.

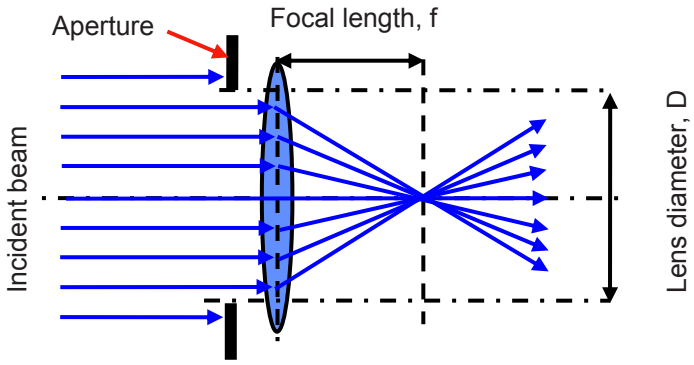


Figure 4.8 *Schematics of a perfect lens that according to the Geometrical Optics model will focus parallel input rays to a single point in the focal plane. The most important lens parameters are the focal length, f, and the aperture diameter, D.*

In Geometrical Optics, the size of the image is simply determined by the quality of the lens. If the lens is designed and manufactured perfectly, then Geometrical Optics predicts that the image is a point in the focal plane. It follows that a larger lens would create a less well-defined point image, because a larger lens surface would have more imperfections that a smaller one. This loss of resolution with increasing aperture is what we observe in many practical systems, e.g. cameras that are limited by lens imperfections in the range of apertures that are typically used.

In optical microsystems the situation is the opposite. The limited apertures required by miniaturization increases diffraction to the point where it typically dominates over the effects of lens imperfections. To understand the effect of the lens aperture on resolution, consider the formula we have derived for the Gaussian-waist radius created by a lens in the limit of short focal lengths

$$\omega_{image} = \frac{f/z_{R1}}{\sqrt{1+f^2/z_{R1}^2}}\omega_{lens} \xrightarrow[\frac{f}{z_{R1}}\to 0]{} \frac{f}{z_{R1}}\omega_{lens} = \frac{f\lambda}{\pi\cdot\omega_{lens}} \tag{4.49}$$

To proceed we must relate the beam diameter at the lens to the lens diameter. We will see shortly that if the aperture equals three times the beam radius, then the beam truncated by the aperture will behave as a Gaussian beam for practical purpose, so we set $D=3\omega_{lens}$. Likewise we say that the image diameter is three times the image beam radius. The image beam diameter created by a strong lens (short focal length) can then be expressed

$$d_{image} = 3\omega_{image} = \frac{3f\lambda}{\pi\cdot\omega_{lens}} = \frac{9f\lambda}{\pi\cdot D} \approx 2.9\cdot\frac{f}{D}\lambda \tag{4.50}$$

In contrast to the predictions made by Geometrical Optics, we have found that the point-spread function or resolution of a lens is of finite extent as shown in Fig. 4.9. The size of the focus is fundamentally limited by the f-number and the wavelength. The inverse dependence of focal-spot size on lens diameter is of course a challenge for designers of optical microsystems that ideally would have both small apertures and small spot sizes.

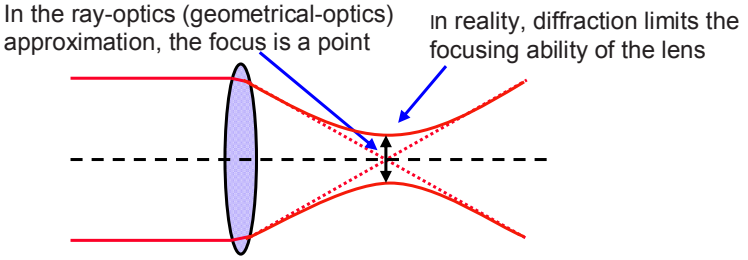


Figure 4.9 *Geometrical Optics predicts that a perfect lens can focus parallel input rays to a single point in the focal plane. Gaussian-Beam theory predicts that the focus will have a beam radius* $\omega_{focus} = f\lambda/(\pi\omega_{lens})$.

Our expression for the Gaussian beam diameter created by a lens is very close to the classical expression $\left(d_{Airy} = 2 \cdot r_{Airy} = 4.4 \cdot \frac{f}{D}\lambda \right)$ found for the diameter of the Airy disc, which is the central lobe of the diffraction-limited pattern created by a homogeneously illuminated lens [1]. The main source of numerical discrepancy between the two expressions is simply the differences in the definitions of the diameters. The correspondence of the formulae we have derived and the result of classical diffraction calculations show that the paraxial Gaussian theory correctly predicts lens focusing even in the non-paraxial, large convergence-angle limit.

4.4.1 Focusing into High-Index Media

The fact that the spot size created by a lens is directly proportional to the wavelength makes short wavelength sources the preferred solution for critical imaging and data storage applications. That has driven the development of short-wavelenght sources for optical lithography and for optical-disk (CDs and DVDs) readers. Another way to shorten the wavelength is to use a high index medium. All the formulas we have derived for Gaussian beams are perfectly valid for propagation in any uniform optical medium as long as we use the wavelength in the medium in the expressions. The possibility of achieving smaller spot sizes in-

side high-index optical media therefore motivates us to consider Gaussian Beam propagation across a dielectric interface.

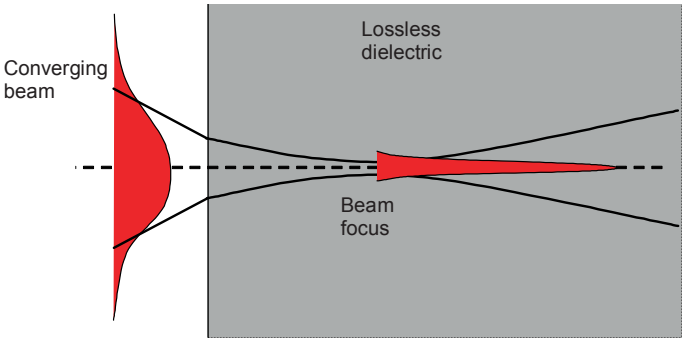


Figure 4.10. A converging Gaussian beam crossing a vacuum-dielectric interface creating a focus the high-index, lossless dielectric material. We find that the beam focus is the same as the one the converging beam would create without the presence of the dielectric.

Consider a Gaussian beam crossing the planar interface between vacuum and an optical material of index n at normal incidence as shown in Fig. 4.10. The converging beam comes to a focus within the dielectric. The beam radius does not change at the interface, but the radius of curvature increases due to refraction as illustrated in Fig. 4.11. Using Snell's law we can write

$$n_1 \cdot \sin\frac{\omega}{R_1} = n_2 \cdot \sin\frac{\omega}{R_2} \Rightarrow R_2 = \frac{\omega}{\sin^{-1}\left(\frac{n_1}{n_2} \cdot \sin\frac{\omega}{R_1}\right)} \tag{4.51}$$

In this paraxial limit, this simplifies to

$$R_2 = \frac{n_2}{n_1} R_1 \tag{4.52}$$

The radius of curvature increases by the ratio of the indexes at the dielectric interface.

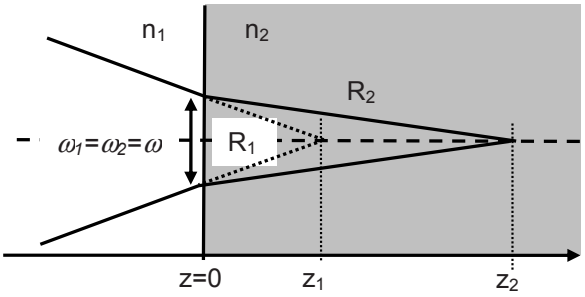


Figure 4.11. The radius of curvature of a Gaussian beam at a dielectric interface increases by a factor equal to the ratio of the indexes.

To understand how this change in radius of curvature affects the beam radius at the focus, we combine the definition of the beam parameter $\left(\frac{1}{q}=\frac{1}{R}-j\frac{\lambda}{\pi\cdot\omega^2}\right)$ with the beam-parameter propagation law $\left(q(z)=q_0+z=j\frac{\pi\cdot\omega_0^2}{\lambda}+z\right)$, and solve for the beam radius at the focus in the absence of the interface

$$\omega_{01}=\frac{\frac{\lambda_0^2}{n_1^2\pi^2\omega^2}}{\frac{1}{R_1^2}+\frac{\lambda^2}{n_1^2\pi^2\omega^4}} \tag{4.53}$$

With the dielectric in place the beam radius changes to

$$\omega_{02}=\frac{\frac{\lambda_0^2}{n_2^2\pi^2\omega^2}}{\frac{1}{R_2^2}+\frac{\lambda^2}{n_2^2\pi^2\omega^4}}=\frac{\frac{\lambda_0^2}{n_2^2\pi^2\omega^2}}{\frac{n_1^2}{n_2^2}\frac{1}{R_1^2}+\frac{\lambda^2}{n_2^2\pi^2\omega^4}}=\omega_{01} \tag{4.54}$$

The interface does not change the beam radius at focus! If, on the other hand, we use the same approach to solve for the distance from the interface to the focus with and without the high-index material, we find that that the distance to the focus increases by the index ratio. The normal-incidence crossing of a dielectric planar interface therefore moves the focus further along the optical axis, but it does not reduce the beam radius at the waist.

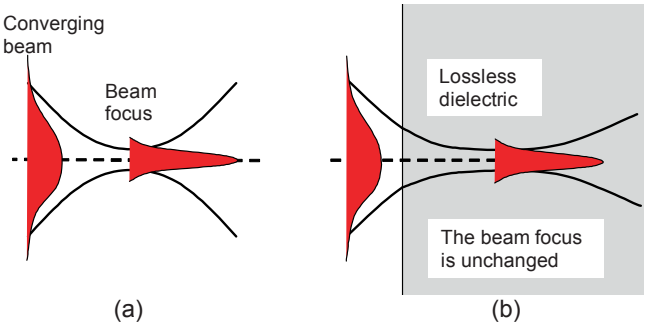


Figure 4.12. *(a) A converging optical beam creates a well-defined focus in air. (b) If a high-index, lossless dielectric material is inserted in the beam pathwith a planar surface perpendicular to the optical axis, then the beam focus moves, but its beam radius does not change.*

This result has important consequences for optical microscopy and lithography. It says that we cannot use the wavelength reduction of high-index materials to improve optical resolution by simply focusing the light into the material through a planar interface. On the other hand it is well known, and of great practical utility, that a lens that is part of the high-index material, or that is placed directly in contact with it (immersion lens), will lead to a reduced beam focus as illustrated in Fig. 4.13.

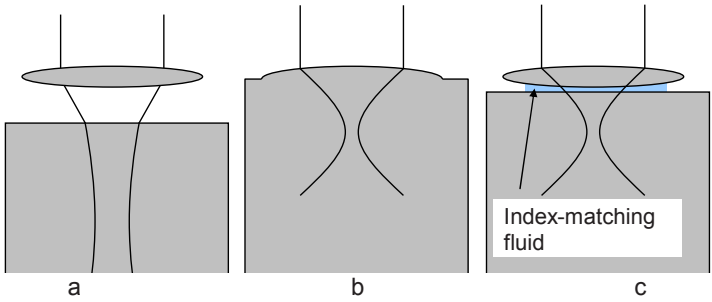


Figure 4.13. *The shorter wavelength of high-index materials does not lead to a reduction of the focus if the focusing lens is placed away from the high index medium (a). If on the other hand, the lens is part of the high index material (b), or placed in contact with it (immersion lens, c) then the focus beam radius is reduced and better optical resolution is achieved.*

The fact that a converging beam will not focus to a smaller spot if it crosses a dielectric interface can also be deduced from the energy-conservation argument we developed in Chapter 2. Consider a Gedanken experiment in which we have a beam that runs parallel to the original converging beam, but with a small offset such that the two beams have a finite cross energy. Now it becomes obvious that

although the position of the foci can change, their size cannot, because it would lead to a change in the cross energy. This argument does not hold if the dielectric surface is not planar, because then the two beams (the real and the Gedanken beams) are no longer identical and parallel inside the dielectric, so even though the cross energy must still remain unchanged, that does not imply that the beam foci are unchanged.

4.5 Projecting Gaussian Beams

We have seen that the Gaussian-beam theory sets fundamental limits on how tightly we can focus an optical beam. Equally important for design and operation of optical microsystems is the fact that an optical beam cannot be perfectly collimated, but instead will converge to a focus before diverging. Consider the distance from a lens to its focus

$$d_2 = \frac{f}{1 + f^2 / z_{R1}^2} \tag{4.55}$$

This distance has a maximum value of $d_{2\max} = \frac{f}{2} = \frac{\pi \cdot \omega_{lens}^2}{2 \cdot \lambda} = \frac{\pi \cdot \omega_{focus}^2}{\lambda}$ for $f = \frac{\pi \cdot \omega_{lens}^2}{\lambda} = \frac{2\pi \cdot \omega_{focus}^2}{\lambda}$ as demonstrated in Chapter 4.3. This maximum value can be thought of as the collimated distance of the Gaussian beam. It limits the distance between focusing elements for a beam of a given beam radius, and its square dependence on the beam radius represents a challenge for designers of optical microsystems.

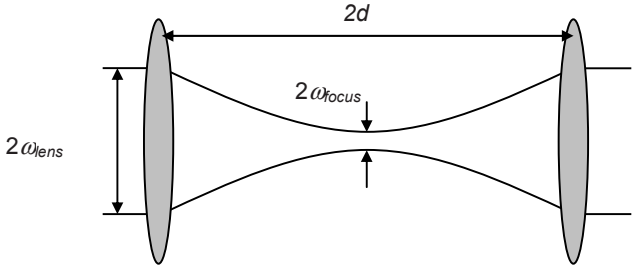


Figure 4.14. A standard problem in optical MEMS design is to design a lens system that maximizes the distance between the lenses for a given lens aperture.

To begin to understand the design tradeoffs, consider the situation shown in Fig. 4.14. The problem here is to design a lens system such that the propagation distance between the lenses is as long as possible given the aperture of the lenses.

This is equivalent to minimizing the lens apertures for a given lens spacing. So we fix the lens spacing and treat the beam radius at the focus as the independent parameter in the optimization.

The beam radius at the lens can be expressed in terms of the beam radius at the focus and the lens separation

$$\omega_{lens} = \sqrt{\omega_{focus}^2 \left[1 + \left(\frac{\lambda \cdot d}{\pi \cdot \omega_{focus}^2} \right)^2 \right]} \tag{4.56}$$

We optimize this expression with respect to variations in the beam radius at the focus

$$\frac{\partial \omega_{lens}}{\partial d} = \frac{2\omega_{focus} \left[1 + \left(\frac{\lambda \cdot d}{\pi \cdot \omega_{focus}^2} \right)^2 \right] + \omega_{focus}^2 \left(\frac{\lambda \cdot d}{\pi} \right)^2 \frac{(-4)}{\omega_{focus}^5}}{(-2)\sqrt{\omega_{focus}^2 \left[1 + \left(\frac{\lambda \cdot d}{\pi \cdot \omega_{focus}^2} \right)^2 \right]}} = 0 \Rightarrow \tag{4.57}$$

$$\omega_{focus} = \sqrt{\frac{\lambda \cdot d}{\pi}} \tag{4.58}$$

$$\omega_{lens} = \sqrt{\frac{2 \cdot \lambda \cdot d}{\pi}} = \sqrt{2} \cdot \omega_{focus} \tag{4.59}$$

This is of course the same result we found by a different method in Chapter 4.3. We will use this result over and over in design of chip-scale fiber optic switches and other optical Microsystems.

4.6 Gaussian Beam "Imaging"

The preceding treatment is quite general because we can apply it to incident Gaussian beams with finite radius of curvature by adjusting the focal length as discussed above. However, it is often useful to have formulas relating beam waists on either side of a lens similar to the ones we use to describe imaging in geometrical optics (Appendix A). Consider the situation depicted in Fig. 4.15, where a diverging Gaussian beam is converted into a converging beam by the lens. This creates a situation in which the beam waist at a distance d_1 in front of the lens is imaged to a waist at a distance d_2 behind the lens.

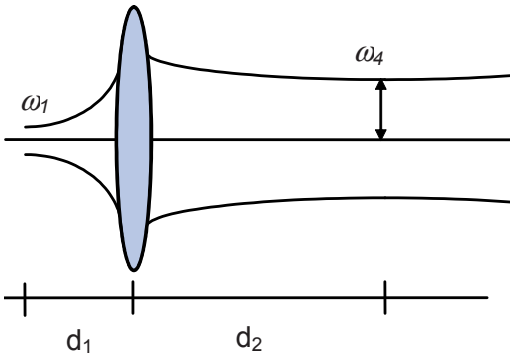


Figure 4.15 Gaussian beam imaging

Adopting the expression we found earlier (Eq. 4.35 derived in the calculation following Fig. 4.5), we can write

$$j\frac{\pi\cdot\omega_4^2}{\lambda}=\frac{\left(1-\frac{d_2}{f}\right)j\frac{\pi\cdot\omega_1^2}{\lambda}+d_1+d_2-\frac{d_1d_2}{f}}{1-j\frac{\pi\cdot\omega_1^2}{f\lambda}-\frac{d_1}{f}} \tag{4.60}$$

Equating imaginary parts of this equation allows us to find the beam radius at the image

$$\omega_2^2=\frac{\omega_1^2}{\left(1-\frac{d_1}{f}\right)^2+\frac{\pi^2\cdot\omega_1^4}{f^2\lambda^2}} \tag{4.61}$$

Similarly, we find the distance d_2 by equating real parts

$$\frac{d_2}{f}=\frac{\frac{d_1}{f}\left(\frac{d_1}{f}-1\right)+\frac{z_R^2}{f^2}}{\left(\frac{d_1}{f}-1\right)^2+\frac{z_R^2}{f^2}} \tag{4.62}$$

This can be rewritten to show the correspondence to the lens law

$$\frac{1}{d_2}+\frac{1}{d_1}\frac{1}{1+\frac{z_R^2}{d_1(d_1-f)}}=\frac{1}{f} \tag{4.63}$$

or in normalized form

$$\frac{z_R}{d_2}+\frac{z_R}{d_1}\frac{1}{1+\dfrac{1}{\left(\dfrac{d_1}{z_R}-\dfrac{f}{z_R}\right)\dfrac{d_1}{z_R}}}=\frac{z_R}{f} \tag{4.64}$$

Here $z_R=\dfrac{\pi\cdot\omega_1^2}{\lambda}$ is the Rayleigh length of the object waist.

We notice these expressions become identical to the familiar one for Geometrical Optics as z_R -> *0*. For the special case that the object waist is exactly one focal length away from the lens we find the following relationships

$$d_1=f\Rightarrow d_2=f \tag{4.65}$$

$$\omega_{x1,y1}=\frac{f\lambda}{\pi\omega_{x2,y2}} \tag{4.66}$$

Notice that the beam radii at the two foci are inversely related. This is what we would expect from Fourier Optics.

4.6.1 Graphical Description of Gaussian Beam "Imaging"

It is instructive to use graphical representations of the formulas we have derived to better understand the nature of Gaussian Beam "imaging". Figure 4.16 shows a plot of the Gaussian beam radius of the "image" as given by the equation $\omega_4^2/\omega_1^2=\left[(1-d_1/f)^2+z_R^2/f^2\right]^{-1}$ derived above.

The solid line in Fig. 4.16 describes the situation when z_R=*0*, which corresponds to the geometrical optics limit. We see that Gaussian beams of a finite Rayleigh length have well defined "images" for all positions of the object. This is not the case for geometrical optics, in which the image disappears when d_1=*f*.

Figure 4.17 shows a plot of the position of the Gaussian "image" as given by the equation $\dfrac{d_2}{f}=\dfrac{(d_1/f)(d_1/f-1)+(z_R/f)^2}{(d_1/f-1)^2+(z_R/f)^2}$. Again we see that Gaussian imaging does not suffer from the singularities predicted by geometrical optics.

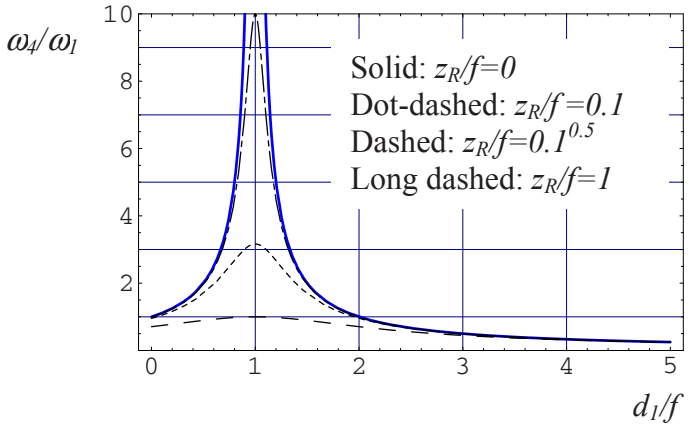


Figure 4.16. Plot of Gaussian beam radius of the image normalized to the beam radius of the object as a function of object-to-lens separation normalized to the focal length (see Fig. 4.9 for definitions).

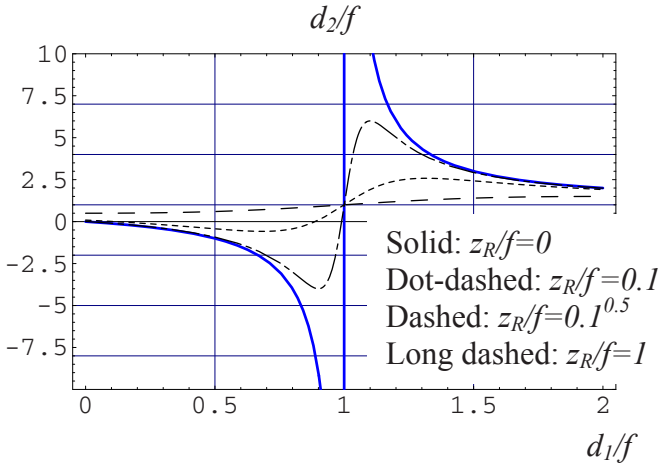


Figure 4.17. Plot of Gaussian image-to-lens distance of the image normalized to the focal length as a function of object-to-lens separation, also normalized to the focal length.

4.7 Truncation of Gaussian Beams

It is clear from the preceding that an ideal Gaussian Beam cannot be realized in practical systems, simply due to the fact that in our mathematical description the Gaussian beam extend to infinity in both dimensions perpendicular to the optical axis. The best we can do in practice is to create an approximation to the Gaussian

Beam. In optical MEMS and integrated optics it is advantageous to keep all apertures as small as possible, so we need to understand how truncation affects Gaussian Beams.

Truncation creates loss in two equally important ways; through direct blocking of the beam by the aperture and by forward scattering into higher-order modes. The forward scattered light also interferes with the light passing the aperture in the Gaussian mode, and this interference changes the beams in ways that are unacceptable for some applications. We will start by calculating the energy loss and then proceed to investigate the effects of diffraction due to finite sized apertures.

4.7.1 Energy Loss Due to Truncation of Gaussian Beams

From the Poynting theorem we know that the energy in a Gaussian beam is

$$W = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\vec{E}\times\vec{H}^{*}\right)\cdot dxdy \tag{4.67}$$

Gaussian beams have electric and magnetic fields that are polarized perpendicular to the direction of propagation, i.e. they are Transversal Electro magnetic (TEM) waves, so as for plane waves, we have the following relationship between fields

$$H_x = \frac{-j}{\omega\cdot\mu_0}\frac{\partial E_y}{\partial z} = \frac{\beta}{\omega\cdot\mu_0}E_y \tag{4.68}$$

Using this relationship, the energy in the Gaussian can be written

$$W = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\beta}{\omega\cdot\mu_0}EE^{*}\cdot dxdy \tag{4.69}$$

This expression allows us to find a simple formula for the energy transfer of a Gaussian Beam that is centered on a rectangular aperture

$$T = \frac{\int_{-d_x/2}^{d_x/2}\int_{-d_y/2}^{d_y/2} e^{-2\frac{x^2+y^2}{\omega^2}}\cdot dxdy}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-2\frac{x^2+y^2}{\omega^2}}\cdot dxdy} \tag{4.70}$$

The solid curve in Fig. 4.18 shows the energy transfer of a Gaussian that is centered on a square aperture. In this plot, the size of the square is normalized to the

beam radius. We see that 99% of the energy in a Gaussian Beam will pass though an aperture with a side that matches the beam radius.

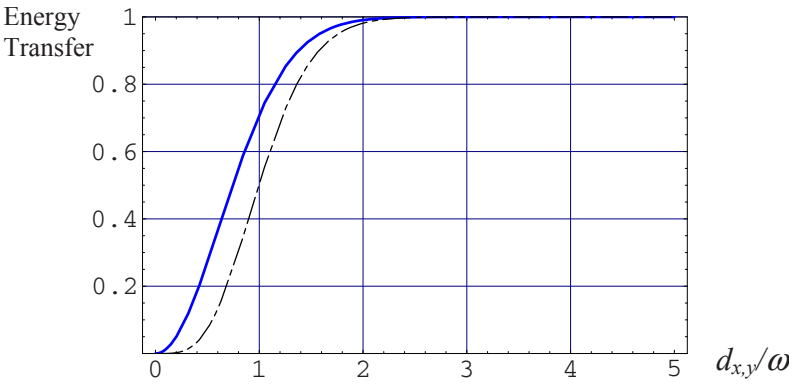


Figure 4.18 *The solid line shows the total energy transfer of a Gaussian beam that has been truncated by a square aperture with a side of $d_{x,y}/\omega$. The dashed line shows the fraction of energy that is left in the fundamental Gaussian mode after truncation by the aperture.*

This energy-transfer calculation doesn't tell the whole story. In many cases we are more interested in how much of the energy is left in the fundamental Gaussian Beam. In other words, we are interested in calculating not only what percentage of the energy is passed by the aperture, but also how much of the passed energy that remains in the fundamental Gaussian mode and how much that is transferred into higher order Gaussian modes. From the point of view of the fundamental Gaussian Beam, both the blocked energy and the transferred energy are lost.

To calculate the total loss of truncation, we express the truncated field in terms of the fundamental and higher-order modes.

$$\vec{E}_{total} = \sum_{n=0}^{\infty} c_n \vec{E}_n \tag{4.71}$$

Again we use the Poynting theorem to express the propagating power in mode n. Keeping in mind the orthogonality of the Gaussian modes, we find the following expression for the expansion coefficients

$$\int_A \left(\vec{E}_{truncated} \times \vec{H}_n^*\right) \cdot dA = \int_A \left(\sum_{n=0}^{\infty} c_n \vec{E}_n \times \vec{H}_n^*\right) \cdot dA = c_n \int_A \left(\vec{E}_n \times \vec{H}_n^*\right) \cdot dA$$

$$\Rightarrow c_n = \frac{\int_A \left(\vec{E}_{truncated} \times \vec{H}_n^*\right) \cdot dA}{\int_A \left(\vec{E}_n \times \vec{H}_n^*\right) \cdot dA} \tag{4.72}$$

Here we have simplified the expression by substituting *dA* for *dx*·*dy*.

We normalize the expansion coefficients to find a transfer function of the truncated field into the Gaussian-mode fields

$$t_n = \frac{\int_A \left(\vec{E}_{truncated} \times \vec{H}_n^*\right) \cdot dA}{\sqrt{\int_A \left(\vec{E}_{truncated} \times \vec{H}_{truncated}^*\right) \cdot dA} \sqrt{\int_A \left(\vec{E}_n \times \vec{H}_n^*\right) \cdot dA}} \tag{4.73}$$

The power transfer is the square of the field transfer

$$T_n = t_n t_n^* = \frac{\left[\int_A \left(\vec{E}_{truncated} \times \vec{H}_n^*\right) \cdot dA\right]^2}{\int_A \left(\vec{E}_{truncated} \times \vec{H}_{truncated}^*\right) \cdot dA \cdot \int_A \left(\vec{E}_n \times \vec{H}_n^*\right) \cdot dA} \tag{4.74}$$

Again we use the fact that Gaussian Beams are TEM waves to simplify the expressions

$$t_n = \frac{\int_A \left(E_{truncatedl} E_n^*\right) \cdot dA}{\sqrt{\int_A \left(E_{truncated} E_{truncated}^*\right) \cdot dA} \sqrt{\int_A \left(E_n \vec{E}_n^*\right) \cdot dA}} \tag{4.75}$$

$$T_n = t_n t_n^* = \frac{\left[\int_A \left(E_{truncated} E_n^*\right) \cdot dA\right]^2}{\int_A \left(E_{truncated} E_{truncated}^*\right) \cdot dA \cdot \int_A \left(E_n E_n^*\right) \cdot dA} \tag{4.76}$$

Comparing this to the total energy transfer through an aperture, we find that the transfer coefficient into the fundamental mode is simply the square of the total-energy transfer coefficient. This is shown in the dashed line in Fig. 4.18.

In optical system design we are typically more concerned with the energy that is left in the Fundamental Gaussian than in the total energy, so the dashed line of Fig. 4.18 is the more significant. It shows that if we chose the aperture size equal to twice the beam radius, then we have that 98% of the energy is left in the Fundamental Gaussian Beam, 1% is scattered into higher order modes, and 1% is blocked by the aperture. The blocked light might be reflected or absorbed.

The results presented in Fig. 4.18 are important in themselves, but even more significant are the concepts of projections onto Gaussians and over-lap integrals that are introduced in the derivation of Fig. 4.18. We will use the type of projection demonstrated in this calculation over and over again to solve problems of light propagation through complex systems, e.g. fiber couplers, optical scanners, and fiber-optic switches that are the subjects of the next several chapters. To readers who are unfamiliar with these concepts, it is therefore well worth the effort to carefully study this calculation and familiarize themselves with all the steps of the derivation.

4.7.2 Far-field of Truncated Gaussian Beams – Frauenhofer Diffraction

The forward-scattered light from an aperture interferes with light in the fundamental Gaussian and will in many cases create field variations that are detrimental to system operation. These types of effects are particularly difficult in systems with multiple apertures. To understand forward scattering and quantify its effect on system performance, we must develop models for diffraction of beams after truncation by apertures.

The truncated Gaussian will propagate (diffract) very differently from a complete Gaussian. To find the profile of the truncated Gaussian after it has propagated to a plane (the output plane) a certain distance away from the aperture, we can in principle write the truncated Gaussian as a sum of the fundamental and higher order Gaussians. We then use the simple laws of Gaussian Beam propagation to find

the shape and phase of each of the components of the sum at the output plane. Finally the components are re-summed to find the resulting beam profile at the output plane. This procedure is, however, impractical because of the large number of elements we have to sum to get a good approximation to a truncated Gaussian. Instead the preferred method for such diffraction calculations is to use the Huygens-Fresnel diffraction integral, typically in the Fresnel approximation [2].

The truncated Gaussian is rather complex both in the near field close to the diffracting aperture and in the far field. Even moderate truncation leads to significant modulation of the Gaussian profile. The modulation of the profile, including the on-axis intensity, varies along the propagation path in complex patterns that are characteristic of the aperture. Due to the importance of truncated Gaussian Beams in laser technology, these effects have been thoroughly studied and documented [3].

The complicated near-field effects of truncated Gaussians described by Fresnel diffraction are for several reasons of limited interest in optical microsystems. Many such systems, including fiber switches and confocal microscopes, use spatial mode filters (pin holes or single-mode optical fibers) that reject all but the fundamental Gaussian of interest. In systems where forward scattered light is important we can most often use the simpler Frauenhofer diffraction theory to model the relevant effects with sufficient accuracy.

The standard criterion for validity of Frauenhofer diffraction is

$$z > \frac{2D^2}{\lambda} \tag{4.77}$$

where D is the linear extent of the aperture, and λ is the wavelength. Optical microsystems that operate at visible and near-infrared wavelengths with well-confined beams and apertures less than 100 um typically meet this criterion. Systems with lager apertures, e.g. laser scanners, are almost always designed with Fourier lenses that bring the far-field to their focal plane, so that the conditions for Frauenhofer diffraction are automatically fulfilled.

In Fraunhofer Diffraction Theory the field distribution in a plane perpendicular to the optical maxis is described in terms of the field in an input plane a distance z away [2]

$$E(x_{out}, y_{out}) = \frac{e^{jkz} \cdot e^{j\frac{k}{2z}\left(x_{out}^2 + y_{out}^2\right)}}{j\lambda z} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} E(x_{in}, y_{in}) \cdot e^{-j\frac{2\pi}{\lambda z}(x_{out}x_{in} + y_{out}y_{in})} dx_{in} dy_{in} \tag{4.78}$$

Comparing this expression to the Fourier transform defined as

$$F(g)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\cdot e^{-j2\pi\left(x\cdot f_x+y\cdot f_y\right)}dxdy \tag{4.79}$$

we see that the output field is, to within a multiplicative factor, the Fourier Transform of the input field evaluated at

$$f_x=\frac{x_{out}}{\lambda z} \tag{4.80}$$

$$f_y=\frac{y_{out}}{\lambda z} \tag{4.81}$$

It is sometimes more convenient to express the diffracted field in terms of angular coordinates

$$E\left(\theta_x,\theta_y\right)=\frac{e^{jkz\left(1+\frac{\theta_x^2+\theta_y^2}{2}\right)}}{j\lambda z}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}E\left(x_{in},y_{in}\right)\cdot e^{-j\frac{2\pi}{\lambda}\left(\theta_x x_{in}+\theta_y y_{in}\right)}dx_{in}dy_{in} \tag{4.82}$$

Using the standard formula for the Fourier transform of a Gaussian distribution

$$F\left(e^{-\frac{x^2}{\omega_x^2}-\frac{y^2}{\omega_y^2}}\right)=\pi\omega_x\omega_y e^{-\pi^2\left(\omega_x^2\cdot f_x^2+\omega_y^2\cdot f_y^2\right)} \tag{4.83}$$

we find the following expression for the far field of a Gaussian Beam

$$E_{Gauss}\left(x_{out},y_{out}\right)=\frac{e^{jkz}\cdot e^{j\frac{k}{2z}\left(x_{out}^2+y_{out}^2\right)}}{j\lambda z}F\left\{E(x,y,0)=e^{-\frac{x^2+y^2}{\omega_0^2}}\right\}=$$
$$\frac{e^{jkz}\cdot e^{j\frac{k}{2z}\left(x_{out}^2+y_{out}^2\right)}}{j\lambda z}\pi\omega_0^2 e^{-\frac{\pi^2\omega_0^2}{\lambda^2 z^2}\left(x^2+y^2\right)} \tag{4.84}$$

In angular coordinates this becomes

$$E_{Gauss}\left(\theta_x,\theta_y\right)=e^{-j\frac{\pi}{2}}e^{jkz\left(1+\frac{\theta_x^2+\theta_y^2}{2}\right)}\frac{\omega_0/z}{\theta_{diff}}e^{-\frac{\theta_x^2+\theta_y^2}{\theta_{diff}^2}} \tag{4.85}$$

These formulas for the Gaussian in the far field can be verified by direct substitution of the far-field expressions of the beam radius, radius of curvature, and Gouy phase into the Gaussian-Beam equation.

To understand the effects of truncation on Gaussian Beams we use the Fraunhofer Diffraction integral, $E(\theta_x, \theta_y)$ to plot the far-field intensity profile (keeping in mind that intensity is proportional to the square of the field) of a Gaussian beam that passes through an aperture at its waist. The results are shown in Fig. 4.19. The Fraunhofer Diffraction integral is completely separable in rectangular coordinates if the aperture is rectangular with its sides along the principal coordinate axes. This means that for rectangular apertures the relative field strength along one coordinate axis is not influenced by the size of the aperture along the orthogonal coordinate. We therefore only plot the far field along a single axis with varying degrees of truncation along this same axis.

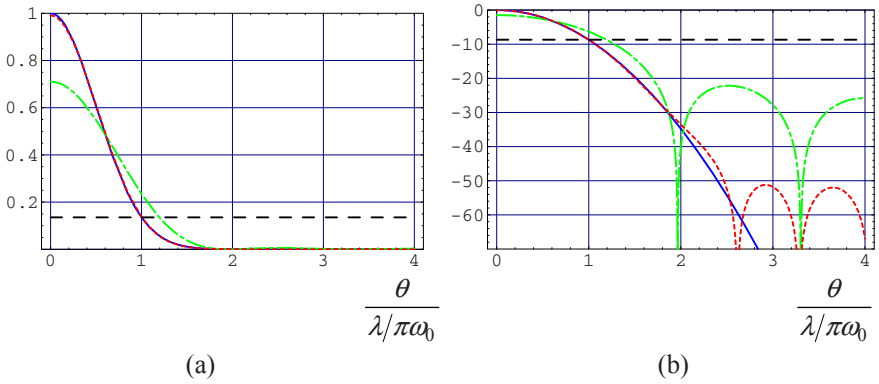


Figure 4.19 *Far-field (angular) profiles of Gaussians truncated in one dimension plotted against angle normalized to the Gaussian far-field angle. The graphs in (a) are on a linear scale, while (b) shows the profile on a dB scale (10 Lg). The solid lines show the complete Gaussian with no truncation, the dashed lines show a beam truncated by an aperture with a width equal to four times the beam radius ($d=4\omega_0$), and the dot-dashed line represent an aperture width equal to twice the beam radius ($d=2\omega_0$).*

Three different aperture sizes are chosen: infinite, twice the beam radius, and four times the beam radius. The resulting beam profiles are plotted as a function of angle normalized to the Gaussian-Beam far-field angle, $\lambda/\pi\omega_0$. The graphs in Fig. 4.19a are on a linear scale to emphasize variation close to the optical axis where the intensity is high, while in Fig. 4.19b they are on a dB scale to show the nulls and side lobes of the truncated profiles.

The graphs show that the effects of truncation depend strongly on aperture size in the range from twice the beam radius to four times the beam radius. The beam that is truncated at its 1/e field ($1/e^2$ intensity) radius has significantly lower on-axis intensity and a wider central lobe than the unobstructed Gaussian. It also has side lobes with peak intensity close to 1% of the on-axis intensity. This means

that the field strength in the side lobes is at the 10% level, which leads to interferences, both in the near field and far field, that are unacceptable in many systems.

The beam that passes through an aperture that is four times larger than the beam radius is indistinguishable from the complete Gaussian Beam when viewed on a linear scale. Side lobes can be observed in the dB-scale graph, but they are at the 10^{-5} intensity level. This corresponds to field strengths on the order of $3 \cdot 10^{-3}$, which is too low to create significant modulation through interference with the fundamental Gaussian. This size aperture is therefore acceptable for all but the most stringent systems.

The graphs of Fig. 4.19 tells us that only systems with relatively low requirements on contrast can tolerate apertures that are equal to or smaller than twice the Gaussian Beam radius. The good news is that the amount of stray light is a strong function of aperture size[a], so we don't have to increase the apertures much to reduce stray-light interference to tolerable levels. At four times the beam radius, the aperture only weakly distorts the Gaussian Beam. This size aperture does not significantly widen or weaken the central lobes, and it creates only in significant side lobes. Most systems in which miniaturization is important are therefore designed with apertures that are between two and four times the beam radius. A good rule of thumb for many designs is to make the apertures three times the beam radius.

4.8 Summary of Gaussian Beams

This chapter extends the plane wave picture developed in the first two chapters by introducing diffraction. The essence of diffraction is that all electro-magnetic beams will converge to a focus of finite size and then diverge. The Fundamental Gaussian Beam can be considered the "best" possible approximation to the geometrical-optics concept of a light ray in the sense that it has the smallest possible product of beam size at the focus and angular spread in the far field.

The mathematical description of Gaussian-Beam propagation can be summarized by the expression

$$E(r,z) = H_l\left(\frac{\sqrt{2}x}{\omega}\right) H_m\left(\frac{\sqrt{2}y}{\omega}\right) \frac{\omega_0}{\omega(z)} \exp\left[-jkz + j\phi(z) - \frac{r^2}{\omega^2(z)} - jk\frac{r^2}{2R(z)}\right] \tag{4.86}$$

[a] In later chapters we will use Gaussian Beams to model electromagnetic waves that in reality do not have quite the same strong exponential dependence on distance from beam center. In such systems we must be careful when applying the rule-of-thumb that has been stated here.

where $H_l(x)$ is the Hermite-Gaussian mode of order *l*, ω_0 is the beam radius at focus, $\omega(z)$ is the beam radius, $\phi(z)$ is the Gouy phase shift, and $R(z)$ is the radius of curvature. The Fundamental Gaussian Beam is of order $l=0$, $m=0$, and $H_0(x)=1$.

The *z*-dependence of the characteristic parameters are described in the following equations

$$\omega(z)=\omega_0\sqrt{1+\left(\frac{z}{z_R}\right)^2} \tag{4.87}$$

$$R(z)=z+\frac{z_R^2}{z} \tag{4.88}$$

$$\phi(z)=(1+l+m)\tan^{-1}\left(\frac{z}{z_R}\right) \tag{4.89}$$

where we for convenience have introduced the Rayleigh Range that is defined as

$$z_R=\frac{\pi\omega_0^2}{\lambda} \tag{4.90}$$

Based on these formulas we find the far-field diffraction angle for the Gaussian beam

$$\theta=\lim_{z\to\infty}\frac{\omega(z)}{z}=\frac{\lambda}{\pi\cdot\omega_0} \tag{4.91}$$

Gaussian beams are solution to the paraxial wave equation, and give an intuitive, and largely correct picture of diffraction. The usefulness of Gaussian-beam theory stems from its mathematical simplicity and the fact that many practical lasers and waveguide devices produce optical fields that to a very high degree of accuracy can be modeled as Gaussian beams. Gaussian modes constitute a complete set of basis function, so they provide means to solve general diffraction problems. This is, however, not always a practical approach.

Using the propagation law for Gaussian beams and the lens law, we can calculate the effect of lenses on the propagation of Gaussian beams, and we find that Gaussian-beam theory corrects several erroneous results of geometrical optics. In particular, Gaussian-beam theory predicts that:

1. There is no such thing as a collimated beam. The best we can do is to create a Gaussian beam with a soft focus and a long Rayleigh range.
2. A focused spot has a finite beam radius, and the focus can at most be half a Rayleigh length in front of the focusing lens.

Figure 4.20 shows a comparison of Gaussian and geometrical optics. Geometrical optics makes several erroneous predictions that are corrected by Gaussian beam theory. Geometrical optics introduces the concept of a collimated beam. We have seen that this violates energy conservation, so it is gratifying that Gaussian-beam theory does not support this concept. Instead we see that a lens placed exactly one focal length away from a beam waist, produces another beam waist exactly one focal length away on the other side of the lens. The beam radius at the focus is inversely related to the beam radius at the original waist, so a rapidly diverging Gaussian beam (i.e. one that originates from a waist with a small beam radius), will produce a large beam radius at the focus. This creates a beam with a soft focus and a long Rayleigh length, i.e. the resulting beam is an approximation of a collimated beam.

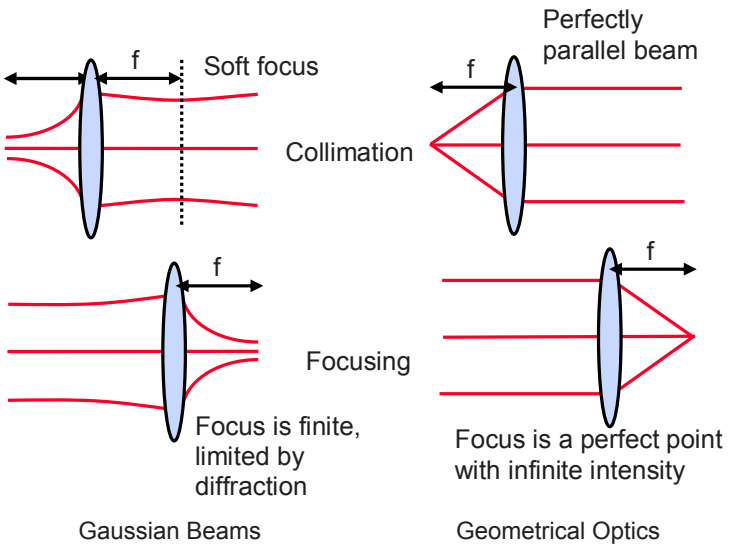


Figure 4.20 *Comparison of Gaussian and geometrical optics. Gaussian Beams are the "best" approximations to Geometrical-Optics rays, but the predictions of Gaussian-Beam theory differ from those of Geometrical optics in important ways as shown.*

Geometrical optics predicts that a perfect lens will focus light to a point, and that this focus can be obtained at any distance from the lens (by adjusting the lens-to-object distance). Gaussian-beam theory corrects this by showing that the focus

can be no further away than half the Rayleigh length of the incident beam, and that the beam radius is finite at the focus.

Gaussian-Beam theory predicts that the waist radius created by a lens in the limit of short focal lengths is given by $\frac{2f\lambda}{\pi \cdot \omega_{lens}} \approx 1.9 \cdot \frac{f}{D}\lambda$, in good agreement with the classical expression for diffraction-limited spot size $\left(2.2 \cdot \frac{f}{D}\lambda\right)$. These expressions show that the spot size is directly proportional to wavelength. The shorter wavelength in high-index materials can, however, only be used to realize a smaller spot-size if the focusing takes place in the high-index material. Gaussian-beam theory also sets fundamental limits on the distance from a lens to its focus ($d_{2\max} = \frac{f}{2} = \frac{\pi \cdot \omega_{focus}^2}{\lambda}$ with $\omega_{lens} = \sqrt{2} \cdot \omega_{focus}$) with important consequences for miniaturization of optics.

A lens of focal length f images a beam waist ω_1 at a distance d_1 in front of the lens to a waist ω_2 at a distance d_2 behind the lens:

$$\omega_2^2 = \frac{\omega_1^2}{\left(1 - \frac{d_1}{f}\right)^2 + \frac{\pi^2 \cdot \omega_1^4}{f^2 \lambda^2}} \tag{4.92}$$

$$\frac{1}{d_2} + \frac{1}{d_1} \frac{1}{1 + \frac{z_R^2}{d_1(d_1 - f)}} = \frac{1}{f} \tag{4.93}$$

For the special case $d_1=f$ we find $d_2=f$ and $\omega_{x1,y1} = \frac{f\lambda}{\pi\omega_{x2,y2}}$.

In practical systems all Gaussian Beams are truncated by finite apertures. The effects of truncation are first to remove energy by through the blocking of the finite aperture and second to scatter energy into higher order modes by the mode-shape change cause by truncation. A much-used rule of thumb is to use apertures that are at least three times the beam radius to avoid significant effect of truncation.

Further Reading

H. Kogelnik, T. Li, “Laser Beams and Resonators”, Applied Optics, vol. 5, no. 10, October 1966, pp. 1550-1567.

B.E.A. Saleh, M.C. Teich, "Fundamentals of Photonics", 2nd edition, Wiley, 2007.

A. Yariv, P. Yeh, "Photonics: Optical Electronics in Modern Communications", 6th edition, Oxford University Press, 2007.

Exercises

Problem 4.1 - Projecting Gaussian beams

Consider a Nd:YAG laser at *1.06 um* wavelength that emits a laser beam that after collimation has a beam radius of *1 m* at its waist, which coincides with a projection lens.

a) What is the longest distance from the lens that you can form another beam waist?
b) What is the corresponding focal length of the lens?

Problem 4.2

You have a 1 mW HeNe laser at *630 nm* wavelength, and you can make lenses that accommodate Gaussian beams with a beam radius of *0.3 m*. Using these components, design an optical system that delivers the maximum intensity at a distance of *1,000 km*.

Problem 4.3

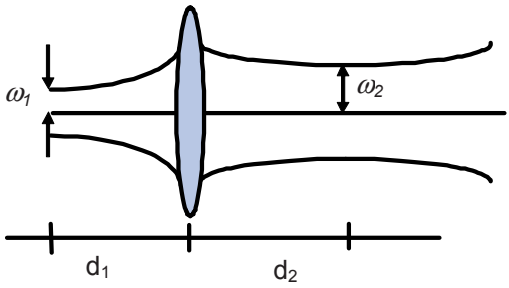
a) What is the minimum radius of curvature for a Gaussian beam, and where does it occur? Express your answer in terms of Rayleigh length of the beam.

You have a laser that has a Gaussian beam output at a wavelength of *1 um*. The waist of the Gaussian beam is at the laser output mirror and has a beam radius of ω_0*=100 um*. You want to use a lens to image the beam waist at a distance *5 m* from the lens.

b) As a practical matter, we want to use a lens with the smallest possible diameter. What should be the focal length of the lens?
c) What is the beams radius of the imaged waist *5 m* from the lens?
d) What focal length would you find if you were using ray-optics to model the imaging? Comment on how well the Gaussian-beam model and the Ray model agree.

Problem 4.4 – Imaging Gaussian beams

Consider the Gaussian beam "imaging" set up shown below.

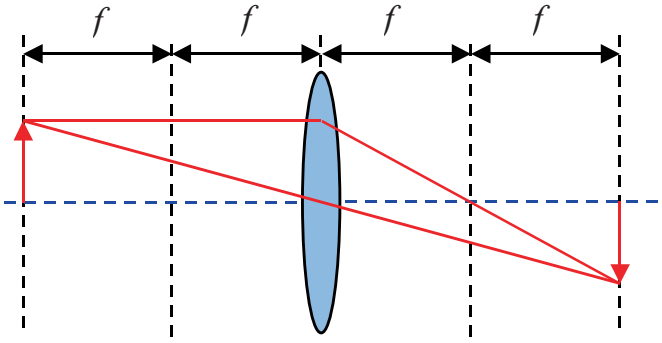


"Imaging" of Gaussian beam waist

a) What is the longest possible distance, d_2, from the lens to the "image". Express your answer in terms of the beam radius at the lens.
b) Under what conditions can a Gaussian beam waist be "imaged" by a lens that is placed less than its focal length away from the waist?
c) Explain physically how this can happen.

Problem 4.5 - Gaussian beams vs. geometrical optics

Geometrical optics (ray tracing) predicts that an object placed *2f* from a positive lens will be imaged to a position *2f* behind the lens with a unity magnification as shown in the figure below.



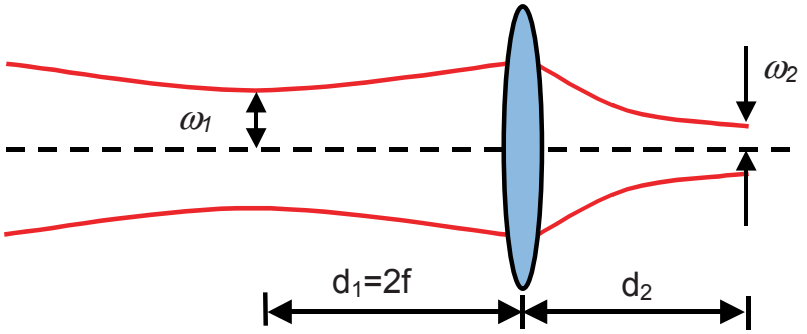
Geometrical optics model of imaging of an object 2f in front of a positive lens. The image appears 2f behind the lens, and the magnification is unity.

In Gaussian optics, this picture is quite different as shown below. According to our formulas for Gaussian beam "imaging", a beam waist located *2f* in front of the lens will be "imaged" at a distance (d_2) behind the lens that is given by:

$\frac{1}{d_2}+\frac{1}{d_1}\frac{1}{1+\frac{\pi^2\cdot\omega_1^4}{2f^2\cdot\lambda^2}}=\frac{1}{f}$. The corresponding image size is: $\omega_2^2=\frac{\omega_1^2}{1+\frac{\pi^2\cdot\omega_1^4}{f^2\lambda^2}}$.

In the no-diffraction limit (λ->*0*), we find that d_2=f and ω_2->0.

Explain this discrepancy between the geometrical-optics picture and the Gaussian picture.

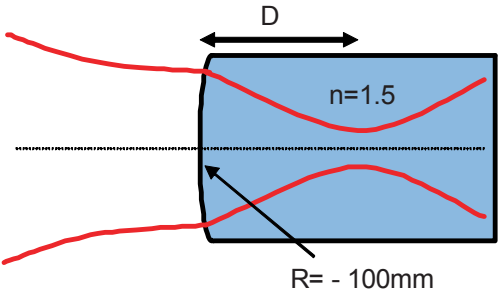


Lens transformation of Gaussian-beams.

Problem 4.7 – Immersion optics

Consider a Gaussian beam with a beam radius of *1 mm* incident on a block of material of index *n=1.5* with a spherical front surface with a radius of *-100 mm*. The incident beam has a flat phase front at the interface, and the wavelength is *1.06 um*. The ABCD matrix of the spherical surface is

$$\begin{vmatrix} 1 & 0 \\ \frac{n-1}{n}\frac{1}{R} & \frac{1}{n} \end{vmatrix}$$



a) What is the beam radius at the waist?
b) What is the distance, *D*, from the spherical surface to the waist?

Problem 4.8 – Gaussian beam overlap

a) Show by detailed calculations (not energy methods) that two co-linear Gaussian beams of different waist size and position maintain their overlap integral as they propagate (i.e. their overlap integral in a plane perpendicular to the optical axis is independent of axial position).
b) Do the same for two crossing Gaussian beams.

References

1. For a derivation and discussion of the Airy's disk pattern, see for example: E. Hecht, "Optics", 3rd edition, Addison-Wesley, 1998.
2 . J.W. Goodman, "Introduction to Fourier Optics, 2nd edition", McGraw-Hill, 1996.
3 For a general discussion of some of the near-field effects of Gaussian-Beam truncation, see Chapter 18 of A.E. Siegman, "Lasers", University Science, Mill Valley, CA 1989.

5: Optical Fibers and Waveguides

5.1 Introduction to Fibers and Waveguides

A very large fraction of all micro and nano-photonics systems are designed to interact with optical waveguides and optical fibers. In this chapter we described analytical concepts and computational tools for the study of guided-wave propagation. In the next chapter we will use these tools for analysis and design of guided-wave optical devices. We start the discussion with a complete derivation of the modes on a slab waveguide. The modes on slab waveguides are only confined in one dimension, while most practical waveguides confine the light in two dimensions orthogonally to the direction of propagation. Nevertheless, slab-waveguide modes demonstrate many of the most important features of guided-wave optics and are good conceptual models for developing intuition about fiber optics.

We then extend the slab-waveguide treatment to optical fibers of cylindrical symmetry, and demonstrate the existence of single-mode fibers. A key finding is that we can develop a Gaussian approximation to the mode profile of the fundamental mode on step-index optical fibers. This model allow us to use many of the tools we developed in Chapter 4 on Gaussian Beams to design and analyze coupling between optical fibers and integrated optics.

In the last part of the chapter we investigate dispersion, i.e. wavelength dependence, of the effective index on optical waveguides and fibers. We consider material dispersion, waveguide dispersion, and modal dispersion, and investigate how these effects influence the propagation of pulses on waveguides and fibers. Again we use a Gaussian approximation, in this case in the time domain as opposed to the spatial domain, to find close-form expressions for pulse broadening and chirp of electromagnetic pulses propagating in the presence of dispersion.

5.2 Geometrical optics description of waveguides

The concept of Total Internal Reflection (TIR) that we have investigated in some detail in Chapter 3, allows us to make a first-order description of dielectric optical waveguides. Consider the symmetric planar slab waveguide shown in Fig. 5.1.

For a guided ray, the incident angle on the cladding must not exceed the critical angle for TIR

$$\sin\theta > \sin\theta_{crit} = n_{clad}/n_{core} \tag{5.1}$$

This means that the maximum incident angle on the facet of the waveguide is

$$\begin{aligned} &\sin\theta_{inc} = NA = n_{core}\sin\left(\frac{\pi}{2} - \theta_{crit}\right) = n_{core}\cos(\theta_{crit}) \\ &\Rightarrow NA = n_{core}\sqrt{1-\sin^2(\theta_{crit})} = \sqrt{n_{core}^2 - n_{clad}^2} \end{aligned} \tag{5.2}$$

The Numerical Aperture (NA), which is the sine of the maximum acceptance angle, is an important parameter for a waveguide.

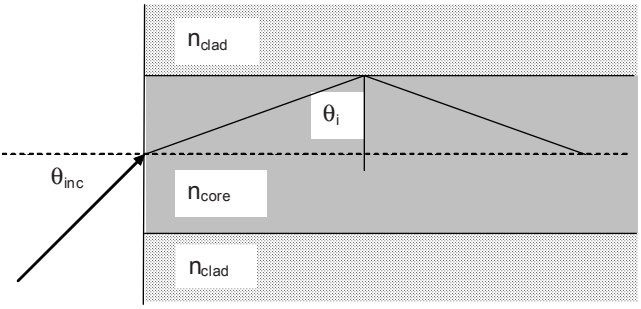


Figure 5.1. For the optical ray to be confined to the core of the slab waveguide, the internal angle, θ_i, must not exceed the critical angle for TIR.

The TIR picture is of limited use in detailed calculations of fiber transmission characteristics. As shown in Fig. 5.1 it is a geometric-optics description that fails to take into account the inevitable spreading of the light "rays" as they propagate on the fiber. We need more sophisticated models for detailed analysis and design.

The TIR model is useful in that it allows us to predict the existence of optical waveguides, and speculate about their advantages. We can imagine making optical waveguides using metallic reflectors, much like we make co-axial and other metallic waveguides for RF-communication. The dielectric waveguide does however have several significant advantages:

1. Much lower loss (metals absorb at optical wavelengths)
2. Simpler and less expensive fabrication, particularly for single-mode waveguides.
3. Better environment for waveguide devices.
4. Lower modal dispersion.

In the following discussion we will concentrate on the dielectric slab waveguide. It is a simple device that allows us to develop an understanding of the important

aspects of waveguide characteristics without being too mathematically complex, and it forms the basis of some important approximate methods for analyzing more complex (and realistic) waveguide structures.

5.3 Three-layered Slab Waveguide

Consider the dielectric stack in Fig. 5.2. We will first investigate TE polarized solutions, i.e. solutions that have their electric fields along the y-direction. These has to fulfill the scalar wave equation

$$\nabla^2 E_y(x,z) + k_0^2 n_i^2 E_y(x,z) = 0 \tag{5.3}$$

in each of the three regions of the waveguide. We are interested in solutions with amplitudes that are independent of z, i.e. solutions of the form

$$E_y(x,z) = E_y(x) e^{-j\beta z} \tag{5.4}$$

where ***β is the longitudinal wave vector***. The term $E_y(x)$ is the mode profile of the waveguide. The equation expresses the fact that we are searching for profiles that propagate without changing.

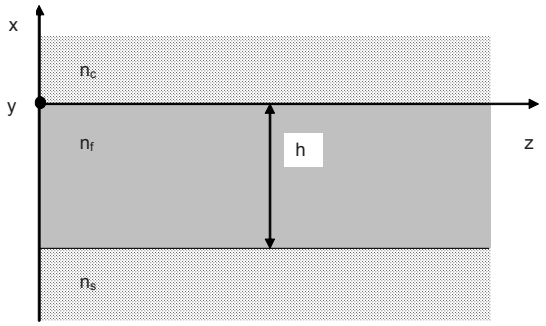


Figure 5.2. The slab waveguide shown in this figure consist of a substrate, a thin dielectric film, and a cladding layer. In our analysis we assume that the substrate and cladding are both infinitely thick, while the film (core) has a thickness h.

Substituting this description of a waveguide mode (Eq. 5.4) into the wave equation (Eq. 5.3), we find

$$\frac{\partial^2}{\partial x^2} E_y(x) + \left(k_0^2 n_i^2 - \beta^2\right) E_y(x) = 0 \tag{5.5}$$

Depending on the values of the wave vector, this transverse wave equation has solutions with exponential x-dependence

$$E_y(x) = E_{y0} e^{\pm\sqrt{\beta^2 - k_0^2 n_i^2}\cdot x} = E_{y0} e^{\pm\gamma x} \quad \text{for} \quad \beta > k_0 n_i \tag{5.6}$$

and oscillatory x-dependence

$$E_y(x) = E_{y0} e^{\pm j\sqrt{k_0^2 n_i^2 - \beta^2}\cdot x} = E_{y0} e^{\pm j\kappa x} \quad \text{for} \quad \beta < k_0 n_i \tag{5.7}$$

The parameter $\kappa = k_0^2 n_i^2 - \beta^2$ is called the ***transversal wave vector***, and $\gamma = \beta^2 - k_0^2 n_i^2$ is the ***attenuation coefficient***. Once we have found one wave vector (i.e. either the longitudinal wave vector, the transversal wave vector, or the attenuation coefficient) in any region for an optical mode, it is trivial to find the others.

We will now postulate that this structure can support a guided mode of the form

$$E_y(x) = \begin{cases} Ae^{-\gamma_c x} & 0 < x \\ B\cos(\kappa_f x) + C\sin(\kappa_f x) & -\mathrm{h} < x < 0 \\ De^{\gamma_s (x+h)} & x < -h \end{cases} \tag{5.8}$$

Notice that we have fields that are decaying in the direction away from the core of the waveguide. At *x=0*, the boundary conditions require that

$$Ae^{-\gamma_c \cdot 0} = B\cos(\kappa_f \cdot 0) + C\sin(\kappa_f \cdot 0) \Rightarrow B = A \tag{5.9}$$

Now we could proceed by writing the equations for the magnetic fields and applying the appropriate boundary, but instead we note that

$$\nabla \times \vec{E} = -\mu j\omega \vec{H} \Rightarrow \vec{z}\left(\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y}\right) = -\mu j\omega H_z \Rightarrow H_z = \frac{j}{\mu\omega}\frac{\partial E_y}{\partial x} \tag{5.10}$$

At *x=0*, this becomes

$$-A\gamma_c e^{-\gamma_c \cdot 0} = -B\kappa_f \cos(\kappa_f \cdot 0) + C\kappa_f \sin(\kappa_f \cdot 0) \Rightarrow C = A\frac{\gamma_c}{\kappa_f} \tag{5.11}$$

Using these two expressions for *B* and *C*, we can write the condition for continuity of the electric field at *x=-h* in the following way

$$A\cos\left(-\kappa_f \cdot h\right) - A\frac{\gamma_c}{\kappa_f}\sin\left(-\kappa_f \cdot h\right) = De^{-\gamma_s \cdot 0}$$
$$\Rightarrow D = A\left[\cos\left(\kappa_f \cdot h\right) + \frac{\gamma_c}{\kappa_f}\sin\left(\kappa_f \cdot h\right)\right] \tag{5.12}$$

The total field can then be written

$$E_y(x) = \begin{cases} A\,e^{-\gamma_c x} & 0 < x \\ A\left[\cos\left(\kappa_f x\right) - \frac{\gamma_c}{\kappa_f}\sin\left(\kappa_f x\right)\right] & -\mathrm{h} < x < 0 \\ A\left[\cos\left(\kappa_f h\right) + \frac{\gamma_c}{\kappa_f}\sin\left(\kappa_f h\right)\right]e^{\gamma_s(x+h)} & x < -h \end{cases} \tag{5.13}$$

Now we apply the final boundary condition, which says that $\frac{\partial E_y}{\partial x}$ must be continuous at $x=-h$, to arrive at the eigenvalue equation for the longitudinal wave vector

$$\left.\frac{\partial E_y}{\partial x}\right|_{x=-h} =$$
$$A\left[\cos\left(\kappa_f \cdot h\right) - \gamma_c \sin\left(\kappa_f \cdot h\right)\right] = A\left[\cos\left(\kappa_f \cdot h\right) + \frac{\gamma_c}{\kappa_f}\sin\left(\kappa_f \cdot h\right)\right]\gamma_s \tag{5.14}$$
$$\Rightarrow \tan\left(h\kappa_f\right) = \frac{\gamma_c + \gamma_s}{\kappa_f\left(1 - \frac{\gamma_c\gamma_s}{\kappa_f^2}\right)}$$

This transcendental equation, and the corresponding one for TM polarized waves

$$\tan\left(h\kappa_f\right) = \frac{\kappa_f\left(\frac{n_f^2}{n_s^2}\gamma_s + \frac{n_f^2}{n_c^2}\gamma_c\right)}{\kappa_f^2 - \frac{n_f^4}{n_s^2 n_c^2}\gamma_s\gamma_s} \tag{5.15}$$

must be solved numerically or graphically to find the eigenvalues for β_{TE} and β_{TM}. These eigenvalues define the guided modes of the slab waveguide. Once they are known, we can find the field distribution of the modes. Notice that these two

equations give us the values of the phase and group velocities of the guided modes. We will investigate this aspect more carefully when we consider waveguide dispersion in the next chapter.

5.3.1 Numerical Solutions to Eigenvalue Equations

The right and left-hand side of the eigenvalue equation for the longitudinal wave vector for TE polarized guided waves are shown in Fig. 5.3. The waveguide parameters are: Thickness: *h=5 micron*, Cladding index: $n_c = 1.4$, Film index: $n_f = 1.5$, Substrate index: $n_s = 1.45$, Wavelength: $\lambda=1\ \mu m$

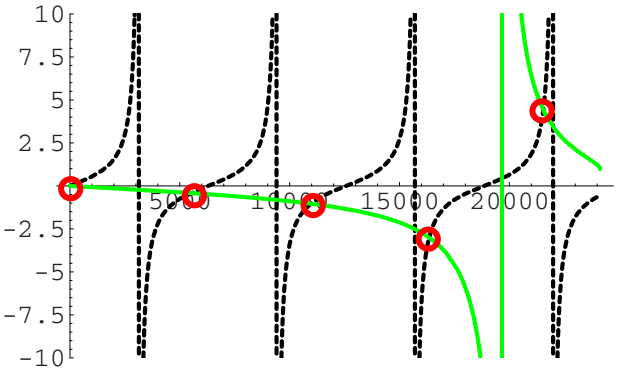


Figure 5.3. *This graphical solution is created by plotting the two sides of the Eigenvalue equation for TE polarized modes as functions of the transversal wave vector. Five candidate solutions (not all of them of practical interest) are indicated by red circles. Waveguide parameters: h=5 um, n_c=1.4, n_f=1.5, n_s=1.45*

From the figure we see that there are five possible solutions to the Eigenvalue equation. Notice that the thickness of the waveguide only influences the green curves, which represent the right hand side of the equation. We can therefore change the number of solutions by simply changing the waveguide thickness.

Let's start by investigating the solution at κ=0. In this case we have a constant E-field in the film. Combined with the boundary condition requiring that the derivative of the E-field be zero at the film-cladding interface, this leads to the conclusion that only a trivial solution (all field components are zero) can exist for κ=0. The other four solutions are all non-trivial. There is no significant difference between the solutions in the upper and lower halves of the plane.

5.3.2 TM Solutions

The TM solutions are found by graphically solving the eigenvalue equation for the TM longitudinal wave vector. The nature of these solutions is very similar to the TE solutions as can be seen from Fig. 5.4, in which the TE and TM graphs are compared.

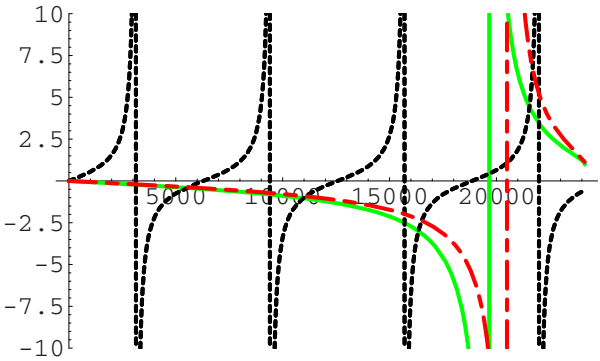


Figure 5.4. *Comparison of the graphical solutions for the TE (solid) and TM (dot-dashed) guided waves. Waveguide parameters: h=5 micron, n_c=1.4, n_f=1.5, n_s=1.45*

5.3.3 Nature of the Solutions

One way to visualize the optical guided modes we have just derived is to consider each mode as consisting of two plane waves with the same longitudinal wave vector and opposite transversal wave vectors of equal magnitude. The two plane waves interfere to create the transversal field distribution of the mode. The decaying part of the mode corresponds to the evanescent fields of Total Internal Reflection. This plane wave picture of optical waveguides is illustrated in Fig. 5.5.

For the interference pattern to not change under propagation, we must have

$$2kn_f h\cos\theta - \Phi_c - \Phi_s = 2\pi \cdot m \tag{5.16}$$

where m is an integer, and Φ_c and Φ_s are the phase shifts of TIR at the cladding and substrate respectively. This expression relates the effective propagation constant along the waveguide to wavelength, so it is the dispersion relationship for the slab waveguide.

The first mode has no nulls in the field distribution, the second has one, and so on. It follows that the n^{th} mode has n-1 nulls. We often call the modes with an even number of nulls the even modes and modes with an odd number of modes the odd

modes. This does not imply that the even and odd modes in general have even and odd symmetry. That is only the case if the waveguide itself is symmetric.

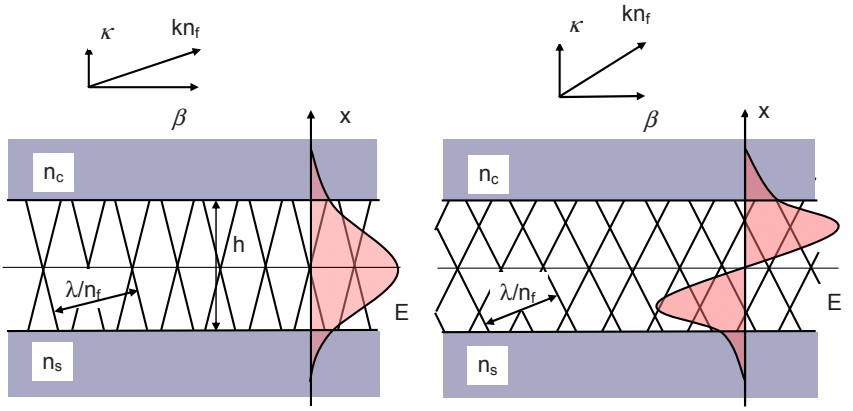


Figure 5.5. *Plane-wave picture of modes on optical waveguides. The fundamental mode (left) has the smallest transversal wavevector, and only one maximum in the field pattern. The first order mode (right) has two maxima in the field profile.*

The fact that different modes have different numbers of nulls gives us a useful way of indexing the modes. We call the mode with no nulls the zero-order mode, and the mode with one null the 1st order mode and so on. With two sets of modes (TE and TM) we index the modes in the following way: TE_0, TE_1, TE_2, TE_3….. TE_n and TM_0, TM_1, TM_2, TM_3….. TM_n.

5.3.4 Number of Modes

It is clear from Fig. 5.3 and 5.4 that the total number of solutions depends strongly on the waveguide thickness, *h*. If we decrease the thickness, the graph of the left hand side of the equation (dashed curve in Fig. 5.3) is stretched out along the κ-axis until the first part of the curve is too close to the x-axis to cross the graph representing the right hand side of the equation. By decreasing the waveguide thickness, we will therefore reach a situation in which there is no solution to the Eigenvalue equation for the longitudinal wavevector, i.e. no modes can propagate except if the waveguide is symmetric. We note that this happens at close to, but not exactly, the same thickness for the TE and TM cases.

A symmetric waveguide is different in this respect. It will have a guided mode for any waveguide thickness. This follows from the fact that when $\gamma_c = \gamma_s$, the right-hand side of the Eigenvalue equation is zero at $\kappa_f = \kappa_{max}$ as illustrated in Fig. 5.6. That guarantees at least one solution even as the thickness-dependent graph is stretched out as a consequence of the reduction of the core thickness.

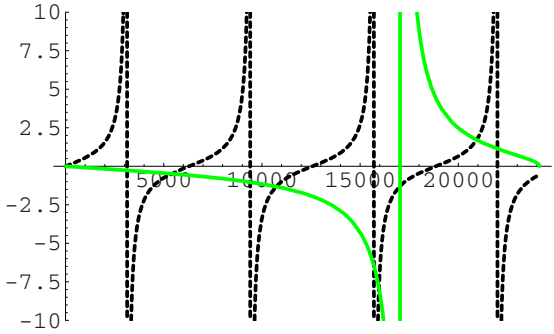


Figure 5.6. Graphical solution of the Eigenvalue equation for TE polarized modes in a symmetric waveguide. The symmetry guarantees a non-trivial solution for any waveguide thickness. Waveguide parameters: h=5 micron, n_c=1.45, n_f=1.5, n_s=1.45

Guided modes must have negative slopes in the outward direction at the core-cladding interface to be exponentially decaying in the cladding. This suggests a pictorial way to understand the difference between guided modes and radiation modes. Consider first the TE_1 mode on a symmetric waveguide as shown in Fig. 5.7. We will describe the modes in terms of their *V*-parameter

$$V = h \cdot \kappa_{\max} = h \cdot k_0 \sqrt{n_f^2 - n_s^2} \tag{5.17}$$

For $V >> \pi$, the field is well confined to the core. As *V* is decreased to π, the confinement factor is also decreased, until, at $V=\pi$, the mode is no longer guided. One way to understand that that as *h* is decreased, the mode doesn't have "room to turn", meaning that the turning points of the mode fall outside the core.

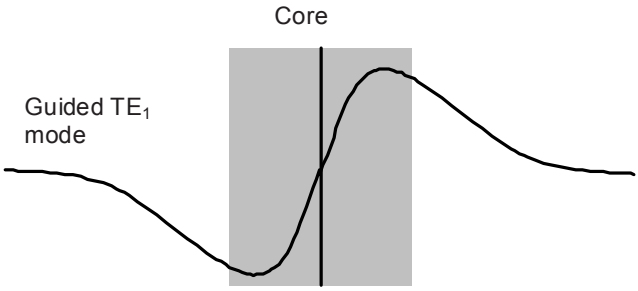


Figure 5.7. When the core is sufficiently wide, the TE_1 mode is guided. As the core width is decreased we reach a point where the turning points (extrema) of the mode profile falls outside the core. At that point the mode is no longer guided.

The TE_0 mode in a symmetric guide, on the other hand, exists for any thickness, because it only has one turning point placed in the center of the guide. This is illustrated in Fig. 5.8. In an asymmetric guide, even the TE_0 mode is cut-off at sufficiently small core thicknesses when the single turning point falls outside the core.

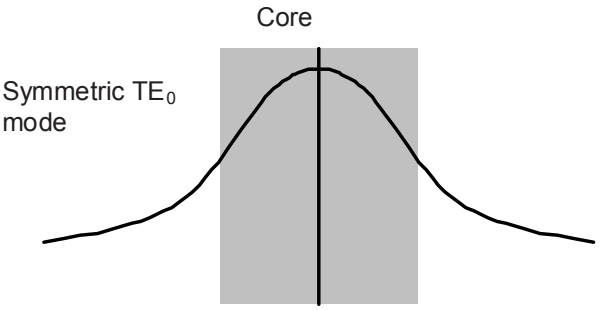


Figure 5.7. Symmetric TE_0 modes are always guided, while asymmetric TE_0 modes may be cut-off if the turning point falls outside the core.

For large core thicknesses we can derive an approximate expression for the number of guided modes supported by the structure. We note that if $\kappa_{max}h$ is larger than $\pi/2$, we are guaranteed at least one mode even in asymmetric guides, and that for each π increase in $\kappa_{max}h$ we have another solution. The total number of solutions is then approximately given by

$$m = Int\left[\frac{h\cdot\kappa_{\max}}{\pi}\right] = Int\left[\frac{h\cdot k_0\sqrt{n_f^2-n_s^2}}{\pi}\right] = Int\left[\frac{V}{\pi}\right] \tag{5.18}$$

5.3.5 Energy carried by a mode

To find the energy carried by a mode we integrate the time averaged Poynting vector over the cross section. For a slab waveguide we only integrate over one dimension to get the power per unit of length

$$P_z = \int_{-\infty}^{\infty} S_z\cdot dx = \int_{-\infty}^{\infty}\left(\vec{E}\times\vec{H}\right)\cdot\vec{z}\cdot dx \tag{5.19}$$

In practice we have waveguides with finite cross sections. The total energy flow can the be found by integration over the full cross section

$$P_z = \iint_A S_z \cdot dx = \iint_A \left(\vec{E} \times \vec{H}\right) \cdot \vec{z} \cdot dx \tag{5.20}$$

The time-averaged power then becomes

$$P_z = \iint_A \left\langle S_z \right\rangle \cdot dx = \frac{1}{2} \mathrm{Re}\left[\iint_A \left(\vec{E} \times \vec{H}^*\right) \cdot \vec{z} \cdot dx \right] \tag{5.21}$$

It can be shown that the modes are orthogonal. If we also normalize the power in each mode to 1 W, we can write the orthonormalization condition as

$$\frac{1}{2} \iint_A \left(\vec{E}_l \times \vec{H}_m^*\right) \cdot \vec{z} \cdot dx = \delta_{lm} \tag{5.22}$$

where δ_{lm} is the Kronecker delta. For TE and TM modes we get

$$\text{TE modes:} \quad \frac{\beta_m}{2\omega\mu} \iint_A \vec{E}_l \cdot \vec{E}_m^* \cdot \vec{z} \cdot dx = \delta_{lm} \tag{5.23}$$

$$\text{TM modes:} \quad \frac{\beta_m}{2\omega\varepsilon} \iint_A \vec{H}_l \cdot \vec{H}_m^* \cdot \vec{z} \cdot dx = \delta_{lm} \tag{5.24}$$

Consider the energy flow in an arbitrary optical field on the waveguide. The field can be expanded in guided and radiation modes as discussed above. It follows that the energy flow in the guide is the sum of the energy flow in each mode. The energy in the radiation modes will eventually be lost (a negligible amount will propagate in the vicinity of the waveguide). We may therefore discount these modes when considering the power carried by the waveguide, and we reach the conclusion that **the power propagating in a waveguide is the sum of the power in the guided modes.**

5.3.6 Properties of Modes

Our modeling has shown that there are several different types of waveguide modes. For longitudinal wave vectors in the range $kn_s<\beta<kn_f$, the modes of a slab waveguide are discrete and confined to the guiding film with exponentially decaying fields in the substrate and cladding. In the range $kn_c<\beta<kn_s$, the modes are continuous and have oscillatory behavior in the film and substrate, while they decay exponentially in the cladding. These modes are called substrate radiation modes. For $\beta<kn_c$, the modes are oscillatory in all three regions. These modes are simply called radiation modes.

There is a finite set of guided modes and a continuous, infinite set of radiation modes. Each eigenvalue of the longitudinal wavevector, β, corresponds to one or more (in the case of degenerate modes) distinct modes with unique field profiles. The modes of the slab are orthogonal, and they form a complete set so that any field profile can be written as a superposition of modes

$$A(x,y,z) = \sum_{guided} a_i A_i(x,y,z) + \int_{radiation} a(\beta) \cdot A(x,y,z,\beta) d\beta \tag{5.25}$$

where *A(x,y,z)* can be the *E* or the *H* field.

The important properties of modes are:

1. Each eigenvalue of the longitudinal wavevector, β, corresponds to a unique mode or field distribution.
2. Most modes are not guided. These are called radiation modes.
3. A finite number of modes are guided.
4. All modes are orthogonal.
5. Some modes are degenerate, i.e. they have the same longitudinal wave vector, β, but different field distributions.
6. The modes of an optical system form a complete set.
7. The power propagating in a waveguide is the sum of the power in the guided modes.

5.3.7 Normalized propagation parameters

Now we will develop a more general description of slab waveguides. Our approach is again based on graphical solutions of the characteristic equation for the guided modes, but we will generalize the parameters so that our results can be applied to a wide range of slab waveguides.

An asymmetric slab waveguide have four free parameters; the thickness of the core or film, and the refractive indices of the substrate, core and cladding. Together with the wave vector or wavelength of the optical field, this gives five parameters to describe any asymmetric slab waveguide with any monochromatic optical field.

To generalize our description, we will use the following dimensionless parameters for TE modes

Normalized frequency (*V* parameter): $V = k_0 h \sqrt{n_f^2 - n_s^2}$ (5.26)

$$\textbf{The asymmetry parameter:}\quad a = \frac{n_s^2 - n_c^2}{n_f^2 - n_s^2} \tag{5.27}$$

$$\textbf{Normalized effective index:}\quad b = \frac{\left(\frac{\beta}{k_0}\right)^2 - n_s^2}{n_f^2 - n_s^2} = \frac{n_{eff}^2 - n_s^2}{n_f^2 - n_s^2} \tag{5.28}$$

where $n_{eff} = \dfrac{\beta}{k_0}$ is the effective index of the mode.

Now we use these normalized parameters to rewrite the slab-waveguide dispersion relationship (Eq. 5.16) in the following form

$$V\sqrt{1-b} = m \cdot \pi + \tan^{-1}\sqrt{\frac{b}{1-b}} + \tan^{-1}\sqrt{\frac{b+a}{1-b}} \tag{5.29}$$

This expression is plotted for $a=0$ (symmetric waveguide) in Fig. 5.9. We can use this plot of the dispersion relation of the modes of the slab waveguide in generalized parameters to find solutions for specific waveguide structures.

Note on dispersion relations

The graphs of Fig. 5.9 are not the same wavevector (β) vs. natural frequency (ω) plots that we called dispersion diagrams in Chapter 2. They do, however, contain the same information as β–ω diagrams. The reason for choosing other axes is that the deviation from linearity of the β–ω relationship is very slight for most practical optical waveguides and fibers. A straightforward β–ω would not allow the slight, but important, non-linearities to be observed, so other plot parameters designed to highlight the non-linearities are more useful.

If we want to find the longitudinal wavevector at a given frequency in a given slab waveguide, we use the following procedure:

1. Calculate the normalized frequency for the specific waveguide and wavelength
2. Find the normalized index, b, for the desired mode from the plot (the plot shown in Fig. 5.9 is for $a=0$, but similar plots for any arbitrary a can easily be generated).
3. Calculate n_{eff} for the different modes from the b values.
4. The longitudinal wavevector is given by $\beta = k_0\, n_{eff}$

This procedure gives results that are in excellent agreement with direct numerical solutions for specific waveguides.

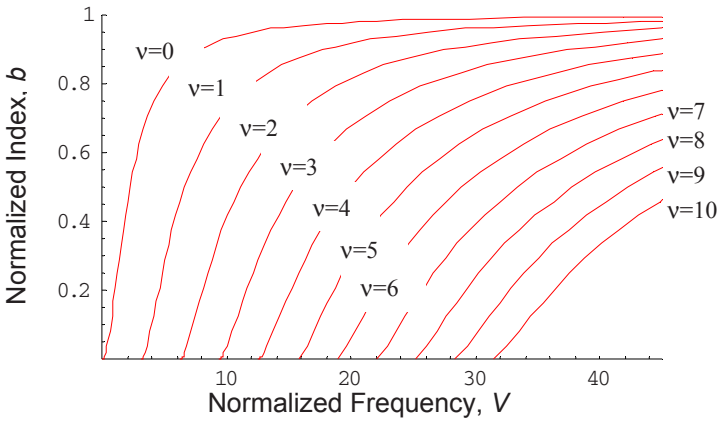


Figure 5.9. Dispersion relation for the TE modes of a symmetric slab waveguide in terms of normalized parameters.

In terms of the normalized parameters we have introduced, the cut-off condition is simply $b = 0$, which leads to

$$V_{cutoff} = \tan^{-1}\left[\sqrt{a}\right] + \nu\pi \tag{5.30}$$

It is much simpler to use this expression than to find the cut-off numerically.

The above development is valid for TE modes. For TM modes the treatment is similar, but we must adjust the asymmetry parameter

$$\textbf{Asymmetry parameter for TM modes:} \quad a = \frac{n_f^2}{n_c^2}\frac{n_s^2 - n_c^2}{n_f^2 - n_s^2} \tag{5.31}$$

Dispersion diagrams like the ones of Fig. 5.9 contain a wealth of information about wave propagation. We will use similar dispersion diagrams to understand the propagation characteristics of a variety of optical waveguides and fibers.

5.4 Optical Fibers

Optical fibers have revolutionized the communications industry due to their low loss and inexpensive fabrication. This second point is often underestimated. Optical fibers are fabricated by pulling preformed rods with carefully designed index profiles into long fibers. This process is well controlled and results in highly uniform core and cladding diameters. The uniformity of the fibers is so good that for most practical purposes we consider the fiber a perfectly cylindrical waveguide. If we also assume that the refractive index changes are binary, i.e. the index values

are constant in the core and cladding and change abruptly at the core-cladding interface, then we have a step-index fiber.

The attraction of the step-index fiber model is that the modes can be found analytically. We will not go through the tedious derivation of the mode profiles, but we will freely use the results because of the insight they offer into the nature of wave propagation on optical fibers. When applying the analytical solutions to practical situations, we have to keep in mind that real fibers do not have step-function refractive-index variations.

5.4.1 Modes in Step-Index Optical Fibers

A step-index optical fiber is shown in Fig. 5.10. In the mathematical treatment of the step-index fiber it is customary to assume that the cladding is infinite, so the fiber can be characterized by three parameters: The core radius, a, and the core and cladding refractive indices, n_{core} and $n_{cladding}$. Together with the wave vector of the optical field, these parameters completely determine the propagation of light on the fiber. Just as for the dielectric-slab waveguide, it is very convenient to capture these four parameters in the normalized frequency

$$V = k_0 a\sqrt{n_{core}^2 - n_{cladding}^2} = \frac{2\pi \cdot a}{\lambda}\sqrt{n_{core}^2 - n_{cladding}^2} \tag{5.32}$$

Note that a here is the radius of the core. In the fiber literature, this normalized frequency is referred to as the *V-parameter* or the *V-number*. We will use both names interchangeably.

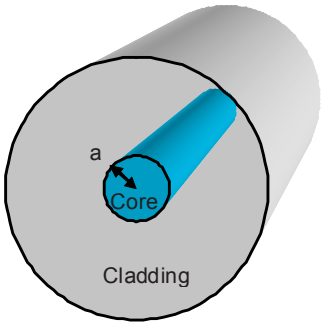
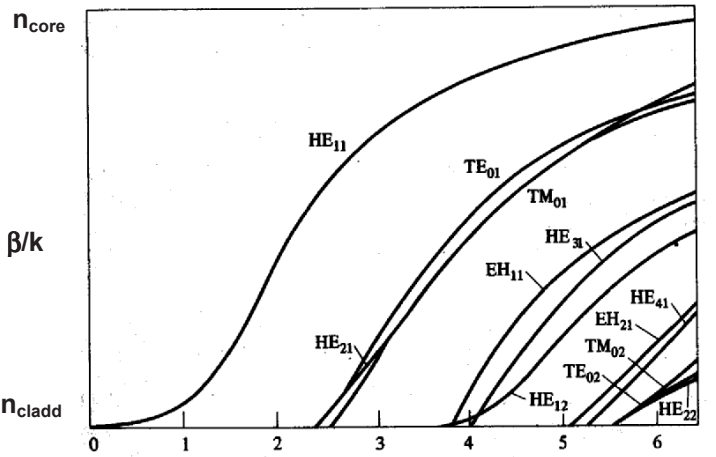


Figure 5.10. Geometry of a step-index cylindrical waveguide (optical fiber) with core radius a. The cladding diameter (125 um for standard single mode fiber) is chosen large enough that we can consider it infinite in our calculations.

The analytical description of the modes allows the dispersion relationship for the step-index fiber of Fig. 5.10 to be computed. The results for a number of the lowest order modes are shown in Fig. 5.11. The most important things to note in this relatively complex dispersion diagram are that (1) a step-index fiber supports at least one mode for any value of the analytical solutions V-parameter, and (2) the

fiber support only one mode, the HE_{11} mode, for *V*-parameters smaller than 2.405. The HE_{11} mode is really two orthogonally polarized modes. Due to the circular symmetry of the fiber, these two modes are degenerate, i.e. they have the exact same effective index at any analytical solutions number.



$$V = k_0 a\sqrt{n_{core}^2 - n_{cladding}^2}$$

Figure 5.11 *Normalized longitudinal wavevector (or propagation constant) for the lower order modes of a step-index fiber as a function of normalized frequency. Reprinted from [1] with permission.*

The inequality *V<2.405* is called the cut-off condition. From the definition of the *V-parameter*, we see that single-mode operation require long wavelengths, small fiber radii, and small index contrasts. The cut-off condition is often rewritten in terms of the wavelength

$$k_0 a\sqrt{n_{core}^2 - n_{clad}^2} < 2.405 \Rightarrow \lambda > \lambda_c = \frac{2\pi \cdot a\sqrt{n_{core}^2 - n_{clad}^2}}{2.405} \tag{5.33}$$

For typical fiber parameters (a = 3.955 μm, n_{core} = 1.4514 and $n_{cladding}$ = 1.4469), we find that the cut-off wavelength is 1.18 um. In other words, a standard single mode fiber (SMF) is single-mode only at wavelengths beyond 1.18 um. That includes the two most important fiber-optical wavelength ranges around 1.55 um and 1.3 um.

5.4.2 Linearly Polarized Modes

The modes of step-index fibers are very complex, to the point that their analytical descriptions are of limited value. Weakly guiding optical fibers (i.e. fibers with a small refractive-index difference between the core and cladding), on the other

hand, have modes that are, to a good approximation, transverse electro-magnetic (TEM) waves. These approximate TEM modes are simple to visualize and good models for understanding lower-order modes on real fibers.

In the weakly-guided approximation, the amplitude of the electric field of the fiber modes can be described in the following way

$$\text{Core (r < a):}\ E_x = E_0 J_\nu(\kappa \cdot r)\cos(l\phi) \tag{5.34}$$

$$\text{Cladding (r > a):}\ E_x = E_0 \frac{J_\nu(\kappa \cdot a)}{K_\nu(\gamma \cdot a)} K_\nu(\gamma \cdot r)\cos(l\phi) \tag{5.35}$$

where J_ν is the Bessel function and K_ν the modified Bessel function of the second kind.

The parameters κ and γ are found from the equations:

$$h\frac{J_{\nu\pm1}(\kappa \cdot a)}{J_\nu(\kappa \cdot a)} = \pm q \frac{K_{\nu\pm1}(\gamma \cdot a)}{K_\nu(\gamma \cdot a)} \tag{5.36}$$

$$a^2\left(\kappa^2 + \gamma^2\right) = V^2 = \left(2\pi\frac{a}{\lambda}\right)\left(n_{core}^2 - n_{cladding}^2\right) \tag{5.37}$$

To be able to plot the modes of a cylindrical waveguide or fiber, we must solve these equations at the wavelengths of interest. Assume the following fiber parameters: a = 3.955 μm, n_{core} = 1.4514 and $n_{cladding}$ = 1.4469. At 1310 nm wavelength we then have V = 2.166, and the characteristic equation has only one solution (ν = 0, κ = 0.39975525 μm^{-1}, γ = 0.374337 μm^{-1}). The fundamental mode (LP_{01}) can then be expressed as:

$$\text{Core } (r < a):\ E_x = E_0 J_0\left(0.39975525\mu m^{-1} \cdot r\right) \tag{5.38}$$

$$\text{Cladding } (r > a):\ E_x = E_0 \frac{J_0\left(0.39975525\mu m^{-1} \cdot a\right)}{K_0\left(0.374337\mu m^{-1} \cdot a\right)} K_0\left(0.374337\mu m^{-1} \cdot r\right) \tag{5.39}$$

As stated above, the LP_{01} mode is very similar to the HE_{11} on weakly guided fibers. Just like the HE_{11}, it is really not a single mode, but rather two orthogonally polarized modes.

At 670 nm wavelength we have V = 4.236. The characteristic equation then has four solutions; LP_{01} (two degenerate polarization modes), LP_{11} (four-fold degenerate), LP_{21} (four-fold degenerate), *and* LP_{02} (two degenerate polarization modes). These modes are shown in Figs. 5.12 through 5.15.

The two circularly symmetric modes in these graphs are each two orthogonal polarization modes, and the two other modes consist of four degenerate modes. In addition to the one shown and its orthogonal polarization, there are two orthogonally polarized modes with a $sin\phi$ or $sin2\phi$ angular dependence. Counting all polarizations and both helical polarities of the LP_{11} and LP_{21} modes, we find a total of 12 modes for $V = 4.236$. Not all of these produce distinguishable intensity patterns.

LP_{01} mode ($\nu = 0$, $\kappa = 0.488154$, $\gamma = 0.953337$):

$$\text{Core } (r < a): \; E_x = E_0 J_0(\kappa \cdot r) \tag{5.40}$$

$$\text{Cladding } (r > a): \; E_x = E_0 \frac{J_0(\kappa \cdot a)}{K_0(\gamma \cdot a)} K_0(\gamma \cdot r) \tag{5.41}$$

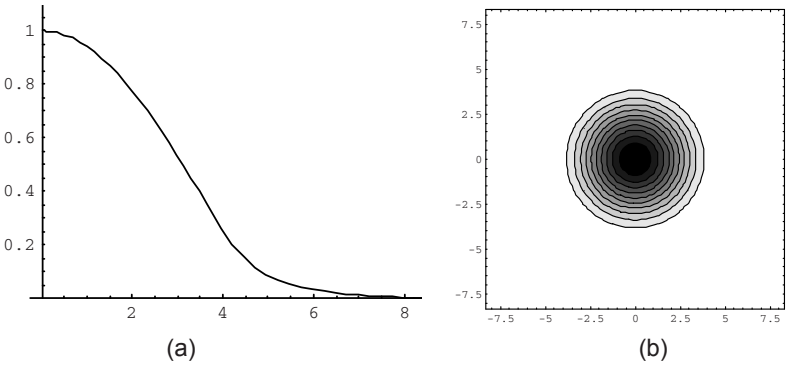


Figure 5.12. (a) Electrical field amplitude of LP_{01} as a function of radius (in μm) at 670 nm wavelength. (b) Contour plot of LP_{01} mode at 670 nm wavelength.

LP_{11} mode ($\nu = 1$, $\kappa = 0.76807$, $\gamma = 0.746468$):

$$\text{Core } (r < a): \; E_x = E_0 J_1(\kappa \cdot r)\cos\phi \tag{5.42}$$

$$\text{Cladding } (r > a): \; E_x = E_0 \frac{J_1(\kappa \cdot a)}{K_1(\gamma \cdot a)} K_1(\gamma \cdot r)\cos\phi \tag{5.43}$$

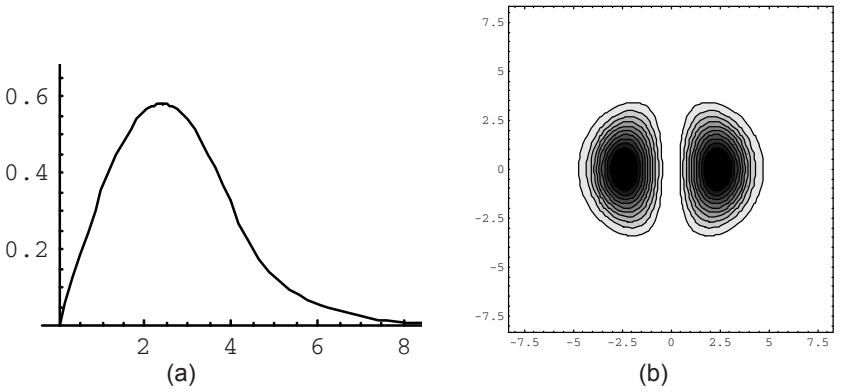


Figure 5.13. *(a) Electrical field amplitude of LP_{11} as a function of radius (in μm) at 670 nm wavelength. (b) Contour plot of LP_{11} mode at 670 nm wavelength.*

LP_{21} ($\nu = 2$, $\kappa = 1.00806325$, $\gamma = 0.361877$):

$$\text{Core } (r < a): \; E_x = E_0 J_2(\kappa \cdot r)\cos(2\phi) \tag{5.44}$$

$$\text{Cladding } (r > a): \; E_x = E_0 \frac{J_2(\kappa \cdot a)}{K_2(\gamma \cdot a)} K_2(\gamma \cdot r)\cos(2\phi) \tag{5.45}$$

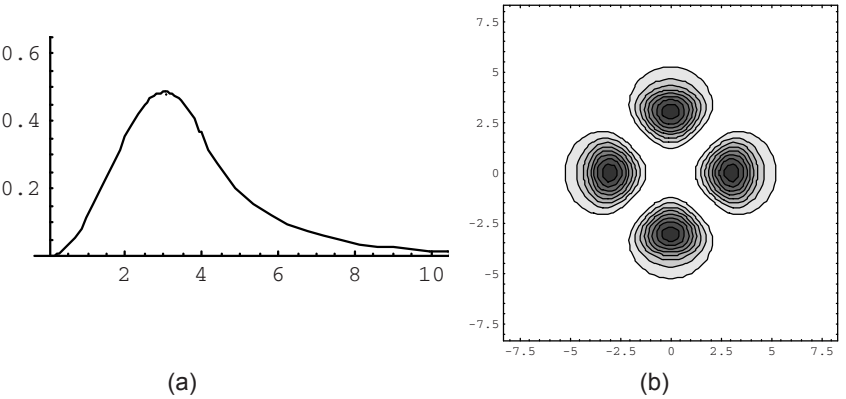


Figure 5.14 *(a) Electrical field amplitude of LP_{21} as a function of radius (in μm) at 670 nm wavelength. (b) Contour plot of LP_{21} mode at 670 nm wavelength.*

LP_{02} ($\nu = 0$, $\kappa = 1.0482116$, $\gamma = 0.219998$):

$$\text{Core } (r < a): \; E_x = E_0 J_0(\kappa \cdot r) \tag{5.46}$$

$$\text{Cladding } (r > a): \; E_x = E_0 \frac{J_0(\kappa \cdot a)}{K_0(\gamma \cdot a)} K_0(\gamma \cdot r) \tag{5.47}$$

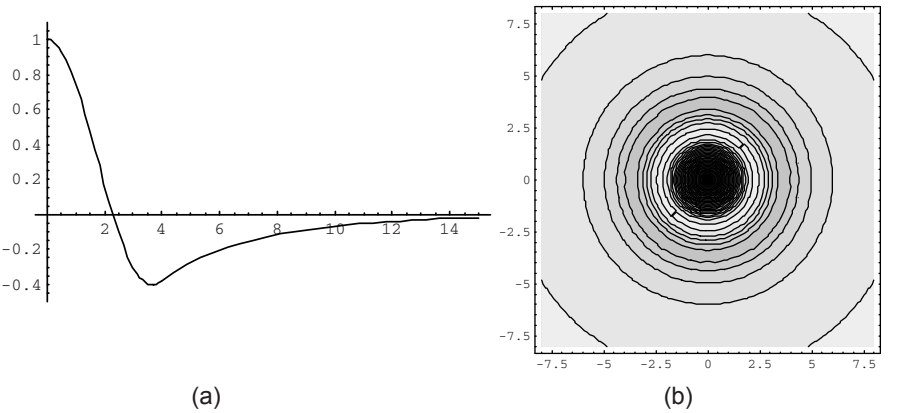


Figure 5.15. (a) Electrical field amplitude of LP_{02} as a function of radius (in μm) at 670 nm wavelength. (b) Contour plot of LP_{02} mode at 670 nm wavelength.

5.4.3 The Fundamental Mode of a Cylindrical Waveguide

The HE_{11} mode of the cylindrical waveguide is guided for all values of the normalized frequency (V-number), and when the analytical solutions V-number is less than 2.405, the HE_{11} mode is the only guided mode. The HE_{11} mode is of special importance, because single-mode operation is the preferred way to use optical fibers in high-capacity communication systems. The details of the field distribution of the HE_{11} mode is therefore important in a variety of calculations of mode propagation, dispersion, coupling, switching, cross talk, and modulation on optical fibers and waveguide devices.

The difficulty of these calculations is often substantial because of the complex nature of the HE_{11} mode. Fortunately, the Gaussian approximation to the HE_{11} mode shape is sufficiently accurate for the majority of fiber mode calculations. In the Gaussian approximation, the mode field is

$$\vec{E}(r) = E_x \exp\left[-\left(\frac{r}{\omega}\right)^2\right] \tag{5.48}$$

where the beam 1/e-field beam radius (1/e^2 intensity), ω, is chosen to give the best match to the HE_{11} mode shape. For a step-index fiber with radius a, the best-match beam radius is

$$\frac{\omega}{a} = 0.65 + 1.619 \cdot V^{-1.5} + 2.87 \cdot V^{-6} \tag{5.49}$$

This Gaussian approximation is compared to the LP_{01} mode in Fig. 5.16 with the following fiber parameters: V-number: $V=2.166$, Core radius: $a=3.955\mu m$, Wavelength: $\lambda=1.310\ \mu m$.

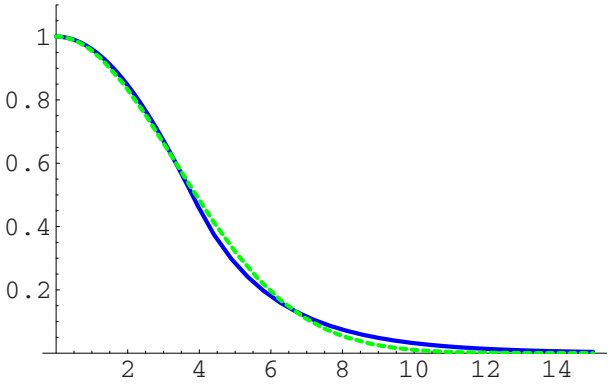


Figure 5.16. Comparison of LP_{01} mode (solid) and the Gaussian approximation (dashed) to the HE_{11} mode. We see that the two mode shapes are well matched, but that the Gaussian falls off quicker at large radii. That is of importance in some cross-talk calculations.

5.4.4 Power Confinement

As for the slab waveguide, the total power in a mode is calculated by integrating

$$\langle S_z \rangle = \frac{1}{2}\mathrm{Re}\left(\vec{E} \times \vec{H}\right) \tag{5.50}$$

over the cross sectional area of the waveguide. In the weakly guiding limit, the ratio of the power carried in the core to the total power is

$$\frac{P_{core}}{P_{total}} = \left(1 - \frac{\kappa^2 a^2}{V^2}\right)\left(1 - \frac{K_\nu^2(\gamma \cdot a)}{K_{\nu+1}(\gamma \cdot a)K_{\nu-1}(\gamma \cdot a)}\right) \tag{5.51}$$

It follows that the ratio of the power carried in the cladding to the total power can be expressed as

$$\frac{P_{clad}}{P_{total}} = 1 - \frac{P_{core}}{P_{total}} \tag{5.52}$$

As in the slab waveguide we studied before, the power confinement factor for a given mode is very low at the cut-off for the mode, and increases with increasing V-number above cut-off.

5.5 Dispersion

Dielectric optical waveguides have much higher bandwidth than coaxial cables and other metal waveguides used for radio-frequency communication, because higher frequencies are strongly attenuated on metallic waveguides. The maximum frequency and bandwidth of metallic guides are therefore quite limited. Dielectric waveguides, on the other hand, have very low loss, so the bandwidth is not limited by the range of frequencies that can propagate without excess attenuation, but rather by pulse spreading or dispersion.

There are three types of phenomena that create dispersion on waveguides; material dispersion, waveguide dispersion, and modal dispersion. These will be present in different amounts on different types of waveguides. Metallic waveguides, for example, have little material dispersion or waveguide dispersion, because the fields are propagating in vacuum or air. In optical waveguides, the material is typically glass, and the distribution of fields in the core and cladding depends on wavelength, so both material and waveguide dispersion are important.

Modal dispersion is caused by the fact that different modes have different velocities, so it can be eliminated by using single mode waveguides. We often say that modern optical fiber is single mode, so we should expect these types of waveguides to be free of modal dispersion. The standard single mode fiber is, however, not single mode, but supports two nominally degenerate modes of orthogonal polarization. In practice, these modes are not completely degenerate, and polarization mode dispersion (PMD) does in fact contribute to signal degradation on modern single mode fibers.

We will now first consider the three sources of dispersion separately, and then we will describe their combined effects on the communication characteristics of opti-

cal waveguides. Finally, we will investigate approaches and devices designed to negate the negative effects of dispersion on signal fidelity.

5.5.1 Material Dispersion

Dispersion in a solid is caused by the fact that the index of refraction of the solid has a non-linear dependence on frequency or wavelength. A linear dependence does not lead to pulse spreading because the group velocity is constant. To quantify our discussion of material dispersion, we develop a simple model of for the response of a solid to an electric field. Consider the classical electron model or Lorenz model of Fig. 5.17 below.

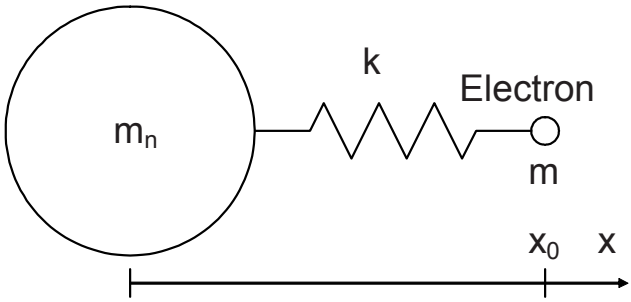


Figure 5.17 Classical Electron Model (CEO) of the atomic response to an applied optical field.

Making the assumption that the mass of the nucleus is so much larger than the electron mass that we can consider the nucleus stationary, we find the equation of motion of the electron by summing the forces that are acting on it

$$m\frac{d^2x(t)}{dt^2}+\gamma\frac{dx(t)}{dt}+kx(t)=-eE(t) \tag{5.53}$$

Here we have introduced a loss term to account for absorption or emission of electromagnetic energy.

We now rewrite this equation in terms of the resonance frequency of the spring mass system, $\omega_0=\sqrt{\frac{k}{m}}$

$$\frac{d^2x(t)}{dt^2}+\gamma\frac{dx(t)}{dt}+\omega_0^2x(t)=-\frac{e}{m}E(t) \tag{5.54}$$

With a harmonically varying optical field driving the dipole, we expect a harmonic response.

$$E(t) = E(\omega)e^{j\omega t} \qquad \Rightarrow \qquad x(t) = x(\omega)e^{j\omega t} \tag{5.55}$$

We can then rewrite the equation in phasor form

$$(j\omega)^2 x(\omega) + j\omega\gamma x(\omega) + \omega_0^2 x(\omega) = -\frac{e}{m}E(\omega) \tag{5.56}$$

with the solution

$$x(\omega) = \frac{-\frac{e}{m}}{(j\omega)^2 + j\gamma\omega + \omega_0^2}E(\omega) \tag{5.57}$$

The electric displacement can then be expressed

$$\vec{D} = \varepsilon_0\vec{E} + \vec{P} = \varepsilon_0\vec{E} - Ne\vec{x} = \varepsilon_0\vec{E} + \frac{\frac{e^2}{m}}{(j\omega)^2 + j\gamma\omega + \omega_0^2}\vec{E} \tag{5.58}$$

We see that the relative dielectric constant can be expressed

$$\frac{\varepsilon}{\varepsilon_0} = 1 + \frac{N\frac{e^2}{\varepsilon_0 m}}{(j\omega)^2 + j\gamma\omega + \omega_0^2} = \left(1 + \frac{N\frac{e^2}{m}\left(\omega_0^2 - \omega^2\right)}{\varepsilon_0\left[\left(\omega_0^2 - \omega^2\right)^2 + \gamma^2\omega^2\right]} - j\frac{N\frac{e^2}{m}\gamma\omega}{\varepsilon_0\left[\left(\omega_0^2 - \omega^2\right)^2 + \gamma^2\omega^2\right]}\right) \tag{5.59}$$

Finally we find the refractive index as the square root of the relative dielectric constant

$$n = \sqrt{\frac{\varepsilon}{\varepsilon_0}} = \left(1 + \frac{N\frac{e^2}{m}\left(\omega_0^2 - \omega^2\right)}{\varepsilon_0\left[\left(\omega_0^2 - \omega^2\right)^2 + \gamma^2\omega^2\right]} - j\frac{N\frac{e^2}{m}\gamma\omega}{\varepsilon_0\left[\left(\omega_0^2 - \omega^2\right)^2 + \gamma^2\omega^2\right]}\right)^{0.5} \tag{5.60}$$

In a typical solid we have several electron states with different resonance frequencies contributing to the dielectric constant, which then can be expressed

$$\varepsilon = \varepsilon_0 + \sum_{i=1}^{Z} \frac{f_i N \frac{e^2}{m}}{(j\omega)^2 + j\gamma\omega + \omega_0^2} \tag{5.61}$$

where f_i is the oscillator strength of the resonance.

This information is expressed in the empirical *Sellmeier equation*

$$n^2 - A = \sum_k \frac{G_k \lambda^2}{\lambda^2 - \lambda_k^2} \tag{5.62}$$

The Sellmeier coefficients are available for many optical materials of interest. The most common materials can be conveniently looked up on Wikipedia.

5.5.1.1 Frequency Dependent Dielectric Constant

The dielectric constant, ε, is defined through the constitutive relation

$$\vec{D}(\vec{r},t) = \varepsilon \vec{E}(\vec{r},t) \tag{5.63}$$

This time-domain equation has no frequency dependent terms. To introduce frequency dependence we turn to the phasor representation of plane waves.

$$\begin{aligned} &\vec{E}(\vec{r},t) = \vec{x} \cdot E_0(\vec{r}) \cdot \cos(\omega t - \phi_x(\vec{r})) = \frac{1}{2} E_x(\vec{r}) e^{j(\omega t - \phi_x(\vec{r}))} + c.c. = \\ &\mathrm{Re}\left[E_x(\vec{r}) e^{j(\omega t - \phi_x(\vec{r}))}\right] = \mathrm{Re}\left[E_x(\vec{r}) e^{j\omega t}\right] \end{aligned} \tag{5.64}$$

where the phase factor, $exp[-j\theta_x]$, has been included in the field amplitude in the last expression.

In phasor notation we drop the explicit taking of the real part.

$$\vec{E}(\vec{r},t) = \vec{E}(\vec{r},\omega) e^{j\omega t} \tag{5.65}$$

where E_x is a complex quantity with six components (three vector components each with amplitude and phase). Maxwell's equations are linear so we can write an arbitrary time waveform as a sum of monochromatic plane waves

$$\vec{E}(\vec{r},t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \vec{E}(\vec{r},\omega) \cdot \exp[j(\omega t - kz)] d\omega \tag{5.66}$$

where again the electric field symbol in the integral is a phasor, i.e. a complex quantity. The constitutive relation can now be written

$$\vec{D}(\vec{r},\omega) = \varepsilon(\omega) \cdot \vec{E}(\vec{r},\omega) \tag{5.67}$$

In the time domain this becomes

$$\begin{aligned}\vec{D}(\vec{r},t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\vec{D}(\vec{r},\omega)\cdot\exp\left[j(\omega t-kz)\right]d\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\varepsilon(\omega)\cdot\vec{E}(\vec{r},\omega)\cdot\exp\left[j(\omega t-kz)\right]d\omega = \varepsilon(t)\otimes\vec{E}(\vec{r},t)\end{aligned} \tag{5.68}$$

where $\varepsilon(t)$ is the impulse response of the dielectric constant.

Only if the impulse response is a delta function do we recover the "standard" constitutive relation. Vacuum has a delta-function impulse response, but all real materials have impulse responses of finite duration. When we talk about the relative dielectric constant of a material, we therefore always mean the proportionality constant relating the phasor or Fourier component of the electric displacement to the phasor or Fourier component of the electric field. If we are considering arbitrary waveforms in the time domain, we must use the Fourier integral or the convolution with the impulse response. Similar considerations are valid for the permitivity.

Now that we have established that we are interested in the dielectric constant in the Fourier or frequency domain, let's see what values this quantity can take. Clearly the dielectric constant can be real and larger than unity. This is the standard value for most optical waveguide materials. By considering the Lorenz model for the polarization of a solid, we also realize that the dielectric constant can be less than unity and even negative. That happens when the polarization of the material is out of phase with the applied optical field. A negative value of the dielectric constant can be used as a functional definition of a metal.

The dielectric constant may also be complex. This is the case in materials with gain or loss. The sign of the imaginary part of the dielectric constant for a lossy material **depends on the convention we choose for the phasors**. One way to see this is to consider the complex Poynting theorem

$$\begin{aligned}&-\int_{surface}\left(\vec{E}\times\vec{H}^*\right)\cdot\vec{n}\,dv = \\ &\int_{volume}\left[j\omega\left(\varepsilon_0\vec{E}\cdot\vec{E}^*+\mu_0\vec{H}\cdot\vec{H}^*+\vec{E}\vec{P}^*+\mu_0\vec{H}\vec{M}^*\right)+\vec{E}\cdot\vec{J}^*\right]\cdot dv = \\ &\int_{volume}\left[j\omega\left(\varepsilon\vec{E}\cdot\vec{E}^*+\mu_0\vec{H}\cdot\vec{H}^*+\mu_0\vec{H}\vec{M}^*\right)+\vec{E}\cdot\vec{J}^*\right]\cdot dv\end{aligned} \tag{5.69}$$

The term $j\omega\varepsilon EE^*$ is positive, which corresponds to energy loss, if the imaginary part of the dielectric constant is negative. With the conventions we have chosen, lossy materials will therefore have a negative imaginary part of their dielectric

constant. Another way to see this is to compare the loss term due to the dielectric constant to the current density in Ampere's law

$$\nabla \times \vec{H} = j\omega\vec{D} + \vec{J} = j\omega\varepsilon\vec{E} + \vec{J} \tag{5.70}$$

Again we see that a negative imaginary part of the dielectric constant is required to give the real part of the $j\omega D$ term the same sign as the current density term.

5.5.1.2 Group Delay Caused by Material Dispersion

The frequency dependencies of the waveguide materials lead to variation of the group velocity as a function of frequency. This variation means that the different frequencies of a pulse move at different speeds, leading to pulse spreading. The details of how pulses are distorted by dispersion are analyzed in detail in Section 5.6. In this section we will just give a first-order description.

Recall that the group velocity is given by

$$v_g = \frac{d\omega}{dk} \tag{5.71}$$

The inverse of the group velocity is the group delay, which can be expressed

$$\begin{aligned} \tau_g = \frac{1}{v_g} = \frac{d\beta}{d\omega} = \frac{\partial\beta}{\partial k}\frac{dk}{d\omega} + \frac{\partial\beta}{\partial n}\frac{dn}{d\omega} = \frac{n}{c}\frac{\partial\beta}{\partial k_0} + \frac{\omega}{c}\frac{dn}{d\omega} \approx \\ \frac{n}{c} + \frac{\omega}{c}\frac{dn}{d\omega} = \frac{1}{c}\left(n + \omega\frac{dn}{d\omega}\right) = \frac{N_g}{c} \end{aligned} \tag{5.72}$$

where N_g is the group index. Here we have ignored the term $\frac{\partial\beta}{\partial k_0}$, which describes how the wave vector depends on the wavelength in the absence of material dispersion. This effect is called waveguide dispersion, and we will get back to it in the next section.

The group index can be expressed

$$N_g = n + \omega\frac{dn}{d\omega} = n - \lambda\frac{dn}{d\lambda} \tag{5.73}$$

The term $dn/d\lambda$ is negative in most materials in the wavelength range of interest. We therefore say that we have *normal dispersion* when the group index is larger than the refractive index.

The spread of a pulse of bandwidth $\Delta\lambda$ traveling a distance L is approximately

$$\Delta\tau = \frac{L}{c}\Delta N_g = \frac{L}{c}\frac{dN_g}{d\lambda}\Delta\lambda \tag{5.74}$$

From the equation for the group index, we find

$$\frac{dN_g}{d\lambda} = \frac{d}{d\lambda}\left(n - \lambda\frac{dn}{d\lambda}\right) = -\lambda\frac{d^2n}{d\lambda^2} \tag{5.75}$$

so pulse spread becomes

$$\Delta\tau_m = -L\cdot\Delta\lambda\cdot\frac{\lambda}{c}\frac{d^2n}{d\lambda^2} = L\cdot\Delta\lambda\cdot D \tag{5.76}$$

where D is the material dispersion. We see that it is proportional to the second derivative of the index with respect to the wavelength.

5.5.2 Waveguide Dispersion

Going back to Eq. 5.72, but this time ignoring material dispersion, we find that the group delay becomes

$$\tau_g = \frac{1}{v_g} = \frac{d\beta}{d\omega} = \frac{\partial\beta}{\partial k}\frac{dk}{d\omega} + \frac{\partial\beta}{\partial n}\frac{dn}{d\omega} = \frac{n}{c}\frac{\partial\beta}{\partial k_0} + \frac{\omega}{c}\frac{dn}{d\omega} \approx \frac{1}{c}\frac{\partial\beta}{\partial k} \tag{5.77}$$

In the absence of material dispersion, the pulse spread can then be expressed

$$\Delta\tau_w = \frac{1}{c}\Delta\left(\frac{d\beta}{dk}\right) \approx \frac{1}{c}\frac{d^2\beta}{dk^2}\Delta k = \frac{k^2}{2\pi c}\frac{d^2\beta}{dk^2}\Delta\lambda \tag{5.78}$$

We typically don't have closed-form solutions for β (we find it by solving the eigenvalue equation numerically or graphically), so we must evaluate $\Delta\tau_g$ numerically. It is instructive, however, to consider the conceptual dispersion diagrams of typical guided modes as shown in Fig. 5.18. Here we have greatly exaggerated the difference between the core and cladding index, to clarify the dependence of the phase and group velocities on frequency.

If we follow an individual mode in Fig. 5.18, we see that at low frequencies just above cut-off, the mode is loosely bound to the core and therefore has an index close to that of the cladding. As the frequency increases, the mode is better and better confined, so its phase and group index approaches that of the core. In between these extremes, the curve goes through a point where the slope, and therefore the group index, is maximized. At that point the curvature (second derivative) is zero, and so is the pulse spread (to first order). There is also a point where the curvature, and therefore the pulse spread, is maximized. This means that the

pulse spreading is maximied of modes that are transitioning between being loosely-bound and well-confined to the core.

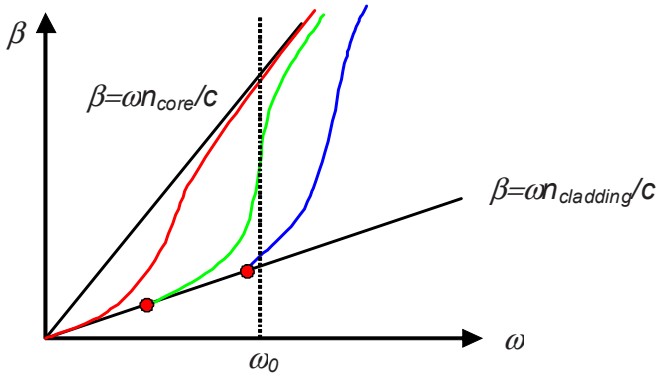


Figure 5.18. Schematic diagram of guided modes in a slab waveguide over an extended range of wavelengths. The group index for a mode starts at a value equal to the cladding index for low frequencies, goes through a minimum (where the dispersion is zero), and approaches the core value for high wave vectors. (This figure is not to scale. It is conceptual and not meant for accurate graphical solutions.)

The conceptual descriptions of the longitudinal wavevector, group velocity, and dispersion in Fig. 5.18 are valid for most standard wave guides. It allows us to make the following observations:

1. The waveguide wavevector, β, is very close to a linear function of the vacuum wave vector ($\beta = n\, k_0$)
2. The effective group index increases from the cladding to a value higher than the core index, before it asymptotically approaches the core value.
3. The dispersion first decreases to a minimum negative value, goes through a null at the inflection point, increases to a maximum positive value before it asymptotically approaches zero for well confined modes.

5.5.3 Modal Dispersion

In addition to the material and waveguide dispersion described in sections 5.5.1 and 5.5.2, multimode waveguides will also experience modal dispersion. This arises from the fact that different modes have different propagation speeds. An optical field that is coupled into a multimode waveguide will excite multiple modes that propagate at different velocities, causing pulse spread and distortion.

The delay difference caused by modal dispersion is independent of the signal bandwidth. This is contrary to material and waveguide dispersion that are both proportional to the signal bandwidth (Eqs. 5.76 and 5.78). The effect of modal dispersion is, however, bandwidth dependent, because pulse spreading *relative to*

the pulse length is proportional to the bandwidth, so high-bandwidth signals are more adversely affected.

To understand the range of group velocities on a typical waveguide, we again consider Fig. 5.18. It shows that the group velocity is a complex function of the mode number. As we go from low to high mode numbers at a given frequency (i.e. we are following a vertical line downwards in the diagram), we see that the group velocity decreases from the phase velocity of the core to a minimum velocity, and then increases again to the phase velocity of the cladding.

The full range of group velocities therefore extend from the minimum velocity of the intermediate modes to the high velocity of the high-order modes. The high-order modes are, however, very loosely bound to the core, so they tend to have high loss in the presence of waveguide imperfections. In practical waveguides we can therefore ignore the high-order modes. The lowest group velocities we consider are therefore those of the low-order modes, i.e. modes that are well confined and therefore have a group index close to the refractive index of the core. The full range of group indecies therefore extends from the core index on the low side to the group index of the minimum-velocity, intermediate modes on the high side.

To calculate the highest group velocity at a given frequency, we would need an accurate description of the waveguide dispersion diagram. This level of sophistication is not required in most practical situations. Typically it suffices to use a crude geometrical-optics model as shown in Fig. 5.19. Here we say that slowest mode on the fiber is the one that is propagating at the Total-Internal-Reflection angle, given by

$$\theta_{cr} = \sin^{-1}\left(\frac{n_{cladding}}{n_{core}}\right) \tag{5.79}$$

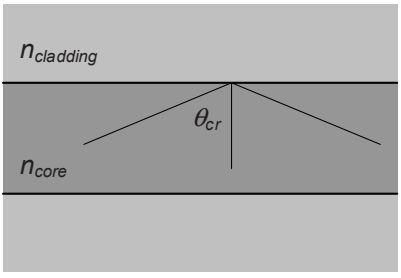


Figure 5.19. Simple geometrical-optics model for calculating the dependence of group velocity on mode number. According to the model the lowest-order modes travel along the axis of the optical waveguide (shortest path), while the modes close to cut-off travel at the TIR cut-off angle.

Within this model, the temporal pulse spread is

$$\Delta\tau = \tau_{high-order} - \tau_{low-order} = \frac{n_{core}}{c}\frac{1}{\sin\theta_{cr}} - \frac{n_{core}}{c} = \frac{n_{core}}{c}\left(\frac{n_{core}}{n_{cladding}} - 1\right) \approx \frac{n_{core}}{c} - \frac{n_{cladding}}{c} \tag{5.80}$$

This simple expression is very useful for calculating modal dispersion, although we should keep in mind that it really isn't correct that the modes at cut-off propagate slower than the ones that are well confined. In fact, if we calculate the delays of the extreme modes, based on the equation

$$\tau = \frac{d\beta}{d\omega} = \frac{n_f}{c} + k_0\frac{dn_f}{d\omega} \tag{5.81}$$

we find that the temporal pulse spread is

$$\Delta\tau = \tau_{high-order} - \tau_{low-order} = \frac{n_{cladding}}{c} + k_0\frac{dn_{cladding}}{d\omega} - \frac{n_{core}}{c} - k_0\frac{dn_{core}}{d\omega} \approx \frac{n_{cladding}}{c} - \frac{n_{core}}{c} \tag{5.82}$$

This calculation yields the same absolute value, but the opposite sign. This equation does in fact give the correct delay difference between the extreme modes, but Eq. 5.80 is the one that gives the practically-useful results for the typical case where the highest-order modes are attenuated to insignificance by waveguide imperfections.

A final point to note about modal dispersion is that it is somewhat mitigated by mode coupling. We will study mode coupling in detail in Chapter 6, but for now we assume that small perturbations will couple energy between the modes of our waveguides. Energy from the extreme modes will therefore continuously be coupled into modes of closer to average group velocity. This coupling has the consequence that instead of being proportional to the propagation length, modal dispersion has a length dependence of $\sqrt{L \cdot L_c}$, where L_c is the characteristic coupling length for the particular coupling mechanism that dominates the waveguide.

5.5.4 Total dispersion – Simultaneous Material, Modal and Waveguide Dispersion

Material and waveguide dispersion act on individual modes, while modal dispersion acts on an ensemble of modes. When we consider a single mode, we need only calculate the material and waveguide dispersion. To first order the pulse broadening due to material and waveguide dispersion can simply be added. We see this by calculating the wavelength derivative of the group delay

$$\tau_g = \frac{d\beta}{d\omega} = \frac{\partial\beta}{\partial k}\frac{dk}{d\omega} + \frac{\partial\beta}{\partial n}\frac{dn}{d\omega} = \frac{n}{c}\frac{\partial\beta}{\partial k_0} + \frac{\omega}{c}\frac{dn}{d\omega} \Rightarrow \tag{5.83}$$

$$\begin{aligned}\frac{d\tau_g}{d\lambda} &= \frac{d}{d\lambda}\left(\frac{n}{c}\frac{\partial\beta}{\partial k_0} + \frac{\omega}{c}\frac{dn}{d\omega}\right) = \frac{1}{c}\frac{d}{d\lambda}\left(n\cdot\frac{\partial\beta}{\partial k_0} - \lambda\frac{dn}{d\lambda}\right) \\ &= \frac{1}{c}\left(\frac{\partial\beta}{\partial k}\frac{dn}{d\lambda} + n\frac{d}{d\lambda}\left(\frac{\partial\beta}{\partial k_0}\right) - \frac{dn}{d\lambda} - \lambda\frac{d^2n}{d\lambda^2}\right) \\ &= \frac{1}{c}\left(\frac{\partial\beta}{\partial k}\frac{dn}{d\lambda} - \frac{dn}{d\lambda} - \frac{k^2}{2\pi}\frac{\partial^2\beta}{\partial k^2} - \lambda\frac{d^2n}{d\lambda^2}\right) \approx \frac{1}{c}\left(-\frac{k^2}{2\pi}\frac{\partial^2\beta}{\partial k^2} - \lambda\frac{d^2n}{d\lambda^2}\right)\end{aligned} \tag{5.84}$$

The approximation we make by adding the material and waveguide dispersion is to neglect the term $(\partial\beta/\partial k - 1)dn/d\lambda$.

Normally the material dispersion dominates over waveguide dispersion, particularly in modern fiber with small index variations (which minimized waveguide dispersion). However, close to the dispersion minimum, which in glass is close to 1.3 μm wavelength, the waveguide dispersion becomes important. Of particular interest is to use the waveguide dispersion to exactly compensate for the material dispersion at a chosen wavelength. Indeed, it is possible to shift the dispersion minimum of optical fiber from 1.3 μm to 1.55 μm, which is the most popular wavelength for fiber optical communication due to the absorption minimum and the existence of efficient, low-noise optical amplifiers in this wavelength range. We will look closer at the principles of waveguide-dispersion engineering after we have developed a detailed model for pulse propagation in the presence of dispersion.

When modal dispersion is present, it tends to dominate the other dispersion mechanisms. In multimode systems, we therefore usually neglect material and waveguide dispersion. In some cases, however, we would like to calculate the total pulse spread caused by all types of dispersion. If we assume that both the single mode dispersion and the modal dispersion both lead to broadened pulses of Gaussian shape, we find that the total pulse broadening is given by

$$\tau_{total} = \sqrt{\left(\tau_{material} + \tau_{waveguide}\right)^2 + \tau_{\text{modal}}^2} \tag{5.85}$$

5.6 Pulse Spreading on Fibers

We will now take a more detailed look at what happens when a pulse of light propagates on a waveguide or fiber. Assume that the pulse has a Gaussian profile in time

$$\vec{E}(x,y,0,t) = u_0(x,y)\,\mathrm{Re}\left[\exp\left(-\alpha\cdot t^2 + j\omega_0 t\right)\right] \tag{5.86}$$

where $u_0(x,y)$ is the transverse profile of the mode. The function $exp(-\alpha t^2)$ represents the envelope of the pulse, while $exp(j\omega t)$ is a rapidly varying optical oscillation.

The envelope can be expressed in terms of its Fourier transform

$$\begin{aligned} F(\Omega) &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\exp\left(-\alpha\cdot t^2\right)\exp(-j\Omega t)dt = \frac{1}{2\sqrt{\alpha\cdot\pi}}\exp\left(-\frac{\Omega^2}{4\alpha}\right) \\ \Rightarrow \exp\left(-\alpha\cdot t^2\right) &= \int_{-\infty}^{\infty}F(\Omega)\exp(j\Omega t)d\Omega \end{aligned} \tag{5.87}$$

Using this equation, the Gaussian pulse can be written

$$\vec{E}(x,y,0,t) = u_0(x,y)\exp(j\omega_0 t)\int_{-\infty}^{\infty}F(\Omega)\exp(j\Omega t)d\Omega \tag{5.88}$$

where we have dropped the explicit taking of the real part. This expression shows that a pulse can be viewed as a sum of harmonics in much the same way that an optical beam can be viewed as a sum of plane waves. In fact, there are strong analogies between dispersion (pulse spreading) and diffraction of propagating optical fields.

To find the field at a different location, we must multiply each frequency component of the pulse with a phase factor, $exp[-j\beta z]$, where the propagation constant is a function of frequency, $\beta(\omega_0+\Omega)$. The description of the pulse then takes the form

$$\vec{E}(x,y,z,t) = u_0(x,y)\int_{-\infty}^{\infty}F(\Omega)\exp\left[j(\omega_0+\Omega)t - \beta\cdot z\right]d\Omega \tag{5.89}$$

This description can be used for pulses with well-defined wave vectors, i.e. pulses of plane waves, “Bessel beams” (this is a class of optical beams that do not change their profile while propagating – just like plane waves, these beams contain infinite energy and cannot be normalized, but are useful as basis functions for expressing physically realizable beam profiles), and guided modes, which is our focus here. For pulses that are also experiencing diffraction (which all physically

realizable beams propagating in free-space do), we must consider this effect in combination with pulse spreading.

In general, the longitudinal wave vector can be a complex function of wavelength in the range of wavelengths contained in the optical pulse, particularly for guided modes, where we must consider both material and waveguide dispersion. In practice, however, optical pulses have a relatively small fractional bandwidth, so we can expand the wave vector in a Taylor series

$$\beta(\omega_0+\Omega)=\beta(\omega_0)+\left.\frac{d\beta}{d\omega}\right|_{\omega_0}\Omega+\frac{1}{2}\left.\frac{d^2\beta}{d\omega^2}\right|_{\omega_0}\Omega^2+\ldots\ldots \tag{5.90}$$

Using this approximation, the pulse is

$$\begin{aligned}&\vec{E}(x,y,z,t)=u_0(x,y)\exp[j(\omega_0 t-\beta_0 z)]\cdot\\&\int_{-\infty}^{\infty}F(\Omega)\exp\left[j\Omega\left(t-\frac{z}{v_g}\right)-j\frac{1}{2}\frac{d}{d\omega}\left(\frac{1}{v_g}\right)\cdot\Omega^2\cdot z\right]d\Omega\end{aligned} \tag{5.91}$$

Substituting the Fourier transform of the pulse, we find

$$\begin{aligned}&\vec{E}(x,y,z,t)=u_0(x,y)\exp[j(\omega_0 t-\beta_0 z)]\cdot\\&\frac{1}{2\sqrt{\alpha\pi}}\int_{-\infty}^{\infty}\exp\left[j\Omega\left(t-\frac{z}{v_g}\right)-\Omega^2\left(\frac{1}{4\alpha}+jz\frac{1}{2}\frac{d}{d\omega}\left(\frac{1}{v_g}\right)\right)\right]d\Omega\end{aligned} \tag{5.92}$$

We see that this expression is identical to the initial Fourier transform of the Gaussian pulse once we make the following substitutions

$$t'=t-\frac{z}{v_g} \tag{5.93}$$

$$\frac{1}{4\alpha'}=\frac{1}{4\alpha}+j\frac{z}{2}\frac{d}{d\omega}\left(\frac{1}{v_g}\right) \tag{5.94}$$

The pulse can then be expressed

$$\vec{E}(x,y,z,t) = u_0(x,y)\exp\left[j(\omega_0 t - \beta_0 z)\right]\cdot$$
$$\sqrt{\frac{1}{1+j2z\alpha\dfrac{d}{d\omega}\left(\dfrac{1}{v_g}\right)}}\cdot\exp\left[-\frac{\left(t-\dfrac{z}{v_g}\right)^2\left(\dfrac{1}{\alpha}-jz2\dfrac{d}{d\omega}\left(\dfrac{1}{v_g}\right)\right)}{\dfrac{1}{\alpha^2}+4\left(\dfrac{d}{d\omega}\left(\dfrac{1}{v_g}\right)z\right)^2}\right] \tag{5.95}$$

Comparing this to the pulse at *z=0* (eq. 5.86), we find that there are three changes: The pulse is displaced, broadened, and has acquired chirp. The displacement is trivial. The pulse is centered at $t=z/v_g$, i.e. it has moved or propagated at the group velocity. This is of course the expected result. The broading and chirp are more interesting and will be discussed in detail

5.6.1 Pulse Broadening

Equation 5.95 shows that the pulse has been broadened while propagating. If we define the pulse width, τ, as the Full-Width-at-Half-Maximum (FWHM) of the intensity of the pulse (the field squared), we find

$$\exp\left[-2\left(\frac{\tau}{2}\right)^2\Bigg/\frac{1}{\alpha}+4\alpha\left(\frac{d}{d\omega}\left(\frac{1}{v_g}\right)z\right)^2\right]=\frac{1}{2}\Rightarrow \tag{5.96}$$

$$\tau=\sqrt{2\ln 2}\sqrt{\frac{1}{\alpha}+4\alpha\cdot z^2\left(\frac{d}{d\omega}\frac{1}{v_g}\right)^2} \tag{5.97}$$

We see that the broadening of the pulse is a function of the derivative of the group delay with respect to frequency. This is exactly what we would expect from our earlier treatment.

Observing that the initial pulse width is $\tau_0=\sqrt{2\ln 2/\alpha}$, we can express the pulse width as

$$\tau=\tau_0\sqrt{1+\left(\frac{4z\ln 2}{\tau_0^2}\frac{d}{d\omega}\left(\frac{1}{v_g}\right)\right)^2} \tag{5.98}$$

At large distances this simplifies to

$$\tau = z\frac{4\ln 2}{\tau_0}\frac{d}{d\omega}\left(\frac{1}{v_g}\right) \tag{5.99}$$

In terms of the earlier-defined dispersion parameter, $D = -\frac{2\pi \cdot c}{\lambda^2}\left(\frac{d^2\beta}{d\omega^2}\right)$, the pulse width can be expressed

$$\tau = \tau_0\sqrt{1+\left(\frac{2\ln 2}{\pi \cdot c}\frac{D\lambda^2}{\tau_0^2}z\right)^2} \tag{5.100}$$

In general, the wave vector of a mode is a function of the materials of the core and cladding of the guide, and the wavelength of the light. The functional relationships depend strongly on the geometry of the guide. Expressing the wavevector in terms of the mode index, we can write

$$\beta = nk_0 = n(n_{core}, n_{clad}, \omega)\frac{\omega}{c} \tag{5.101}$$

In a weakly guiding fiber, i.e. a fiber in which $n_{core} \approx n_{clad}$, we assume that

$$\frac{\partial n_{core}}{\partial \omega} \approx \frac{\partial n_{clad}}{\partial \omega} \equiv \left.\frac{\partial n}{\partial \omega}\right|_{material} \tag{5.102}$$

This is a reasonable assumption if the core and cladding materials are essentially the same, with only a small index difference caused by impurity doping in the core. With this assumption, the group delay simplifies to

$$\tau_g = \frac{1}{v_g} = \frac{d\beta}{d\omega} = \frac{\partial\beta}{\partial k}\frac{dk}{d\omega} + \frac{\partial\beta}{\partial n}\frac{dn}{d\omega} = \frac{n}{c}\frac{\partial\beta}{\partial k_0} + \frac{\omega}{c}\frac{dn}{d\omega} \tag{5.103}$$

To find the dispersion, we take the derivative with respect to the wavelength

$$\begin{aligned}\frac{d\tau_g}{d\lambda} &= \frac{d}{d\lambda}\left(\frac{n}{c}\frac{\partial\beta}{\partial k_0} + \frac{\omega}{c}\frac{dn}{d\omega}\right) = \frac{1}{c}\frac{d}{d\lambda}\left(n\cdot\frac{\partial\beta}{\partial k_0} - \lambda\frac{dn}{d\lambda}\right)\\ &\approx \frac{1}{c}\left(-\frac{k^2}{2\pi}\frac{\partial\beta^2}{\partial^2 k} - \lambda\frac{d^2n}{d\lambda^2}\right)\end{aligned} \tag{5.104}$$

The conclusion is that for weakly guiding waveguides, we can simply add the material and waveguide dispersion. As discussed before, we can utilize this phe-

nomenon to create dispersion shifted and dispersion flattened fibers. This is conceptually illustrated in Fig. 5.20.

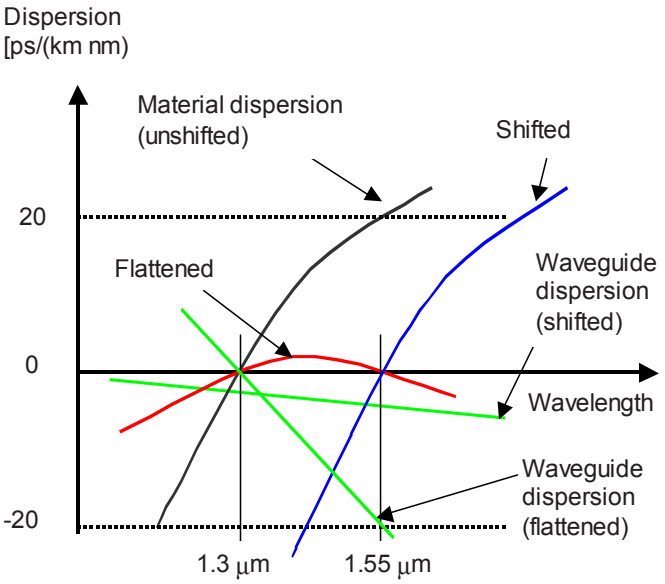


Figure 5.20. *Schematic graphs of group-velocity dispersion on weakly guiding fibers. The total dispersion of a single mode is the sum of material and waveguide dispersion. By correctly engineering the waveguide dispersion, dispersion shifted and dispersion-flattened fibers can be designed.*

5.6.2 Frequency Chirp

Going back to our expression for the Gaussian pulse after propagation

$$\vec{E}(x,y,z,t) = u_0(x,y)\exp\left[j(\omega_0 t - \beta_0 z)\right]\cdot \sqrt{\frac{1}{1+j2z\alpha\dfrac{d}{d\omega}\left(\dfrac{1}{v_g}\right)}}\cdot\exp\left[-\frac{\left(t-\dfrac{z}{v_g}\right)^2\left(\dfrac{1}{\alpha}-jz2\dfrac{d}{d\omega}\left(\dfrac{1}{v_g}\right)\right)}{\dfrac{1}{\alpha^2}+4\left(\dfrac{d}{d\omega}\left(\dfrac{1}{v_g}\right)z\right)^2}\right] \tag{5.105}$$

we see that in addition to the spatial off set and the pulse broadening, we also have a z-dependent phase term in the expression. The total phase is indeed given by

$$\phi(z) = \omega_0 t - \beta_0 z + \frac{\left(t - \frac{z}{v_g}\right)^2 \left(2\frac{d}{d\omega}\left(\frac{1}{v_g}\right)\right)}{\frac{1}{\alpha^2} + 4\left(\frac{d}{d\omega}\left(\frac{1}{v_g}\right)z\right)^2} z \tag{5.106}$$

The instantaneous frequency of the pulse is

$$\omega(z) = \frac{d}{dt}\phi(z) = \omega_0 + \frac{4\frac{d}{d\omega}\left(\frac{1}{v_g}\right)\cdot z}{\frac{1}{\alpha^2} + 4\left(\frac{d}{d\omega}\left(\frac{1}{v_g}\right)z\right)^2}\left(t - \frac{z}{v_g}\right) \tag{5.107}$$

We see that the pulse does not have a uniform instantaneous frequency after propagation, but that the frequency varies throughout the pulse. A pulse with a varying frequency is said to be chirped. Positive dispersion results in a pulse that has lower frequencies (i.e. longer wavelengths, red shift) at its leading edge, and higher frequencies at its trailing edge. This is illustrated in Fig. 5.21.

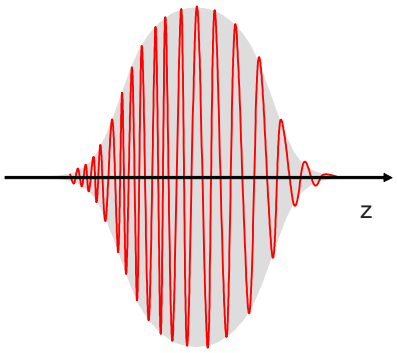


Figure 5.21. Schematic of chirped pulse on fiber with positive dispersion.

5.6.3 Dispersion Compensation

Our treatment of dispersion shows that it is a reversible process, i.e. the pulse broadening caused by positive dispersion can be undone by negative dispersion. This type of dispersion compensation is becoming increasingly common in fiber optic communication systems.

Typically, the installed fiber will have a specific dispersion at the wavelength in question, and the fiber that is used to "undo" the dispersion is placed in a switching center or other service building. The compensating fiber is therefore not contributing to transmission of the signals. Some networks have alternating lengths of fiber with opposite dispersion. This is preferable to simply using low dispersion, or dispersion flattened fiber, because dispersion helps mitigate non-linear optical effects that create cross talk in WDM channels.

Alternatively, dispersion can be controlled by spectral filtering (or spectral phase delay) as in the Heritage-Weiner[2] pulse compressor of Fig. 5.22 (see Chapter 13 for further discussion of this device).

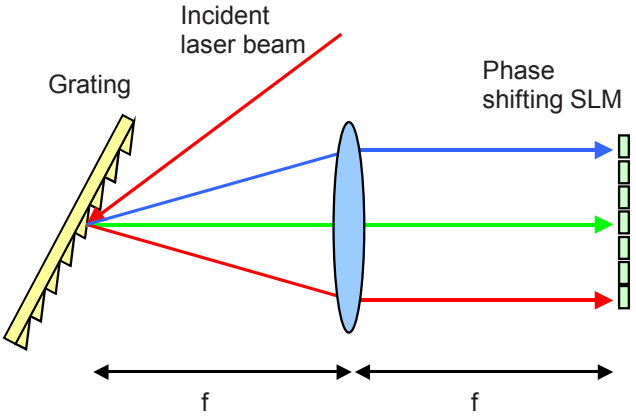


Figure 5.22. Schematic drawing of the Heritage-Weiner femto-second pulse compressor. The short-pulse (broad band) laser beam is dispersed on an array of phase shifters that can be programmed to compensate for the dispersion that the pulse has suffered.

5.6.4 Dispersion Expressed in Normalized Propagation Parameters

Finding the dispersion relationship is tedious in any of the waveguide structures we have studied so far, and those are the simple ones! We must solve the characteristic equation for the longitudinal wave vector for a range of wavelengths or frequencies using graphical or numerical means. Once that is done, we can find the appropriate derivatives. This procedure must be repeated for every mode we are interested in. Once found, the dispersion relation of a waveguide structure can be depicted in several ways, but the most common is to plot the normalized index

$$b = \frac{\left(\frac{\beta}{k_0}\right)^2 - n_{clad}^2}{n_{core}^2 - n_{clad}^2} = \frac{n_{eff}^2 - n_{clad}^2}{n_{core}^2 - n_{clad}^2} \tag{5.108}$$

vs. the normalized frequency $V = k_0 h\sqrt{n_{core}^2 - n_{clad}^2}$.

Earlier we showed how we could generate such a plot for the asymmetric waveguide by rewriting the characteristic equation in terms of normalized parameters. Similar plots for particular optical waveguides and fibers can be generated by finding solutions to the characteristic equations for these structures. These plots of normalized index vs. normalized frequency contain the dispersion relation of the waveguide.

To see how, consider the group delay

$$\tau_g = \frac{d\beta}{d\omega} = \frac{d\beta}{dk_0}\frac{dk_0}{d\omega} = \frac{1}{c}\frac{d\beta}{dk_0} = \frac{1}{c}\frac{d\beta}{dV}\frac{dV}{dk_0} \tag{5.109}$$

The definition of the normalized frequency results in

$$\frac{dV}{dk_0} = h\sqrt{n_{core}^2 - n_{clad}^2} = \frac{V}{k_0} \Rightarrow \tag{5.110}$$

$$\tau_g = \frac{1}{c}\frac{V}{k_0}\frac{d\beta}{dV} \tag{5.111}$$

The definition of the normalized index allows us to write

$$\beta^2 = bk_0^2\left(n_{core}^2 - n_{clad}^2\right) + k_0^2 n_{clad}^2 = k_0^2 n_{clad}^2 + k_0^2 n_{clad}^2 \cdot b\left(\frac{n_{core}^2 - n_{clad}^2}{n_{clad}^2}\right) \Rightarrow \tag{5.112}$$

$$\begin{aligned} &\beta^2 \approx k_0^2 n_{clad}^2 + k_0^2 n_{clad}^2 \cdot 2b\left(\frac{n_{core} - n_{clad}}{n_{clad}}\right) = \\ &k_0^2 n_{clad}^2 \cdot \left(1 + 2b\frac{n_{core} - n_{clad}}{n_{clad}}\right) \Rightarrow \end{aligned} \tag{5.113}$$

$$\beta^2 \approx k_0^2 n_{clad}^2 \cdot \left(1 + 2b\Delta\right) \tag{5.114}$$

where

$$\Delta = \frac{n_{core} - n_{clad}}{n_{clad}} \tag{5.115}$$

Practical fibers have low index differences so we can to a good approximation write the longitudinal wave vector as a binomial expansion

$$\beta \approx k_0 n_{clad} \sqrt{1 + 2b\Delta} \approx k_0 n_{clad} \left(1 + b\Delta\right) \tag{5.116}$$

Using this approximation we find the following expression for the group delay

$$\tau_g = \frac{1}{c}\frac{V}{k_0}\frac{d\beta}{dV} = \frac{1}{c}\frac{V}{k_0}\frac{d}{dV}\left(k_0 n_{clad}\left(1 + b\Delta\right)\right) \tag{5.117}$$

Ignoring the weak wavelength dependence of Δ, this evaluates to

$$\begin{aligned}
\tau_g &= \frac{1}{c}\frac{V}{k_0}\frac{d\left(k_0 n_{clad}\right)}{dV} + \frac{1}{c}\frac{V}{k_0}\frac{d\left(k_0 n_{clad} b\Delta\right)}{dV} \\
&= \left(\frac{n_{clad}}{c} + \frac{V}{c}\frac{dn_{clad}}{dV}\right) + \frac{1}{c}\frac{V}{k_0}\frac{d\left(k_0 n_{clad} b\Delta\right)}{dV} \\
&= \frac{1}{c}\left(n_{clad} + k_0 \frac{dn_{clad}}{dk_0}\right) + \frac{1}{c}\frac{V}{k_0}\frac{d\left(k_0 n_{clad} b\Delta\right)}{dV}
\end{aligned} \tag{5.118}$$

The first term describes the material dispersion that we have discussed before, and the second term gives the waveguide dispersion in terms of the normalized parameters.

Using the approximation

$$V = k_0 a \sqrt{n_{core}^2 - n_{clad}^2} \approx k_0 h \cdot n_{clad} \sqrt{2\Delta} \tag{5.119}$$

the last term (waveguide dispersion) can be rewritten as

$$\begin{aligned}
\tau_{guide} &= \frac{1}{c}\frac{V}{k_0}\frac{d\left(k_0 n_{clad} b\Delta\right)}{dV} = \frac{1}{c}\frac{V}{k_0}\sqrt{\frac{\Delta}{2}}\frac{d\left(k_0 n_{clad} b\sqrt{2\Delta}\right)}{dV} \\
&= \frac{1}{c}\frac{V}{hk_0}\sqrt{\frac{\Delta}{2}}\frac{d\left(Vb\right)}{dV} = \frac{1}{c} n_{clad} \Delta \frac{d\left(Vb\right)}{dV}
\end{aligned} \tag{5.120}$$

Expressions for *d(Vb)/dV* can be found analytically for the LP modes and numerically for the exact fiber modes. Some modes (LP_{0n}) have zero waveguide-based group delay close to cut-off. These modes tend to have large losses due to bend etc. so they play no significant role in practical fiber optical communications.

Away from cut-off the dispersion is a function of the azimuthal mode number as demonstrated by Marcuse

$$\frac{d(Vb)}{dV} = \frac{2(\nu - 1)}{\nu} \tag{5.121}$$

Considering only modes with $\nu > 1$ (modes with $n=0$, and $n=1$ have extreme modal delays, but, as noted above, tend to be strongly attenuated close to cut-off), we find the following modal dispersion

$$\begin{aligned} \Delta\tau_{guide} &= \frac{1}{c} n_{clad} \Delta \left[\left(\frac{d(Vb)}{dV} \right)_{slow} - \left(\frac{d(Vb)}{dV} \right)_{fast} \right] \\ &= \frac{1}{c} (n_{core} - n_{clad}) \left[\frac{2(\nu_{\max} - 1)}{\nu_{\max}} - \frac{2(2-1)}{2} \right] \\ &= \frac{1}{c} (n_{core} - n_{clad}) \left[1 - \frac{2}{\nu_{\max}} \right] = \frac{1}{c} (n_{core} - n_{clad}) \left[1 - \frac{\pi}{V} \right] \end{aligned} \tag{5.122}$$

We have used the relation $\nu_{max}=2V/\pi$, to derive the last expression. In the large V-number limit, this is identical to the simplified expression for modal dispersion we found earlier.

5.6.5 Single-Mode Dispersion Expressed in Normalized Parameters

The dispersion for a single mode is given by

$$d\tau_{guide} = \frac{1}{c} n_{clad} \Delta \frac{d(Vb)}{dV} dV \tag{5.123}$$

Substituting the V-number definition and $\frac{d\lambda}{\lambda} = -\frac{dk_0}{k_0}$, we find

$$\begin{aligned} \frac{d\tau_{guide}}{d\lambda} &= \frac{d\tau_{guide}}{dV} \left(-\frac{ak_0 \sqrt{n_{core}^2 - n_{clad}^2}}{\lambda} \right) \\ &= \left(-\frac{ak_0 \sqrt{n_{core}^2 - n_{clad}^2}}{\lambda} \right) \frac{1}{c} (n_{core} - n_{clad}) \frac{d^2(Vb)}{dV^2} \\ &= -\frac{n_{core} - n_{clad}}{\lambda \cdot c} V \frac{d^2(Vb)}{dV^2} \end{aligned} \tag{5.124}$$

where again we have neglected the dependence of Δ on wavelength.

To evaluate the dispersion for single mode fibers we use Gloge's expression for the LP_{01} mode

$$\kappa \cdot a = \frac{\left(1+\sqrt{2}\right)V}{1+\left(4+V^4\right)^{0.25}} \tag{5.125}$$

together with the definition $b = 1 - \left(\frac{\kappa \cdot a}{V}\right)^2$ to plot the normalized index (b) and it's derivatives ($d(Vb)/dV$ and $Vd^2(Vb)/dV^2$). The results are shown in Figure 7.5.

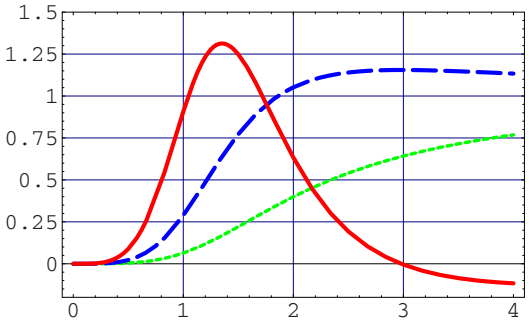


Figure 5.24. Normalized index (dotted), group delay term (d(Vb)/dV - dashed), and dispersion term (Vd2(Vb)/dV2 - solid) of the LP_{01} mode.

At low frequency the fundamental mode is not well confined, and the normalized index, group delay, and dispersion are all zero. At high frequencies, the mode is well confined, and the normalized index approaches unity, while the dispersion term approaches zero. The waveguide dispersion is maximized close to where the slope of the group delay has its maximum.

The LP modes are good approximations to the exact modes of the cylindrical waveguide, but real optical fibers have more complex index profiles, for which we cannot find analytical solutions or even good analytical approximations. To find the dispersion of practical fibers, we must therefore resort to numerical calculations of the wave vector as a function of wavelength.

5.6.6 Single Mode Fiber Design

To ensure single mode operation of a fiber, we must have

$$V = k_0 a \sqrt{n_{core}^2 - n_{clad}^2} < 2.405 \tag{5.126}$$

This requirement can be met by a small core size or by a small index difference. A small core size complicates coupling, termination, and splicing of the fiber, while a small index difference leads to losses because the fundamental mode is not well confined. A compromise between practical fiber handling and loss must therefore be found.

The optical-fiber designer also must also carefully consider the dopants used to create the index profile. Impurities that led to excess absorption are unacceptable. Modern fibers have Germanium doped cores to (slightly) increase the index over that of pure SiO_2. (Doping here means adding some GeO_2 to the SiO_2).

To have the design flexibility to optimize mode-field diameter, confinement, and dispersion while maintaining single-mode operation, fiber with slightly more complex index profiles than the step-index fiber we have studied, have been developed. One example is the so-called W-profile fiber. The exact wavevectors, mode diameters, and dispersion of these fibers must be found numerically.

5.7 Fiber Calculation Example

Consider a single-mode, step-index fiber with the following characteristics: $a = 4$ μm, $n_{core} = 1.45$ and $n_{cladding} = 1.446$.

a) What is the normalized frequency at 1550 nm wavelength?

$$V = \frac{2\pi}{\lambda} a\sqrt{n_{core}^2 - n_{clad}^2} = \frac{2\pi}{1.55\cdot 10^{-6}} \cdot 4.0\cdot 10^{-6}\sqrt{1.45^2 - 1.446^2} = 1.75 \qquad (5.127)$$

b) How many guided modes does the fiber support?

Two degenerate HE_{11} modes.

c) What is the waveguide dispersion, D_{guide}, for this fiber?

$$D_{total} = \frac{d\tau_g}{d\lambda} \approx \frac{1}{c}\left(-\frac{k^2}{2\pi}\frac{\partial^2\beta}{\partial k^2} - \lambda\frac{d^2 n}{d\lambda^2}\right) \Rightarrow \qquad (5.128)$$

$$D_{guide} = \left.\frac{d\tau_g}{d\lambda}\right|_{guide} = -\frac{1}{c}\frac{k^2}{2\pi}\frac{\partial^2\beta}{\partial k^2} = -\frac{n_{core} - n_{clad}}{c\lambda}\left[V\frac{d^2(Vb)}{dV^2}\right] \qquad (5.129)$$

From the Fig. 5.24 we see that the expression in square-parenthesis is approximately 0.97, so

$$D = -\frac{0.004}{3\cdot 10^8 m/s \cdot 1550nm}\cdot 0.97 = -8.3\, ps/(km\cdot nm) \tag{5.130}$$

d) A transform-limited, Gaussian pulse with a FWHM width of 10 ps is launched on the fiber. What is the pulse length after it has propagated 100 km on the fiber? Assume that the material dispersion at 1.55 nm is 10ps/(km·nm) for both the core and cladding materials.

The total dispersion is 10 ps/(km·nm) - 8.3ps/(km·nm) = 1.7 ps/(km·nm). The width of the pulse in wavelength space is

$$\Delta\lambda = \frac{4\ln 2}{\tau_0}\cdot\frac{\lambda^2}{2\pi\cdot c} = \frac{4\ln 2}{10\cdot 10^{-12} s}\cdot\frac{\left(1550\cdot 10^{-9} m\right)^2}{2\pi\cdot 3\cdot 10^8 m/s} = 0.35\cdot 10^{-9} m \tag{5.131}$$

$$\tau = \sqrt{\tau_0^2 + \left(\Delta\lambda\cdot D\cdot z\right)^2} = \sqrt{(10)^2 + \left(0.35\cdot 1.7\cdot 100\right)^2}\cdot 10^{-12} s = 60.3\cdot 10^{-12} s \tag{5.132}$$

e) Design a fiber to compensate for the dispersion caused by propagation as described in d). The compensating fiber should also be a step-index fiber of the same material, but may have a different core size.

As for any design problem, this does not have a unique solution. The approach is to shrink the core radius until waveguide dispersion has a larger absolute value than the material dispersion. This happens for a range of V-values around 1.2 – 1.3. Once the fiber parameters are chosen, the length of fiber must be adjusted such that the net dispersion of the two fibers in series is zero.

5.8 Summary of Fibers and Waveguides

Fiber optics forms the infrastructure for all optical communication devices, including those that are implemented in Optical MEMS and Nanophotonics. This chapter lays the foundation for understanding of guided-wave and optical-fiber components. The first part of the chapter is devoted to building an intuitive understanding of guided-wave physics through a detailed derivation and discussion of slab two-dimensional) waveguides. The second part of the chapter extends, without detailed derivations, the slab-waveguide picture to cylindrical waveguides or optical fibers. The basic characteristics, as well as the important technological features, of modern optical fibers are described. The last part of the chapter is focused on dispersion (wavelength dependence) of wave propagation on optical fibers. Here we use make use of a formalism very similar to the Gaussian-

Beam theory we used in Chapter 4 to described spatial diffraction to model the temporal behavior of guided-wave optical pulses. Our Gaussian-pulse model allows us to derive compact analytical formulas for pulse spreading on optical fibers. The most important mathematical models used in waveguide calculations are summarized in the following.

The total field of a TE guided mode on a slab waveguide can be written

$$E_y(x,z)=\begin{cases} A\, e^{-\gamma_c x}\cdot \mathrm{e}^{-\mathrm{i}\beta z} & 0<x \\ A\left[\cos(\kappa_f x)-\dfrac{\gamma_c}{\kappa_f}\sin(\kappa_f x)\right]\cdot \mathrm{e}^{-\mathrm{i}\beta z} & -\mathrm{h}<x<0 \\ A\left[\cos(\kappa_f h)+\dfrac{\gamma_c}{\kappa_f}\sin(\kappa_f h)\right]e^{\gamma_s(x+h)}\cdot \mathrm{e}^{-\mathrm{i}\beta z} & x<-h \end{cases} \tag{5.133}$$

where β is the longitudinal wave vector, $\kappa_f = k_0^2 n_i^2 - \beta^2$, is the ***transversal wave vector***, and $\gamma_i = \beta^2 - k_0^2 n_i^2$ are the ***attenuation coefficients***.

The transversal wavevector is determined by the eigenvalue equation

$$\tan(h\kappa_f)=\frac{\gamma_c+\gamma_s}{\kappa_f\left(1-\dfrac{\gamma_c\gamma_s}{\kappa_f^2}\right)} \qquad \text{(TE)} \tag{5.134}$$

$$\tan(h\kappa_f)=\frac{\kappa_f\left(\dfrac{n_f^2}{n_s^2}\gamma_s+\dfrac{n_f^2}{n_c^2}\gamma_c\right)}{\kappa_f^2-\dfrac{n_f^4}{n_s^2 n_c^2}\gamma_s\gamma_s} \qquad \text{(TM)} \tag{5.135}$$

These equations must be must be solved numerically or graphically to find the eigenvalues for β_{TE} and β_{TM}, which define the guided modes of the slab waveguide

The planewave picture of the guided modes leads to the dispersion equation

$$2kn_f h\cos\theta-\Phi_c-\Phi_s=2\pi\cdot m \tag{5.136}$$

where m is an integer, and Φ_c and Φ_s are the phase shifts of TIR at the cladding and substrate respectively. Graphical solutions show that as the ratio h/λ increases, the number of guided modes supported by the guide also increases. For sufficiently small h/λ ratios, a symmetric guide will have only one mode while an

asymmetric guide will support no guided modes. For a given h/λ ratio, there is a finite set of guided modes and a continuous, infinite set of radiation modes.

The important properties of modes are:

1. Each eigenvalue of the longitudinal wavevector, β, corresponds to one unique mode (field distribution), or a unique set of degenerate modes.
2. There is a finite number of guided modes.
3. Most modes are not guided. These are the radiation modes.
4. The modes form an orthogonal, complete set, which means that any field profile can be written as a superposition of modes

$$A(x,y,z) = \sum_{guided} a_i A_i(x,y,z) + \int_{radiation} a(\beta)\cdot A(x,y,z,\beta)d\beta \tag{5.137}$$

where $A(x,y,z)$ can be the E or the H field.

Guided Modes on Step-Index Fibers - LP modes

The modes on weakly guiding fibers (n_{core}~n_{clad}) can be approximated as linearly polarized modes, which are degenerate combination of the exact modes.

The Fundamental Mode of a Cylindrical Waveguide

The HE_{11} mode of the cylindrical waveguide is guided for all wavelengths. **When the V-number is less than 2.405, the HE_{11} mode is the only guided mode!**
The following Gaussian approximation to the HE_{11} mode is sufficiently accurate for most fiber calculations:

$$\vec{E}(r) = E_x \exp\left[-\left(\frac{r}{\omega}\right)^2\right] \tag{5.138}$$

where

$$\frac{\omega}{a} = 0.65 + 1.619\cdot V^{-1.5} + 2.87\cdot V^{-6} \tag{5.139}$$

Dispersion in homogeneous media:

In a homogeneous medium, the group velocity, group delay, and group index can be expressed

$$v_g = \frac{d\omega}{dk} = \left[\frac{dk}{d\omega}\right]^{-1} = \left[\frac{d}{d\omega}\left(\frac{\omega n}{c}\right)\right]^{-1} = \frac{c}{n - \lambda\dfrac{dn}{d\lambda}} \tag{5.140}$$

$$\tau_g = \frac{d\left(\frac{n\omega}{c}\right)}{d\omega} = \frac{n}{c} + \frac{\omega}{c}\frac{dn}{d\omega} = \frac{1}{c}\left(n + \omega\frac{dn}{d\omega}\right) \tag{5.141}$$

$$N_g = \frac{c}{v_g} = c \cdot \tau_g \Rightarrow N_g = n + \omega\frac{dn}{d\omega} - = n - \lambda\frac{dn}{d\lambda} \tag{5.142}$$

The derivative $dn/d\lambda$ is negative (normal dispersion) in most materials of interest for optical waveguides. The spread of a pulse of bandwidth $\Delta\lambda$ traveling a distance L is

$$\Delta\tau = -L \cdot \Delta\lambda \cdot \frac{\lambda}{c}\frac{d^2 n}{d\lambda^2} = L \cdot \Delta\lambda \cdot D \tag{5.143}$$

where D is the material dispersion.

Waveguide Dispersion

The group delays and pulse spreads of guided modes in media without material dispersion are

$$\tau_g = \frac{d\beta}{d\omega} = \frac{1}{c}\frac{\partial\beta}{\partial k} \tag{5.144}$$

$$\frac{d\tau_g}{d\lambda} = \frac{1}{c}\frac{d}{d\lambda}\left(\frac{\partial\beta}{\partial k}\right) = -\frac{k^2}{2\pi \cdot c}\frac{\partial\beta^2}{\partial^2 k} \tag{5.145}$$

Modal Dispersion

The longitudinal wave vector of the lowest order mode of a waveguide is approximately equal to the wavevector in the core material, and the highest-order guided mode of practical interest can be though of as propagating at the TIR angle, so the maximum delay difference is

$$\Delta\tau \approx \frac{n_{core}}{c} - \frac{n_{cladding}}{c} \tag{5.146}$$

Modal dispersion is mitigated by mode coupling, which means that instead of being proportional to propagation length, modal dispersion has a length dependence of $\sqrt{L \cdot L_c}$, where L_c is the characteristic coupling length.

Total Dispersion

Assuming that single mode dispersion and modal dispersion both lead to broadened pulses of Gaussian shape, the total pulse broadening is

$$\tau_{total} = \sqrt{\left(\tau_{material} + \tau_{waveguide}\right)^2 + \tau_{\mathrm{mod}al}^2} \tag{5.147}$$

Gaussian Pulse Propagation

A Gaussian pulse of the form

$$\vec{E}(x,y,0,t) = u_0(x,y)\mathrm{Re}\left[\exp\left(-\alpha \cdot t^2 + j\omega_0 t\right)\right] \tag{5.148}$$

has an intensity FWHW of $\tau_0 = \sqrt{\dfrac{2\ln 2}{\alpha}}$, and its spectrum is given by its Fourier transform

$$F(\Omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\left(-\alpha \cdot t^2\right)\exp(-j\Omega t)dt = \frac{1}{2\sqrt{\alpha \cdot \pi}}\exp\left(-\frac{\Omega^2}{4\alpha}\right) \tag{5.149}$$

$$\Rightarrow \vec{E}(x,y,0,t) = u_0(x,y)\exp(j\omega_0 t)\int_{-\infty}^{\infty} F(\Omega)\exp(j\Omega t)d\Omega \tag{5.150}$$

It follows that the intensity FWHM of the spectrum is

$$\Omega_0 = 2\sqrt{\alpha \cdot 2\ln 2} = \frac{4\ln 2}{\tau_0} \tag{5.151}$$

After propagation through a distance z, the pulse takes the form

$$\vec{E}(x,y,z,t) = u_0(x,y)\int_{-\infty}^{\infty} F(\Omega)\exp\left[j(\omega_0 + \Omega)t - \beta \cdot z\right]d\Omega \tag{5.152}$$

The longitudinal wave vector is a function of wavelength, but in practice optical pulses have relatively small fractional bandwidths, so we neglect higher-than-quadratic terms and obtain

$$\vec{E}(x,y,z,t) = u_0(x,y)\exp[j(\omega_0 t - \beta_0 z)]\cdot$$
$$\sqrt{\frac{1}{1+j2z\alpha\dfrac{d}{d\omega}\left(\dfrac{1}{v_g}\right)}}\cdot\exp\left[-\frac{\left(t-\dfrac{z}{v_g}\right)^2\left(\dfrac{1}{\alpha}-jz2\dfrac{d}{d\omega}\left(\dfrac{1}{v_g}\right)\right)}{\dfrac{1}{\alpha^2}+4\left(\dfrac{d}{d\omega}\left(\dfrac{1}{v_g}\right)z\right)^2}\right] \tag{5.153}$$

Comparing this to the pulse at *z=0*, we see that the pulse is displaced by z/v_g, it is broadened, and it is chirped. The broadening can be expressed

$$\tau = \tau_0\sqrt{1+\left(\frac{4z\ln 2}{\tau_0^2}\frac{d}{d\omega}\left(\frac{1}{v_g}\right)\right)^2} = \tau_0\sqrt{1+\left(\frac{2\ln 2}{\pi\cdot c}\frac{D\lambda^2}{\tau_0^2}z\right)^2} \tag{5.154}$$

where $\tau_0 = \sqrt{2\ln 2/\alpha}$ is the FWHM of the transformed-limited pulse, and $D = -\dfrac{2\pi\cdot c}{\lambda^2}\left(\dfrac{d^2\beta}{d\omega^2}\right)$ is the dispersion. The pulse has acquired a frequency chirp that is manifest in the expression for the instantaneous frequency of the pulse

$$\omega(z) = \frac{d}{dt}\phi(z) = \omega_0 + \frac{4\dfrac{d}{d\omega}\left(\dfrac{1}{v_g}\right)\cdot z}{\dfrac{1}{\alpha^2}+4\left(\dfrac{d}{d\omega}\left(\dfrac{1}{v_g}\right)z\right)^2}\left(t-\frac{z}{v_g}\right) \tag{5.155}$$

The frequency varies throughout the pulse.

Further Reading

C.R. Pollock, Fundamentals of Optoelectronics", Richard D. Irwin, inc. 1995.
J.C. Palais, "Fiber Optic Communication", 5th edition, Prentice Hall, 2005.

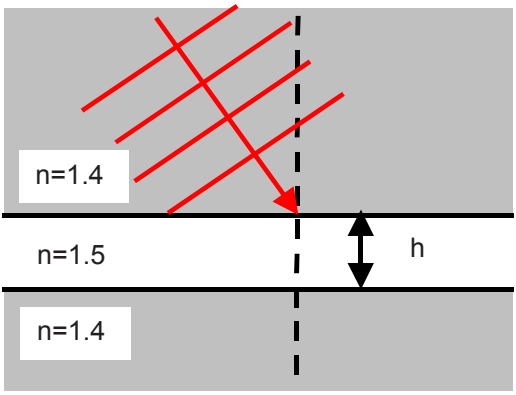
Exercises

Problem 5.1 – Boundary Conditions

a. What are the boundary conditions for the *E* field of a *TE* mode at the film/cladding interface in a symmetric slab waveguide?
b. What are the boundary conditions for the *E* field of a *TM* mode at the film/cladding interface in a symmetric slab waveguide?
c. Is it possible to have an asymmetric field distribution on a symmetric slab waveguide? Explain.

Problem 5.2 – Slab Waveguide

Consider the structure shown below.



Plane wave incident on high-index film.

a. How many guided modes will the structure support ($h=3\lambda$)?
b. At what incident angle, if any, will the plane wave be phase matched to the fundamental TM mode of the waveguide ($h=3\lambda$)?

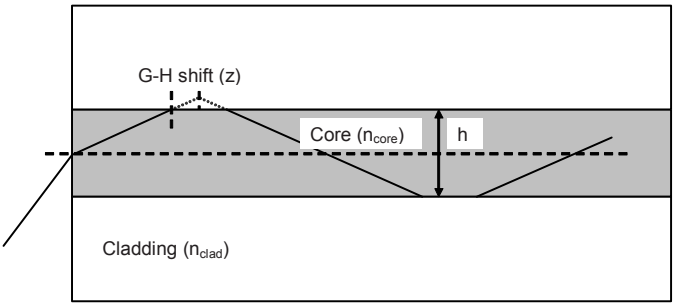
Problem 5.3 - Modal Dispersion and the Goos-Hanchen Shift

The modal dispersion on multimode waveguides are sometimes erroneously modeled as being caused by the difference in path length of a ray propagating along the axis of the waveguide and a ray that is propagating at the TIR angle.

a. Using this overly simplified model, what is the maximum difference in time delay per unit of length of a step -index optical waveguide in terms of its core thickness, core index, cladding index, and the wavelength of the propagating light?

b. Calculate the maximum difference in time delay per unit of length of the same optical waveguide when the effects of the G-H shift are included. Assume that the thickness, *h*, equals 50 wavelengths. Explain how you model the path of the ray in the cladding, and why you chose this model.

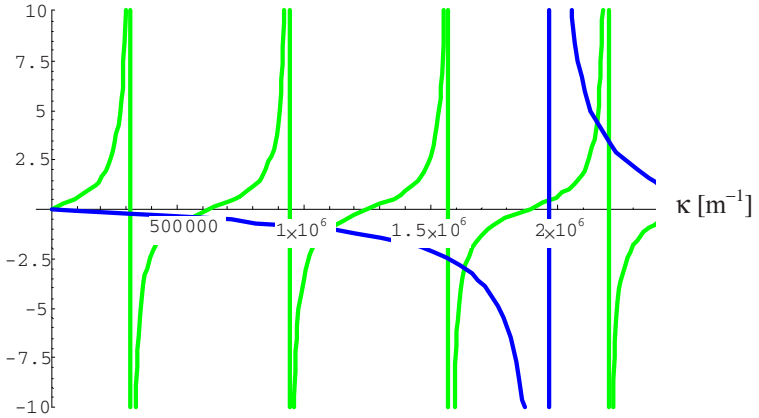
c. Compare the models of a) and b) by plotting the delay through the waveguide as a function of the angle the ray makes with the axis from zero to the TIR angle for a waveguide with the following parameters: $h=50\ \lambda$, $n_{core}=1.5$, *and* $n_{clad}=1.45$. Based on these plots, comment on the physical realism of the two models. Are there angle ranges for which one or the other model is unphysical?



Step-index waveguide.

Problem 5.4 – Waveguide Modes:

The figure shows a graphical solution of the characteristic equation for the TE modes of a dielectric slab waveguide of thickness *5 micron* at *1 micron* wavelength. The refractive index of the core is *1.5*.



a. What is a mode? Explain in terms that can be understood by people not trained in physics or engineering.
b. How many guided-TE modes does this waveguide support? How many TM modes? Label the TE modes on the graph.
c. Sketch the transversal mode profile of the three lowest order TE modes.
d. Use the graph to find the value of the longitudinal wavevector for the highest order mode.
e. Is this a symmetric waveguide? Explain.
f. Use the graph to find the *V*-number of the waveguide.
g. What happens to the modes of the waveguide as the wavelength increases?

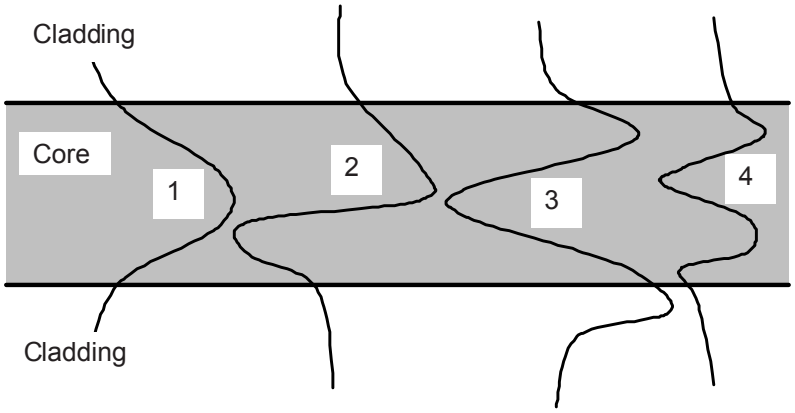
Problem 5.5 - Confinement of Waveguide Modes

Consider a symmetric slab waveguide with a core index of 1.5 and cladding index of 1.45.

a. What thickness of the core results in the smallest width (FWHM) of the lowest order *TE* mode? Find the solution graphically by plotting the FWHM (normalized to λ) as a function of the waveguide thickness (also normalized to λ).
b. What thickness of the core results in the smallest group velocity of the lowest order *TE* mode?
c. What thickness of the core results in the smallest waveguide dispersion of the lowest order TE mode?

Problem 5.6 – Combinations of Modes

a. Which of the profiles in the figure represent realizable guided modes? Explain.



The cladding regions have uniform indices, while the index may vary across the core.

b. We measure the intensity distribution on a waveguide as a function of length (in practice we would do this by cutting the waveguide back and measure each cross section), and we find a periodic variation. There are no backwards propagating waves on the guide. Explain what is going on. What is the significance of the period of the intensity fluctuations?

Problem 5.7 – Single-Mode Optical Fibers

a. Explain what it means that a fiber is "single-mode". How must a fiber be designed to have this property?
b. What is the main advantage of single-mode fiber over multimode fiber? (Explain)

Problem 5.8 – Single-Mode Operation

We have a step-index fiber with the following parameters: Core index: $n_{core}=1.45$, cladding index: $n_{clad}=1.446$, core radius: $a=5\ um$.

a. Is the fiber single-mode at 1.3 um wavelength? (Explain)

We are designing a fiber for single mode operation at 1.55 μm wavelength. The core index is 1.45, and the cladding index is 1.446.

b. How large can we make the core radius and still have the fiber be single-mode (two-mode if you consider the two degenerate polarization modes)?

Problem 5.9 – Mode Radius vs. Core Radius

Consider an step-index, optical fiber with $n_{cladding}=1.445$ and $n_{core}=1.45$.

a. Plot the mode radius (normalized to the wavelength) of the fundamental fiber mode as a function of core radius (normalized to the wavelength) for the range 0.1 to 10 wavelengths. Explain the results physically.
b. Assume that the fiber has a core radius of $4\ um$ and that it has an incident field with a circular and constant intensity with a radius of $2\ um$ perfectly centered on the fiber core. What is the coupling efficiency under these conditions?

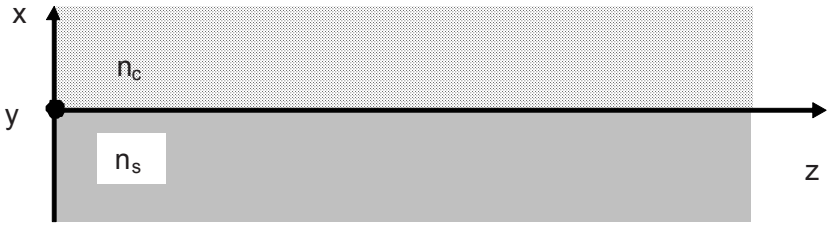
Problem 5.10 – Orthogonal Modes

Orthogonality of fiber modes is a central concept in waveguide theory. Presume that two functions $f(z)$ and $g(z)$, defined for $0 \le z \le h$, are orthogonal if and only if

$\int_0^h f^*(z)g(z)dz = 0$. Assume that *f(z)* and *g(z)* are both harmonic functions with nodes at *z=0* and *z=h*. Show that all modes of this type are orthogonal.

Problem 5.11 – Surface-Plasmon Dispersion Relation

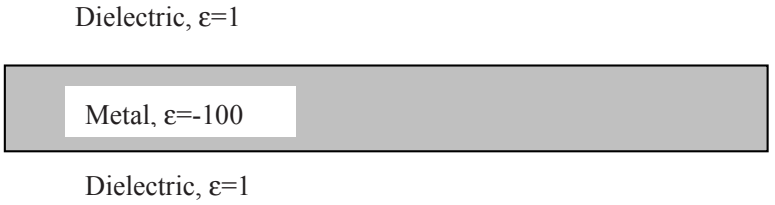
Consider the dielectric interface shown in the figure below. It can be treated as a slab waveguide of zero thickness. Use the approach taken in section 5.3 to find the field distribution and longitudinal wave vector of the surface-plasmon mode that is confined to the interface.



The interface between to semi-infinite dielectrics can support guided TM waves. We'll treat this structure as an asymmetric slab waveguide with zero thickness.

Problem 5.12 – Interacting Surface Plasmons

What types of guided modes are supported by the following structure? Explain qualitatively.



Problem 5.13 – Pulse Propagation on Optical Fibers

Consider a single-mode fiber with $a=5\mu m$, n_{core} = 1.45 and $n_{cladding}$ = 1.44. Assume that the normalized index is related to the normalized frequency by $b = V^2/\left(4+V^2\right)$

a. What is the waveguide dispersion at $\lambda=1\mu m$.

A transform-limited, Gaussian pulse ($\lambda=1\mu m$) with a FWHM width of *10 ps* is launched on the fiber.

b. What is the pulse length after it has propagated *100 km* on the fiber? Assume that the material dispersion at the pulse wavelength is *10ps/km/nm* for both the core and cladding material.

Problem 5.14 – Fiber Dispersion

a. What is the shortest possible Gaussian pulse that can be observed at the output of a single-mode fiber?
b. What is the corresponding initial pulse length? Consider both transform-limited and non-transform-limited input pulses. Express your answers in terms of wavelength, total dispersion, and length of the fiber.

Consider the single-mode fiber of Problem 5.13 ($a=5\mu m$, n_{core} = 1.45, and $n_{cladding}$ = 1.44). You want to launch 10 ps pulses that are chirped such that the output pulse is also 10 ps.

c. What is the longest length of fiber over which you can do this?

Problem 5.15 - Gaussian Pulse Propagation

What is the shortest possible Gaussian pulse that can be observed at the output of a single-mode fiber for a fixed length of the input pulse (i.e. the length is fixed, but the chirp is not)? Express your answers in terms of input pulse length, wavelength, total dispersion and fiber length. Notice that is NOT the same as the last problem. The difference is that here the length of the input pulse is fixed.

Problem 5.16 – Modal Dispersion

Consider a step-index optical fiber with the following parameters: a = 4.0 μm, n_{core} = 1.45 and $n_{cladding}$ = 1.444. The normalized longitudinal wave vectors for the lower order modes can be found from Fig. 511.

a. How many modes does the fiber support at 1.0 μm wavelength? (count all degenerate modes)

A *10 ns* Gaussian pulse is launched on a *10 km* length of this fiber such that all modes carry the same energy. Assume that there is no mode coupling on the fiber, and that the material and waveguide dispersion exactly compensate each other so that the total dispersion is negligible.

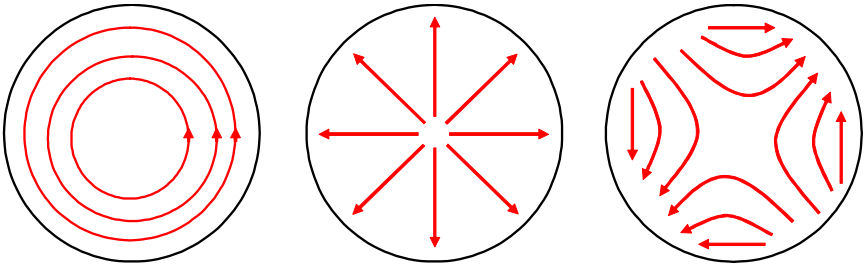
b. Sketch the time waveform of the output power from the fiber. (Detailed calculations of group velocities are not needed. Your sketch should just

indicate the order of arrival and relative power in the pulses on the output.) State your assumptions!

Problem 5.17 - Fiber Modes

The figure shows the transversal electric fields of some of lower order modes on a step-index optical fiber.

Show how these modes can be combined to obtain the LP_{11} modes.



Problem 5.18 – Fiber Polarizers

When a fiber is forced into the shape of a loop, the core deforms into an elliptical shape. Assume that the radius is decreased by δa in the plane of the loop and enlarged by the same amount in the orthogonal direction.

a. Name two effects that will contribute to birefringence of the fiber loop.
b. Which one of the two is dominant?
c. How can you utilize this effect to make a very useful fiber-optic device?

References

1 D.B. Keck, "Optical Fiber Waveguides", in *Fundamentals of Fiber Communications*, 2nd edition, M.K. Bartolski, editor, Academic Press, 1981, p. 18.

2 J.P. Heritage, A.M. Weiner, R.N. Thurston, "Picosecond Pulse Shaping by Spectral Phase and Amplitude Manipulation," Optics Letters 10, 609-611 (1985).

6: Fiber and Waveguide Devices

6.1 Introduction to Fiber and Waveguide Devices

Chapter 5 covered the basics of wave propagation on optical waveguides and fibers. In this chapter we use the concepts from Chapter 5 to develop models for a number of devices and systems that are of importance for microphotonics. The first "device" we consider is simply the optics required for coupling of light into, and out of, optical fibers. We introduce the concept of mode matching and quantify the coupling losses that result from imperfect alignment and matching.

We then develop a perturbation theory, called coupled-mode theory, that is useful for modeling of optical devices based on coupling and interference between two or more modes. Coupled-mode theory is used to describe two fiber devices that are very important in microphotonics and optical MEMS: Directional couplers and Fiber Bragg filters.

Directional couplers are beam splitting, or beam combining, devices with two inputs and two outputs. Their operation is based on evanescent interactions between guided modes. Directional couplers are important in their own right because they comes in many different varieties and are used in one form or another in almost all fiber-optic systems. In addition, evanescent interactions are important in a large number of optical devices other than directional couplers.

We also use coupled mode theory to describe Bragg reflectors. These devices can be implemented both as fiber components and as multi-layer film stacks, and play important roles as filters and reflectors in many optical systems. Bragg reflectors can also be thought of as one-dimensional Photonic Crystals. As such they represent our first introduction to a concept that will be investigated in detail in Chapters 14 and 15 of this book.

We wrap up the chapter with a discussion of optical modulators. It is not a comprehensive treatment, but rather an introduction to central concepts, including device architecture, figures of merit, and modulating effects.

6.2 Coupling to Fibers and Waveguides

To make use of optical fibers and optical waveguides we need to couple light in and out of them. There are several ways to excite propagating modes on optical waveguides, but the most straightforward is *end-fire*, in which we simply launch an optical field at the end of an optical fiber or waveguide. To maximize the coupling, we must maximize the *overlap* of the exciting field and the mode we want to excite [1] as illustrated in Fig. 6.1.

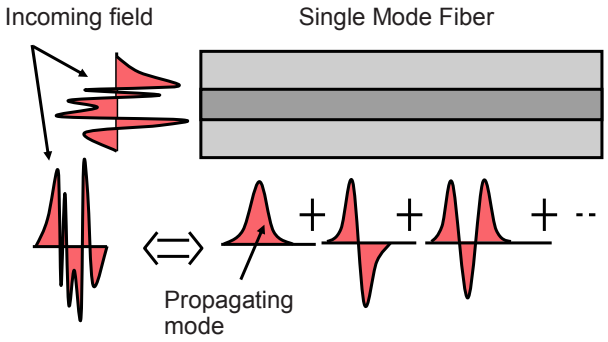


Figure 6.1. End-fire coupling of an optical field to a waveguide. The profile of the incoming field should match the profile of the waveguide mode we wish to excite to maximize the coupling.

Assume that the incoming electric field can be expressed in terms of the guided modes on the fiber as follows

$$\vec{E} = \sum_{guided} c_n \vec{E}_n + \int_{radiation} c_\beta \vec{E}_\beta \cdot d\beta \tag{6.1}$$

Using the Poynting theorem to express the propagating power in mode n, we find the following expression for the expansion coefficients

$$\int_{cross} \left(\vec{E}_{in} \times \vec{H}_n^*\right) \cdot dA = \int_{cross} \left(\left(\sum_{guided} c_n \vec{E}_n + \int_{radiation} c_\beta \vec{E}_\beta \cdot d\beta \right) \times \vec{H}_n^* \right) \cdot dA$$

$$= c_n \int_{cross} \left(\vec{E}_n \times \vec{H}_n^*\right) \cdot dA \Rightarrow c_n = \frac{\int_{cross} \left(\vec{E}_{in} \times \vec{H}_n^*\right) \cdot dA}{\int_{cross} \left(\vec{E}_n \times \vec{H}_n^*\right) \cdot dA} \tag{6.2}$$

Here we have used the fact that the guided modes of an optical fiber are orthogonal. The expression for the expansion coefficients of the incoming field is more useful when normalized

$$t_n = \frac{c_n}{\sqrt{\int_{cross}\left(\vec{E}_{in}\times\vec{H}_{in}^{*}\right)\cdot dA}}\sqrt{\int_{cross}\left(\vec{E}_{n}\times\vec{H}_{n}^{*}\right)\cdot dA}$$
$$= \frac{\int_{cross}\left(\vec{E}_{in}\times\vec{H}_{n}^{*}\right)\cdot dA}{\sqrt{\int_{cross}\left(\vec{E}_{in}\times\vec{H}_{in}^{*}\right)\cdot dA}\sqrt{\int_{cross}\left(\vec{E}_{n}\times\vec{H}_{n}^{*}\right)\cdot dA}} \tag{6.3}$$

The reflections at the waveguide interface complicate matters considerably because the reflection coefficients will depend on the propagation constants of the modes on the two sides of the interface. Finding exact solutions to this problem require that the incoming, as well as the transmitted fields, must be expressed in terms of modes with well-defined propagation constants. In most cases of practical interest, we may simply set the interface reflection and transmission coefficients to

$$r_{interface} = \frac{\beta_{in}-\beta_t}{\beta_{in}+\beta_t} = \frac{n_{eff,in}-n_{eff,t}}{n_{eff,in}+n_{eff,t}} \tag{6.4}$$

$$t_{interface} = \frac{2\beta_{in}}{\beta_{in}+\beta_t} = \frac{2n_{eff,in}}{n_{eff,in}+n_{eff,t}} \tag{6.5}$$

The total coupling coefficient from the incoming field into mode n, is then

$$t_n = \frac{2\beta_{in}}{\beta_{in}+\beta_t}\frac{\int_{cross}\left(\vec{E}_{in}\times\vec{H}_{n}^{*}\right)\cdot dA}{\sqrt{\int_{cross}\left(\vec{E}_{in}\times\vec{H}_{in}^{*}\right)\cdot dA}\sqrt{\int_{cross}\left(\vec{E}_{n}\times\vec{H}_{n}^{*}\right)\cdot dA}} \tag{6.6}$$

The power coupling equals the square of the magnitude of the field coupling

$$T_n = t_n t_n^* = \left(\frac{2\beta_{in}}{\beta_{in} + \beta_t} \right)^2 \frac{\left[\int_{cross} \left(\vec{E}_{in} \times \vec{H}_n^* \right) \cdot dA \right]^2}{\int_{cross} \left(\vec{E}_{in} \times \vec{H}_{in}^* \right) \cdot dA \cdot \int_{cross} \left(\vec{E}_n \times \vec{H}_n^* \right) \cdot dA} \tag{6.7}$$

These coupling formulas are significantly simplified if only the transversal components of the fields are considered. Most guided optical waves are close to Transversal Electro Magnetic (TEM), so ignoring the relatively small longitudinal part of the guided modes lead to insignificant over-estimation of the coupling coefficient in most, if not all, practical situations.

We observe that for TEM waves

$$H_x = \frac{-j}{\omega \cdot \mu_0} \frac{\partial E_y}{\partial z} = \frac{\beta}{\omega \cdot \mu_0} E_y \tag{6.8}$$

The coupling coefficient then becomes

$$t_n = \frac{2\sqrt{\beta_{in}} \cdot \sqrt{\beta_t}}{\beta_{in} + \beta_t} \frac{\int_{cross} \left(E_{in} E_n^* \right) \cdot dA}{\sqrt{\int_{cross} \left(E_{in} E_{in}^* \right) \cdot dA} \sqrt{\int_{cross} \left(E_n \vec{E}_n^* \right) \cdot dA}} \tag{6.9}$$

and the corresponding power coupling is

$$T_n = t_n t_n^* = \frac{4\beta_{in} \cdot \beta_t}{(\beta_{in} + \beta_t)^2} \frac{\left[\int_{cross} \left(E_{in} E_n^* \right) \cdot dA \right]^2}{\int_{cross} \left(E_{in} E_{in}^* \right) \cdot dA \cdot \int_{cross} \left(E_n E_n^* \right) \cdot dA} \tag{6.10}$$

6.2.1 Loss in Single Mode Fiber Splices

An important application of the formalism we have just developed is analysis of loss in single mode fiber splices. Practical fiber splices have several sources of power loss as illustrated in Fig. 6.2. The relative importance of the different sources of misalignment depends on the type of optical waveguide (its size, mode properties etc.) and on the splice technology that is used (fusion splicing, gluing in v-grooves, connectors).

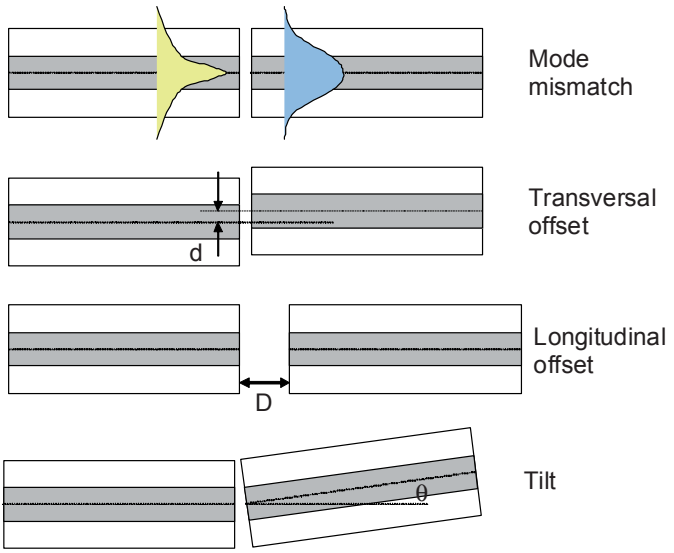


Figure 6.2. Sources of power loss in single-mode fiber splices include mode mismatch, longitudinal and transversal offset, and tilt.

In most practical situations, however, the transversal offset is the most critical. This is especially true for single-mode guides made in high-index materials like semiconductors and ferroelectric materials, because the mode sizes required for low-loss, single-mode operation are small and therefore relatively sensitive to transversal offset.

Single mode fibers are designed to have the largest practical mode size to reduce the transversal-offset sensitivity. Even so, transversal offset tends to dominate the losses in practical fiber connectors. Modern high-quality splicing equipment has reduced transversal offset, and therefore splicing loss, to the point where it is of little significance in splices between identical fibers. Coupling losses between different fibers, and between optical sources and fibers, are still very much an issue in optical communication, however.

To analyze how the different alignment errors influence splice loss, we will use a Gaussian approximation to describe the waveguide modes. Assume that the field on the input fiber can be described

$$E_{in}(r) = E_{0in} \cdot e^{-\frac{r^2}{\omega_{in}^2}} \tag{6.11}$$

Similarly, the mode on the output fiber is

$$E_t(r) = E_{0t} \cdot e^{-\frac{r^2}{\omega_t^2}} \tag{6.12}$$

We assume that the beam radius, ω, is related to the *V-number* and the core radius by the standard formula for step-index fiber found in Chapter 5

$$\frac{\omega}{a} = 0.65 + 1.619 \cdot V^{-1.5} + 2.87 \cdot V^{-6} \tag{6.13}$$

The Gaussian-beam approximation can easily be extended to elliptical beams. The basic Gaussian profile then becomes

$$E_{in}(x,y) = E_{0in} \cdot e^{-\frac{x^2}{\omega_x^2} - \frac{y^2}{\omega_y^2}} \tag{6.14}$$

6.2.2 Coupling Coefficients

The Gaussian beam approximation allows us to find closed form solutions to the loss caused by the alignment errors of Fig. 10.2. As an example, we will find the formula for the coupling coefficient in the presence of longitudinal offset between a waveguide with an elliptical mode profile, and a fiber. A separation, *D*, between the waveguide and fiber leads to the following coupling coefficient (assuming the propagation coefficient is the same on the waveguide and fiber)

$$\begin{aligned} t_n &= \frac{\sqrt{\frac{\omega_{0x}}{\omega_x}\frac{\omega_{0y}}{\omega_y}} \int_{cross} E_{0in} \cdot e^{-x^2\left(\frac{1}{\omega_x^2}+\frac{jk}{2R_x}\right)-y^2\left(\frac{1}{\omega_y^2}+\frac{jk}{2R_y}\right)} E_{0n} \cdot e^{-\frac{x^2}{\omega_{fiber}^2}-\frac{y^2}{\omega_{fiber}^2}} \cdot dA}{\sqrt{\int_{cross} E_{0in}^2 \cdot e^{-\frac{2x^2}{\omega_{0x}^2}-\frac{2y^2}{\omega_{0y}^2}} E_{0in} \cdot dA} \sqrt{\int_{cross} E_{0n}^2 \cdot e^{-\frac{2x^2}{\omega_{fiber}^2}-\frac{2y^2}{\omega_{fiber}^2}} \cdot dA}} \\ &= \frac{2\frac{1}{\omega_{fiber}}\sqrt{\frac{1}{\omega_x}\frac{1}{\omega_y}}}{\sqrt{\frac{1}{\omega_x^2}+\frac{1}{\omega_{fiber}^2}+\frac{jk}{2R_x}}\sqrt{\frac{1}{\omega_y^2}+\frac{1}{\omega_{fiber}^2}+\frac{jk}{2R_y}}} \end{aligned} \tag{6.15}$$

where $k=2\pi/\lambda$. The corresponding expression for the power coupling is

$$T = \frac{4\dfrac{\omega_{fiber}^2}{\omega_x \omega_y}}{\sqrt{\left(1+\dfrac{\omega_{fiber}^2}{\omega_x^2}\right)^2 + \dfrac{k^2\omega_{fiber}^4}{4R_x^2}}\sqrt{\left(1+\dfrac{\omega_{fiber}^2}{\omega_y^2}\right)^2 + \dfrac{k^2\omega_{fiber}^4}{4R_y^2}}} \tag{6.16}$$

In these expressions, the beam radii and radii of curvature are given by

$$R_{x,y} = D\left[1+\left(\frac{\pi\omega_{x0,y0}^2}{\lambda D}\right)^2\right] \tag{6.17}$$

$$\omega_{x,y} = \omega_{x0,y0}\left[1+\left(\frac{\lambda D}{\pi\omega_{x0,y0}^2}\right)^2\right]^{\frac{1}{2}} \tag{6.18}$$

The maximum coupling coefficient (perfect alignment, but imperfect mode matching) between an elliptical Gaussian mode with half axes ω_x and ω_y, and a circular Gaussian mode with mode radius ω_{fiber}, is given by

$$T = \frac{4\omega_{fiber}^2\omega_x\omega_y}{\left(\omega_x^2+\omega_{fiber}^2\right)\left(\omega_y^2+\omega_{fiber}^2\right)} \tag{6.19}$$

Similar calculations allow us to find closed-formed solutions for other types of misalignment. If the fiber axis is tilted with respect to the optical axis of the lens system, the coupling is given by

$$T = \frac{4\omega_{fiber}^2\omega_x\omega_y}{\left(\omega_x^2+\omega_{fiber}^2\right)\left(\omega_y^2+\omega_{fiber}^2\right)} e^{-\frac{2\pi^2\omega_{fiber}^2}{\lambda^2}\left(\frac{(\omega_x\theta_x)^2}{\omega_x^2+\omega_{fiber}^2}+\frac{(\omega_y\theta_y)^2}{\omega_y^2+\omega_{fiber}^2}\right)} \tag{6.20}$$

where θ_x and θ_y are the tilt angles in the x and y directions respectively.

If there is lateral offset between the axis of fiber and the axis of the lens system, the coupling is given by

$$T = \frac{4\omega_{fiber}^2\omega_x\omega_y}{\left(\omega_x^2+\omega_{fiber}^2\right)\left(\omega_y^2+\omega_{fiber}^2\right)} e^{-\frac{2d_x^2}{\omega_x^2+\omega_{fiber}^2}-\frac{2d_y^2}{\omega_y^2+\omega_{fiber}^2}} \tag{6.21}$$

where d_x and d_y are the lateral offsets in the x and y directions respectively.

Notice the opposite dependence on the mode sizes in these last two expressions. We see that a large mode size is relatively strongly affected by angular misalignment, while small modes sizes are more susceptible to lateral offset.

If tilt, offset and separation are present simultaneously, there will be cross-terms. These can be neglected provided that each of the errors (tilt, offset and separation) are small.

6.2.3 Laser to Single-Mode-Fiber Coupling

The formulas for coupling between fibers can also be used to describe coupling between single-spatial mode sources and single-mode waveguides. Unless we are using Anti Reflection (AR) coatings, we must also consider effect of reflections from the waveguide facet. This is neglected in the following examples, but can easily be included if dictated by the practical situation.

To achieve high-efficiency coupling from the laser to the fiber, we must transform the laser mode size to match the fiber mode. An imaging system with a single lens of the right focal length in the correct configuration is sufficient. The set-up is shown in Fig. 6.3.

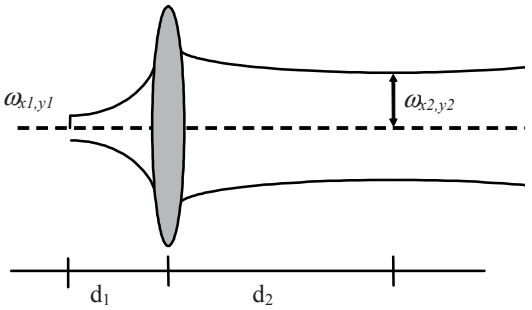


Figure 6.3. Lens transformation of Gaussian beams.

We now consider the effect of the lens on a Gaussian beam. From Chapter 4, we know that the Gaussian beam radii at the waists are related by

$$\omega_2^2 = \frac{\omega_1^2}{(1 - d_1/f)^2 + \dfrac{\pi^2 \omega_1^4}{\lambda^2 f^2}} \tag{6.22}$$

With the laser at the focal point of the lens (Fourier Transform regime), we have

$$d_1 = f \Rightarrow \omega_{x1,y1} = \frac{f\lambda}{\pi\omega_{x2,y2}} \tag{6.23}$$

If we make $d_1 > f$, we will image the beam waist to a position $d_2 > f$. We call this the imaging regime. In this case, the transformed beam radius can be approximated as

$$\omega_2 = \frac{\omega_1}{1 - d_1/f} \tag{6.24}$$

and the distances, d_1 and d_2, are related by the standard formula

$$\frac{1}{d_1} + \frac{1}{d_2} = \frac{1}{f} \tag{6.25}$$

Using these formulas, we can calculate the focal lengths of the lenses required to perfectly match any Gaussian-shaped laser mode to any Gaussian-shaped waveguide mode. Cylindrical lenses can be used to transform for elliptical beams to spherical Gaussians, but in most cases, this is an unwarranted complication. What is typically done is to transform the beam equally in both dimensions such that optimize the coupling is achieved. To see how, consider the following optimization

$$\begin{aligned}
&\frac{\partial T}{\partial \omega_f} = \frac{\partial}{\partial \omega_f} \frac{4\omega_f^2 \omega_x \omega_y}{\left(\omega_f^2 + \omega_x^2\right)\left(\omega_f^2 + \omega_y^2\right)} = 0 \Rightarrow \\
&\frac{8\omega_f \omega_x \omega_y \left(\omega_f^2 + \omega_x^2\right)\left(\omega_f^2 + \omega_y^2\right) - 4\omega_f^2 \omega_x \omega_y \left[2\omega_f \left(\omega_f^2 + \omega_x^2\right) + 2\omega_f \left(\omega_f^2 + \omega_y^2\right)\right]}{\left(\omega_f^2 + \omega_x^2\right)^2 \left(\omega_f^2 + \omega_y^2\right)^2} = 0 \\
&\Rightarrow \left(\omega_f^2 + \omega_x^2\right)\left(\omega_f^2 + \omega_y^2\right) = \omega_f^2 \left(\omega_f^2 + \omega_x^2\right) + \omega_f^2 \left(\omega_f^2 + \omega_y^2\right) \Rightarrow \omega_f = \sqrt{\omega_x \omega_y}
\end{aligned} \tag{6.26}$$

It comes as no surprise that the optimized fiber mode radius is the geometrical mean of the two elliptical half-radii.

6.2.4 Laser-Mode Size Measurements Using the Knife-Edge Method

A very convenient way to measure the size of a Gaussian beam is to chop the beam, detect the light using a photodiode, and observe the rise or fall time of the chopped pulses on an oscilloscope. The advantages of this technique, illustrated in Fig. 6.4, are that it requires only inexpensive, standard optical measurement equipment, and that it allows the relevant beam parameter, i.e. the Gaussian beam radius, to be directly read out on there oscilloscope. The drawbacks are that it only gives the integrated beam profile, so it does not allow detailed investigation of mode shapes. In practice it only works well for known beam shapes.

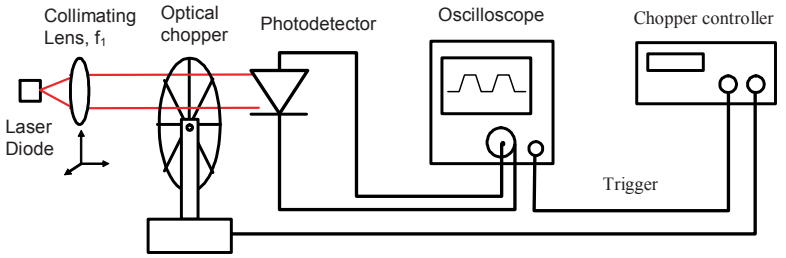


Figure 6.4. Set-up for measurement of laser mode size.

Figure 6.5 illustrates in detail how the shape of the beam influences the rise-time of the chopped signal. It is clear from the illustration that what is measured is the beam size at position of the knife edge, not at the detector.

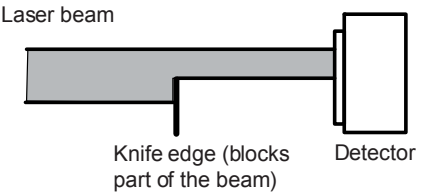


Figure 6.5. Illustration of knife-edge method for measuring the size of a laser beam.

When measuring elliptical beams, we must make sure to chop the beam such that the edge of the chopper intersects the beam along the axis we want to measure. It is practical to use a large photodetector when employing this technique, so we must be careful to chop the beam at sufficiently low frequency that our measured rise or fall time is not influenced by the detector. If the measured rise time is comparable to the rise time of the photo detector, our measurements will be inaccurate.

The beam-half-axis (in the direction of the edge movement) of the laser beam can be found from the following formula

$$\int_{-0.5d_{10-90}}^{0.5d_{10-90}} e^{-2\frac{x^2}{\omega_x^2}} dx = 0.8\sqrt{\frac{\pi}{2}}\omega_x \Rightarrow\Rightarrow \frac{2}{\sqrt{\pi}} \int_{0}^{\frac{\sqrt{2}}{2\omega_x}d_{10-90}} e^{-2\frac{x^2}{\omega_x^2}} d\left(\sqrt{2}\frac{x}{\omega_x}\right) = 0.8$$
$$\Rightarrow Erf\left(\frac{d_{10-90}}{\sqrt{2}\omega_x}\right) = 0.8 \Rightarrow \omega_x = \frac{d_{10-90}}{\sqrt{2}\cdot 0.91} = 0.78 d_{10-90} \quad (6.27)$$

where $d_{10\text{-}90}$ is the 10 to 90% integrated intensity width of the optical beam. This quantity can be expressed in terms of the rise time in the following way

$$\begin{aligned} d_{10-90} &= 2\pi R \cdot f_{chopper} \cdot t_{10-90} \\ \Rightarrow \omega_{x1,y1} &= 0.78 \cdot d_{10-90} = 0.78 \cdot 2\pi R \cdot f_{chopper} \cdot t_{10-90} \end{aligned} \tag{6.28}$$

where $t_{10\text{-}90}$ the 10-to-90% rise (or fall time) of the oscilloscope trace, R is the radius of the chopper where it intersects the laser beam, and f is the chopper frequency.

6.2.5 Coupling from Spatially Incoherent Sources to Multi Mode Fibers

In principle we can calculate a coupling coefficient from each mode of a spatially incoherent source (these modes are by definition delta functions) into each mode of a multimode guide, and integrate the coupled power over the source to find the total coupled power. In practice, however, most multimode waveguides have a large number of modes (there are some notable exceptions, particularly two-mode fibers used as acousto-optic modulators), so we can simply assume that the guide accepts all light incident on its core at below the critical angle.

The power coupling can then be expressed

$$P = \int_{A_f} dA_s \int_{\Omega_f} B(A_s, \Omega_s) d\Omega_s = \int_0^{r\min} \int_0^{2\pi} \left[\int_0^{2\pi} \int_0^{\theta\max} B(\theta, \phi) \sin\theta \cdot d\theta \cdot d\phi \right] d\theta_s \cdot r \cdot dr \tag{6.29}$$

where B is the brightness of the source, r_{min} is the smaller of the source and fiber radii, and θ_{max} is the acceptance angle of the fiber. Applied to the coupling between an LED (a Lambertian source, i.e. the brightness is $B=B_0 cos\theta$), this gives

$$\begin{aligned} P &= \int_0^{r\min} \int_0^{2\pi} \left[\int_0^{2\pi} \int_0^{\theta\max} B_0 \cos\theta \cdot \sin\theta \cdot d\theta \cdot d\phi \right] d\theta_s \cdot r \cdot dr \\ &= \int_0^{r\min} \int_0^{2\pi} 2\pi \cdot B_0 \int_0^{\theta\max} \cos\theta \cdot \sin\theta \cdot d\theta \cdot d\theta_s \cdot r \cdot dr \\ &= \pi \cdot B_0 \int_0^{r\min} \int_0^{2\pi} \sin^2\theta_{\max} \cdot d\theta_s \cdot r \cdot dr = 2\pi^2 \cdot B_0 \int_0^{r\min} NA^2 \cdot r \cdot dr \\ &= \pi^2 \cdot r_{\min}^2 \cdot B_0 \cdot NA^2 \end{aligned} \tag{6.30}$$

The power of an LED is

$$P_s = \pi \cdot r_s^2 \int_0^{2\pi} \int_0^{\pi/2} B_0 \cos\theta \cdot \sin\theta \cdot d\theta \cdot d\phi = \pi^2 \cdot r_s^2 \cdot B_0 \tag{6.31}$$

so if the source is smaller than the fiber, the coupling is

$$P = P_s \cdot NA^2 \tag{6.32}$$

and if the fiber is smaller, we get

$$P = \left(\frac{a}{r_s}\right)^2 P_s \cdot NA^2 \tag{6.33}$$

If the source is smaller than the fiber core, the coupling can be improved by magnifying the source. The magnification reduces the angle spread of the source, so that the net effect is to increase the coupled power.

6.2.6 Coupling between Spatially Coherent Sources and Multimode Fibers

When using a spatially coherent source, we have full control over the mode profile, and in principle, we can couple to any mode, or combination of modes with arbitrarily high coupling efficiency.

6.2.7 Coupling from Spatially Incoherent Sources to Single Mode Fibers

Coupling from spatially incoherent light sources (Lambertian sources) to single mode optical waveguides is very inefficient. This is so because the source is effectively a combination of point sources with randomly varying relative phase. In other words, a spatially incoherent light source is in reality a combination of multiple sources.

As we saw in Chapter 2, it is impossible to combine optical fields, so the best we can do with multiple sources is to pick the strongest and couple it to the single mode fiber. Typical incoherent sources consist of hundreds of independent sources of roughly equal strength, so the coupling to single mode fiber will be less than 1%. In practice we therefore very rarely use LEDs or other spatially incoherent sources, with single mode fibers. An exception is super luminescent diodes, which are used in fiber gyros and other interferometric measurement systems that require sources with low temporal coherence.

6.2.8 Prism Coupling

Plane waves can couple to quasi-guided waves if the two waves are phase matched. This principle can be used to couple into waveguides, but the incoming field must be phase matched to the guided wave, which is not possible for an os-

cillatory wave in the cladding of the guide. The solution is to use an evanescent wave that is phase matched to the guided wave, and that has a finite overlap with the guided mode. This is illustrated in Fig. 6.6.

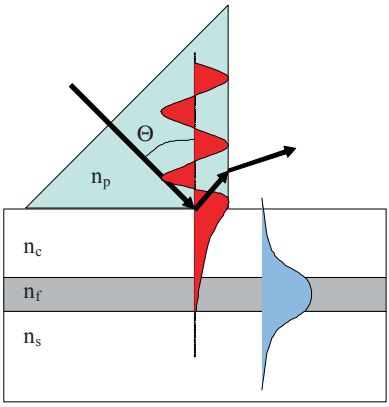


Figure 6.6. *Illustration of prism coupling. The plane wave in the high-index medium is phase matched to the guided wave, and power is exchanged between the plane wave and the guided mode through the evanescent fields.*

For the plane wave and the guided mode to couple, their wave vectors along the guide must be identical (which is to say that the evanescent field of the plane wave is phase matched to the guided wave). This requirement determines the incident angle, θ_{match}, of the plane wave on the prism-cladding interface

$$k_{z,plane} = k_0 n_p \sin\theta_{match} = \beta_{guide} = k_0 n_{eff}$$
$$\Rightarrow \theta_{match} = \sin^{-1}\left[\frac{n_{eff}}{n_p}\right] \tag{6.34}$$

A prism must be used to launch the plane wave at the correct phase-matching angle.

If the plane wave couples to the guided wave, then reciprocity mandates that the guided wave couples to the plane wave. When using a prism to couple into a waveguide, we must therefore align the incoming beam to the end of the prism as shown in Fig. 6.6. If the prism extends beyond the incoming beam, the power will simply couple out of the waveguide again.

Prism coupling can be used to find the modes of a waveguide; either by observing reflection dips in the incident beam as a function of angle, or by measuring the angles at which light is coupled out of the prism (this technique requires that we can excite all the modes of interest on the waveguide). Once the phase match angles

are determined, the wave vectors of the guided modes can be found from the equation above.

Prism coupling is complicated by the need to establish a near perfect optical contact between the prism and the waveguide structure. This is a non-trivial problem that, together with the fact that prisms are bulky and expensive, reduces the utility of prism coupling in practical applications.

6.2.9 Grating Coupling

Gratings can also be used to facilitate phase matching between an incident optical field and a guided mode. To understand why, recall that a grating that is periodic in the *z*-direction, adds a propagation vector

$$K_z = q\frac{2\pi}{\Lambda} \tag{6.35}$$

to the optical field. In this formula q can be any integer (positive, negative, or zero). The additional *k*-vector provided by the grating can be designed to phase match a plane wave in the cladding of a waveguide to a guided mode. This is illustrated in Fig. 6.7. We see that in addition to the diffracted beam that is phase matched to the waveguide mode, there also exist other diffracted modes. Any power that is coupled into these modes is wasted in that it is not transferred into the guided mode. In designing rating couplers we therefore must pay careful attention to all diffraction modes to ensure efficient coupling.

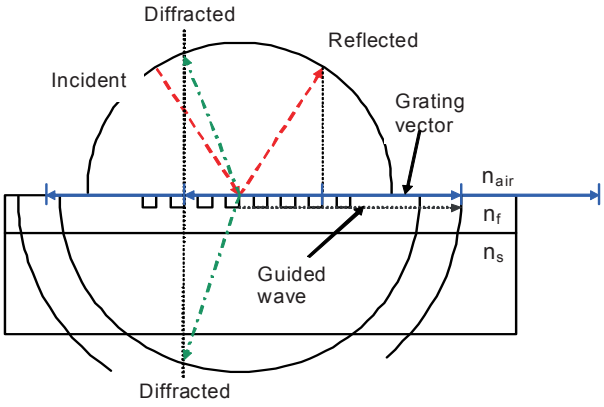


Figure 6.7. Grating coupling. The incident light (dashed arrow) is coupled to several diffracted orders by the k-vectors of the grating (solid arrows). In addition to the reflected wave (dashed), and the guided wave (dotted), we also have one diffracted wave propagating in air (dot-dashed), and one in the substrate (dot-dashed).

Just as in the case of prism coupling, the phase matching provided by gratings can also be used for coupling out of waveguides. Creative design of the grating and waveguide structure will allow most of the light to be coupled into a single output mode as shown in Fig. 6.8.

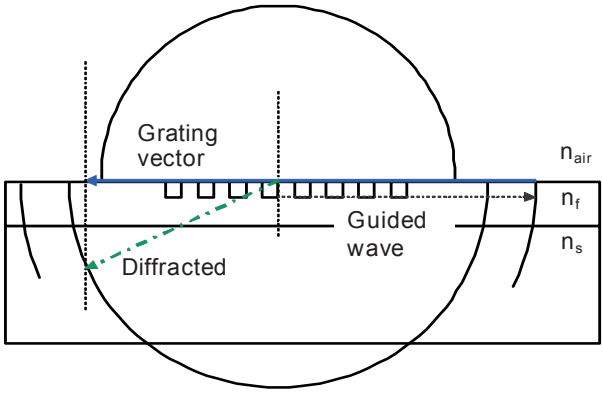


Figure 6.8. *Grating output coupling. The grating vector (solid arrow) is chosen large enough that the guided wave (dotted arrow) only couples to one diffracted output mode (dot-dashed arrow). This wave cannot be coupled out of the substrate without a prism.*

Symmetric gratings tend to put too much power into unwanted diffraction modes, so to achieve efficient coupling, blazed gratings must be used. In any grating the power in the diffracted orders are given by the radiation pattern form the individual element of the gratings as illustrated in Fig. 6.9. A blazed grating in which the grating elements are aligned with the preferred diffraction order can therefore have high diffraction efficiency as shown in Fig. 6.10.

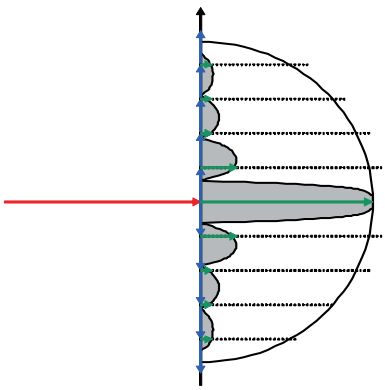


Figure 6.9 *The power in the diffracted orders is determined by the radiation pattern of the grating elements.*

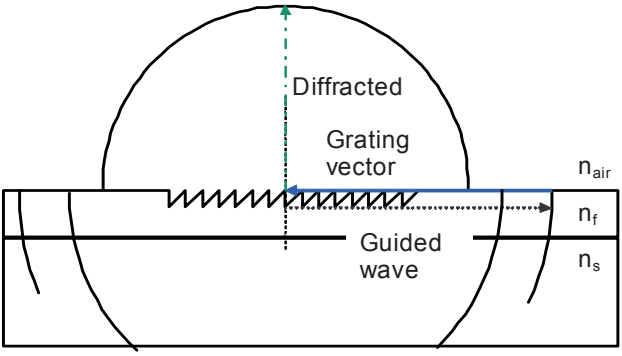


Figure 6.10. Blazed grating output coupler. The preferential radiation of the grating elements leads to high coupling efficiency of the guided wave (dotted arrow) to the output (dot-dashed arrow).

Grating couplers based on the principles described here are becoming increasingly popular for coupling light into high-index guided waves and photonic crystal structures made in silicon and other semiconductors. The main advantages of grating coupling for such applications are that the coupling optics can be miniaturized and that the coupling takes place on a planar surface without requiring access to the cross sectional plane of the waveguides.

6.3. Coupled Optical Modes

Propagating modes on optical waveguides can interact, and therefore be coupled, just as the mechanical and other types of oscillators. Consider the waveguide structure of Fig. 6.11. The individual waveguides of this structure are slab waveguides (i.e. they are of infinite extent in the *y* coordinate that is perpendicular to the plane). Their fields can therefore be expressed in rectangular coordinates without coupling between the fields along the coordinate axes. This greatly simplifies the problem, because it allows us to find each of the field components as a solution to the scalar wave equation.

The full mode structure in the region where the waveguides are in close proximity is considerably more complex, so we are compelled to search for approximations that will give us analytical solutions.

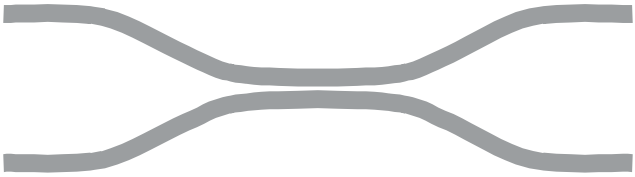


Figure 6.11 *Waveguide structure consisting of two waveguides that are brought into close proximity over a finite distance. In the area where the two waveguides are close, the modes of the waveguides will interact, which means that their propagation vectors will be shifted, and their modes modified.*

Our starting point is the scalar wave equation

$$\nabla^2 E_y(x,z) = \varepsilon\mu \cdot \frac{\partial^2 E_y(x,z)}{\partial t^2} \tag{6.36}$$

that is valid in all the regions of the waveguide. The electric permittivity, or dielectric constant, and the magnetic permeability are both time invariant[a], so the solutions we are interested in are independent of z, i.e. they are of the form

$$E_{yi}(x,z) = \frac{1}{2} A_i \cdot u_{yi}(x) e^{-j(\beta_i z - \omega t)z} + c.c. \tag{6.37}$$

where β is the longitudinal wave vector, A_i is the amplitude of the mode and u_{yi} is the normalized amplitude distribution of the mode.

If we now chose to first consider the fundamental TE mode of a symmetric waveguide, then the mode profile is given by

$$u_y(x,z) = \begin{cases} C\exp[-\gamma_c x] & x > 0 \\ C\left(\cos(\kappa_k x) - \dfrac{\gamma_c}{\kappa_k}\sin(\kappa_k x)\right) & -\mathrm{h} < x < 0 \\ C\left(\cos(\kappa_k h) + \dfrac{\gamma_c}{\kappa_k}\sin(\kappa_k h)\right)\exp[-\gamma_s(x+h)] & x < -\mathrm{h} \end{cases} \tag{6.38}$$

where is C is determined by the requirement

[a] We see examples of time varying permitivities later in our discussions of optical modulators, but even in those cases the time variations are so slow compared to the optical oscillations that we can consider the permitivity time invariant.

$$\int_S u_{yn}(x)\cdot u_{ym}(x)\cdot dx = \frac{2\omega\cdot\mu_0}{\beta_n}\delta_{nm} \tag{6.39}$$

As mentioned above, solving the wave equation on coupled waveguide structures is prohibitively hard and cannot be done analytically. Our approach to this problem is to consider the simpler structure, for which we can find the modes, and introduce the coupling as a perturbation of the polarization of the medium. To do that, we write constitutive relation for electric field as

$$\vec{D} = \varepsilon\vec{E} = \varepsilon_0\vec{E} + \vec{P} = \varepsilon\vec{E} + \vec{P}_{pert} \tag{6.40}$$

which means we can write the wave equation in the following way

$$\nabla^2 E_y(x,z,t) = \varepsilon\mu\cdot\frac{\partial^2 E_y(x,z,t)}{\partial t^2} = \varepsilon\mu\cdot\frac{\partial^2 E_y(x,z,t)}{\partial t^2} + \mu\frac{\partial^2 P_{pert}(x,z,t)}{\partial t^2} \tag{6.41}$$

First we set the perturbation to zero and find the modes of the unperturbed or uncoupled waveguide. The fields of the perturbed guide can be expressed in terms of these unperturbed modes

$$\begin{aligned} E_y(x) &= \frac{1}{2}\sum_i A_i^+ u_i(x)\cdot\exp[-j(\beta_i z-\omega t)] + c.c. \\ &+\frac{1}{2}\sum_i A_i^- u_i(x)\cdot\exp[-j(\beta_i z+\omega t)] + c.c \\ &+\frac{1}{2}\int_{k_0 n_s}^{k_0 n_f} A(\beta,z)\cdot u_\beta(x)\cdot\exp[-j(\beta z+\omega t)]d\beta + c.c \end{aligned} \tag{6.42}$$

where we have included both forward and backward traveling waves.

If we carry the full field expansions, we haven't made any approximations, but we haven't simplified the problem either. To make the problem tractable, we will assume that coupling to the radiation modes is negligible and therefore drop these modes from the expansion. Coupled-mode theory relay on this approximation, so **negligible coupling to radiation modes can therefore be used as a criterion for when to apply coupled mode theory.**

Now substitute the expanded solution back into the wave equation

$$\nabla^2\left\{\begin{array}{l}\frac{1}{2}\sum_i A_i^+ u_i(x)\cdot\exp\left[-j\left(\beta_i z-\omega t\right)\right]+\\ \frac{1}{2}\sum_i A_i^- u_i(x)\cdot\exp\left[j\left(\beta_i z+\omega t\right)\right]+c.c\end{array}\right\}=$$

$$\varepsilon\mu\cdot\frac{\partial^2}{\partial t^2}\left\{\begin{array}{l}\frac{1}{2}\sum_i A_i^+ u_i(x)\cdot\exp\left[-j\left(\beta_i z-\omega t\right)\right]\\ +\frac{1}{2}\sum_i A_i^- u_i(x)\cdot\exp\left[j\left(\beta_i z+\omega t\right)\right]+c.c\end{array}\right\}+\mu\frac{\partial^2 P_{pert}(x,z,t)}{\partial t^2}\Rightarrow$$

$$\sum_i\left[\begin{array}{l}A_i^+\left(-\beta_i^2 u_i(x)+\frac{\partial^2 u_i(x)}{\partial x^2}+\omega^2\mu\cdot\varepsilon\cdot u_i(x)\right)\cdot\exp\left[-j\beta_i z\right]+\\ \left(-2j\beta_i\frac{dA_i^+}{dz}+\frac{d^2A_i^+}{dz^2}\right)u_i(x)\cdot\exp\left[-j\beta_i z\right]+c.c.\end{array}\right] \tag{6.43}$$

$$+\sum_i\left[\begin{array}{l}A_i^-\left(-\beta_i^2 u_i(x)+\frac{\partial^2 u_i(x)}{\partial x^2}+\omega^2\mu\cdot\varepsilon\cdot u_i(x)\right)\cdot\exp\left[-j\beta_i z\right]+\\ \left(2j\beta_i\frac{dA_i^-}{dz}+\frac{d^2A_i^-}{dz^2}\right)u_i(x)\cdot\exp\left[j\beta_i z\right]+c.c.\end{array}\right]$$

$$=2\exp\left[-j\omega t\right]\cdot\mu\frac{\partial^2 P_{pert}(x,z,t)}{\partial t^2}$$

Notice that the first three terms of each of the two summations equals zero. The expression then simplifies to

$$\sum_i\left[\left(-2j\beta_i\frac{dA_i^+}{dz}+\frac{d^2A_i^+}{dz^2}\right)u_i(x)\cdot\exp\left[-j\beta_i z\right]\right]+c.c.$$

$$+\sum_i\left[\left(2j\beta_i\frac{dA_i^-}{dz}+\frac{d^2A_i^-}{dz^2}\right)u_i(x)\cdot\exp\left[j\beta_i z\right]\right]+c.c \tag{6.44}$$

$$=2\exp\left[-j\omega t\right]\cdot\mu\frac{\partial^2 P_{pert}(x,z,t)}{\partial t^2}$$

We also assume slow variations of the amplitudes

$$\left|\frac{d^2A_i}{dz^2}\right|<<\beta_i\left|\frac{dA_i}{dz}\right| \tag{6.45}$$

so

$$\begin{aligned}
&\exp[j\omega t]\frac{1}{2}\sum_i\left[-2j\beta_i\frac{dA_i^+}{dz}u_i(x)\cdot\exp[-j\beta_i z]\right]+c.c.\\
&+\exp[j\omega t]\frac{1}{2}\sum_i\left[2j\beta_i\frac{dA_i^-}{dz}u_i(x)\cdot\exp[j\beta_i z]\right]+c.c\\
&=\mu\frac{\partial^2 P_{pert}(x,z,t)}{\partial t^2}
\end{aligned}\tag{6.46}$$

We multiply this equation with $u_i(x)$, and integrate over the cross section of the guide (from – ∕ to ∕ in x), and use mode orthogonality to arrive at our final result

$$\begin{aligned}
&-\frac{dA_i^+}{dz}\cdot\exp[-j(\beta_i z-\omega\cdot t)]+\frac{dA_i^-}{dz}\cdot\exp[j(\beta_i z+\omega\cdot t)]+c.c\\
&=\frac{-j}{2\omega}\frac{\partial^2}{\partial t^2}\int_{-\infty}^{\infty}P_{pert}(x,z,t)\cdot u_i(x)dx
\end{aligned}\tag{6.47}$$

This equation can be used to treat a variety of waveguide structures with different types of interactions or coupling between guided modes. The exact form of the perturbation will depend on the waveguide structure at hand, but the general form of the coupled-mode equations will be the same.

6.4 Directional Couplers

We will now apply our coupled mode formalism to directional couplers. These devices are very important in integrated optics and optical communications in general. Directional couplers are used as optical modulators, optical power splitters and combiners, sensors, and most importantly they form the basis for waveguide optical switches, which presently represents the most promising approach to all-optical packet switching.

The directional coupler consists of two waveguides, which are brought into close proximity so that their modes overlap. The proximity of the other guide and the evanescent field of modes that may propagate on the other guide, represent perturbations of the modes of each of the waveguides. In general the two guides are not identical (although they often are in practice), but may have different core index and core width as illustrated in Fig. 6.12.

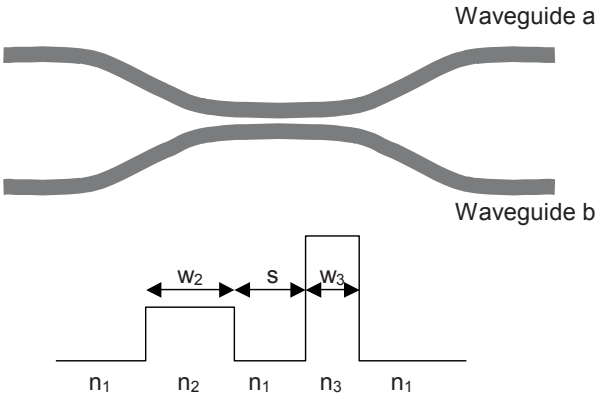


Figure 6.12. Directional coupler consisting of two single-mode waveguides that are brought into close proximity over a finite distance. In the area where the two waveguides are close, the modes of the waveguides will interact, which means that their propagation vectors will be shifted, and their modes modified because of the perturbed polarization resulting from the other guide.

Before we start our mathematical treatment of the directional coupler, let's see what we can learn about it by considering simple energy-conservation arguments, and our intuition about coupled oscillators. The first thing we should ask is whether it is reasonable to use the coupled-mode theory we have derived to describe directional couplers. The basic assumption we made is that we can neglect coupling to radiation modes. Our earlier investigations showed us that we can indeed have energy conservation without considering radiation modes provided that we have an equal number of output and input modes. The directional coupler fulfills this criterion so it seems plausible that we should be able to get accurate descriptions of well-designed directional couplers in spite of the fact that we neglect radiation modes. This is in contrast to the waveguide Y-junction or Y-coupler. When analyzing this device we must consider radiation modes to achieve energy conservation, so the Y-coupler is not a candidate for coupled-mode modeling.

Based on the structure shown in Fig. 6.12, we expect to get energy transfer between the coupled modes of the waveguides, and that the perturbation caused by the coupling will modify the propagation constant and the mode profiles of the individual waveguides. We expect, just like in the case of two coupled mechanical oscillators, to get total energy transfer between the modes when they are degenerate, while two non-degenerate modes will have incomplete energy transfer.

6.4.1 Coupled Mode Description of Directional Couplers

We start our derivation of the coupled mode equations for the directional coupler by describing the field in the coupled-guide structure as a sum of the unperturbed modes

$$E_y = A(z)u_a(x)\exp[j(\omega\cdot t-\beta_a z)]+B(z)u_b(x)\exp[j(\omega\cdot t-\beta_b z)] \tag{6.48}$$

where indexes refer to waveguide *a* and *b* in Fig. 6.12.

Recall that the perturbation is given by

$$\begin{aligned}&\vec{D}=\varepsilon_c\vec{E}=\varepsilon_0\vec{E}+\vec{P}_c=\varepsilon_0\vec{E}+\vec{P}_u+\vec{P}_{pert}=\varepsilon_u\cdot\vec{E}+\vec{P}_{pert}\Rightarrow\\&\vec{P}_{pert}=(\varepsilon_c-\varepsilon_u)\vec{E}\end{aligned} \tag{6.49}$$

where indexes *c* and *u* refer to the coupled and uncoupled waveguide structures respectively.

Substituting the above expression for the field in the coupled guide into this expression for the perturbation polarization results in

$$\begin{aligned}&P_{pert}=(\varepsilon_c-\varepsilon_u)\\&\{A(z)u_a(x)\exp[j(\omega\cdot t-\beta_a z)]+B(z)u_b(x)\exp[j(\omega\cdot t-\beta_b z)]\}\Rightarrow\\&P_{pert}=\varepsilon_0 e^{j\omega\cdot t}\left\{A(z)u_a(x)\left(n_c^2-n_a^2\right)e^{-j\beta_a z}+B(z)u_b(x)\left(n_c^2-n_b^2\right)e^{-j\beta_b z}\right\}\end{aligned} \tag{6.50}$$

where $n_c(x)$ is the index profile of the coupled structure.

Now recall the fundamental coupled mode equation

$$\begin{aligned}&-\frac{dA_i^+}{dz}\cdot\exp[-j(\beta_i z-\omega\cdot t)]+\frac{dA_i^-}{dz}\cdot\exp[j(\beta_i z+\omega\cdot t)]+c.c\\&=\frac{-j}{2\omega}\frac{\partial^2}{\partial t^2}\int_{-\infty}^{\infty}P_{pert}(x,z,t)\cdot u_i(x)dx\end{aligned} \tag{6.51}$$

We are not considering backward propagating waves in the directional coupler (i.e. we expect scattering into backward propagating modes to be negligible), so this simplifies to

$$-\frac{dA}{dz}\cdot\exp[-j(\beta_i z-\omega\cdot t)]+c.c=\frac{-j}{2\omega}\frac{\partial^2}{\partial t^2}\int_{-\infty}^{\infty}P_{pert}(x,z,t)\cdot u_a(x)dx \tag{6.52}$$

Substituting the expression for the polarization perturbation into this equation, and integrating over *x*, gives

$$-\frac{dA}{dz}\cdot\exp\left[-j\left(\beta_a z-\omega\cdot t\right)\right]+c.c=$$
$$\frac{-j}{2\omega}\frac{\partial^2}{\partial t^2}\int_{-\infty}^{\infty}\varepsilon_0 e^{j\omega\cdot t}\left\{\begin{matrix}A(z)u_a(x)\left(n_c^2-n_a^2\right)e^{-j\beta_a z}+\\B(z)u_b(x)\left(n_c^2-n_b^2\right)e^{-j\beta_b z}\end{matrix}\right\}\cdot u_a(x)dx \tag{6.53}$$

$$\frac{dA}{dz}=-\frac{j\omega\cdot\varepsilon_0}{4}\int_{-\infty}^{\infty}\left\{\begin{matrix}A(z)u_a(x)\left(n_c^2-n_a^2\right)+\\B(z)u_b(x)\left(n_c^2-n_b^2\right)e^{-j(\beta_b-\beta_a)z}\end{matrix}\right\}\cdot u_a(x)dx\Rightarrow \tag{6.54}$$

$$\frac{dA}{dz}=-\frac{j\omega\cdot\varepsilon_0}{4}\left\{\begin{matrix}\int_{-\infty}^{\infty}A(z)u_a^2(x)\left(n_c^2-n_a^2\right)\cdot dx+\\\int_{-\infty}^{\infty}B(z)u_b(x)u_a(x)\left(n_c^2-n_b^2\right)e^{-j(\beta_b-\beta_a)z}\cdot dx\end{matrix}\right\} \tag{6.55}$$

We find a similar expression for the mode amplitude in guide *b*. These two equations can be written in the following compact form

$$\frac{dA}{dz}=-jK_{ab}Be^{-j(\beta_b-\beta_a)z}-jM_aA \tag{6.56}$$

$$\frac{dB}{dz}=-jK_{ba}Ae^{-j(\beta_a-\beta_b)z}-jM_bB \tag{6.57}$$

where

$$K_{ab,ba}=\frac{\omega\cdot\varepsilon_0}{4}\int_{-\infty}^{\infty}u_b(x)u_a(x)\left(n_c^2-n_{a,b}^2\right)\cdot dx \tag{6.58}$$

$$M_{a,b}=\frac{\omega\cdot\varepsilon_0}{4}\int_{-\infty}^{\infty}u_{a,b}^2(x)\left(n_c^2-n_{a,b}^2\right)\cdot dx \tag{6.59}$$

The phase terms M_a and M_b in the above equations reflect the fact that the longitudinal propagation vectors of the modes are influenced by the presence of a second guide. We can incorporate this correction to the propagation vector in the field expression as follows

$$E_y=A(z)u_a(x)\exp\left[j\left(\omega\cdot t-\left(\beta_a+M_a\right)z\right)\right]+$$
$$B(z)u_b(x)\exp\left[j\left(\omega\cdot t-\left(\beta_b+M_b\right)z\right)\right] \tag{6.60}$$

This modified field expansion allow us to simplify the coupled-mode equations, which become

$$\frac{dA}{dz} = -jK_{ab}Be^{-j2\delta\cdot z} \tag{6.61}$$

$$\frac{dB}{dz} = -jK_{ba}Ae^{j2\delta\cdot z} \tag{6.62}$$

where

$$2\delta = (\beta_b + M_b) - (\beta_a + M_a) \tag{6.63}$$

Let us now consider these equations from an energy conservation point of view. The first of the two equations can be manipulated to yield

$$\begin{aligned} &\frac{d}{dz}|A|^2 = \frac{d}{dz}AA^* = A\frac{d}{dz}A^* + A^*\frac{d}{dz}A \\ &= AB^*\cdot jK_{ab}^* e^{j2\delta\cdot z} - A^*B\cdot jK_{ab}e^{-j2\delta\cdot z} \end{aligned} \tag{6.64}$$

Similarly

$$\begin{aligned} &\frac{d}{dz}|B|^2 = \frac{d}{dz}BB^* = B\frac{d}{dz}B^* + B^*\frac{d}{dz}B \\ &= BA^*\cdot jK_{ba}^* e^{-j2\delta\cdot z} - B^*A\cdot jK_{ba}e^{j2\delta\cdot z} \end{aligned} \tag{6.65}$$

so

$$\begin{aligned} &\frac{d}{dz}|A|^2 + \frac{d}{dz}|B|^2 = \\ &= AB^*\cdot jK_{ab}^* e^{j2\delta\cdot z} - A^*B\cdot jK_{ab}e^{-j2\delta\cdot z} \\ &+ BA^*\cdot jK_{ba}^* e^{-j2\delta\cdot z} - B^*A\cdot jK_{ba}e^{j2\delta\cdot z} = \\ &AB^*\cdot je^{j2\delta\cdot z}\left(K_{ab}^* - K_{ba}\right) + A^*B\cdot je^{-j2\delta\cdot z}\left(K_{ba}^* - K_{ab}\right) \end{aligned} \tag{6.66}$$

We see that to make this expression zero in general, we need

$$K_{ab}^* = K_{ba} \tag{6.67}$$

The phase of the coupling factor K depend on the choice of the origin of our coordinate system, so without loss of generality we can set

$$K_{ab} = K_{ba} \tag{6.68}$$

Notice that this is in agreement with our earlier formulation of energy conservation, in which we require that the overlap integral is unchanged under propagation though a linear loss-less system.

The above set of equations can be solved by differentiating the first equation with respect to z. Assuming $K_{ab}=K_{ba}=K$, we get

$$\begin{aligned}\frac{d^2A}{dz^2} &= -\left(-j2\delta\right)\cdot jK_{ab}Be^{-j2\delta\cdot z} - jK_{ab}e^{-j2\delta\cdot z}\frac{dB}{dz} \\ &= j2\delta\cdot\frac{dA}{dz} - K_{ab}e^{-j2\delta\cdot z}K_{ba}Ae^{j2\delta\cdot z} \Rightarrow\end{aligned} \tag{6.69}$$

$$\frac{d^2A}{dz^2} - j2\delta\cdot\frac{dA}{dz} - K^2A = 0 \tag{6.70}$$

This differential equation has the solution

$$\begin{aligned}A &= A_1\exp\left[j\left(\delta+\sqrt{K^2+\delta^2}\right)z\right] + A_2\exp\left[j\left(\delta-\sqrt{K^2+\delta^2}\right)z\right] \\ B &= B_1\exp\left[j\left(-\delta+\sqrt{K^2+\delta^2}\right)z\right] + B_2\exp\left[j\left(-\delta-\sqrt{K^2+\delta^2}\right)z\right] \Rightarrow\end{aligned} \tag{6.71}$$

$$\begin{aligned}A &= e^{j\delta\cdot z}\left\{\left(A_1+A_2\right)\cos\left[\sqrt{K^2+\delta^2}\,z\right] + j\left(A_1-A_2\right)\sin\left[\sqrt{K^2+\delta^2}\,z\right]\right\} \\ B &= e^{-j\delta\cdot z}\left\{\left(B_1+B_2\right)\cos\left[\sqrt{K^2+\delta^2}\,z\right] + j\left(B_1-B_2\right)\sin\left[\sqrt{K^2+\delta^2}\,z\right]\right\}\end{aligned} \tag{6.72}$$

With the boundary conditions $A(0)=A_0$ and $B(0)=0$, this becomes

$$A\left(z=0\right) = \left(A_1+A_2\right) = A_0 \tag{6.73}$$

$$\begin{aligned}\frac{dA(z=0)}{dz} &= 0 = j\delta\cdot A_0 + j\left(A_1-A_2\right)\sqrt{K^2+\delta^2} \Rightarrow \\ \left(A_1-A_2\right) &= -\frac{\delta}{\sqrt{K^2+\delta^2}}A_0\end{aligned} \tag{6.74}$$

and

$$B\left(z=0\right) = 0 = \left(B_1+B_2\right) \tag{6.75}$$

$$\frac{dB(z=0)}{dz} = -jKA_0 = -j\delta\cdot(B_1+B_2)+j(B_1-B_2)\sqrt{K^2+\delta^2} \Rightarrow$$
$$(B_1-B_2) = -\frac{KA_0}{\sqrt{K^2+\delta^2}} \tag{6.76}$$

so

$$A = A_0\cdot e^{j\delta\cdot z}\left\{\cos\left[\sqrt{K^2+\delta^2}\,z\right]-j\frac{\delta}{\sqrt{K^2+\delta^2}}\sin\left[\sqrt{K^2+\delta^2}\,z\right]\right\} \tag{6.77}$$

$$B = -jA_0e^{-j\delta\cdot z}\frac{K}{\sqrt{K^2+\delta^2}}\sin\left[\sqrt{K^2+\delta^2}\,z\right] \tag{6.78}$$

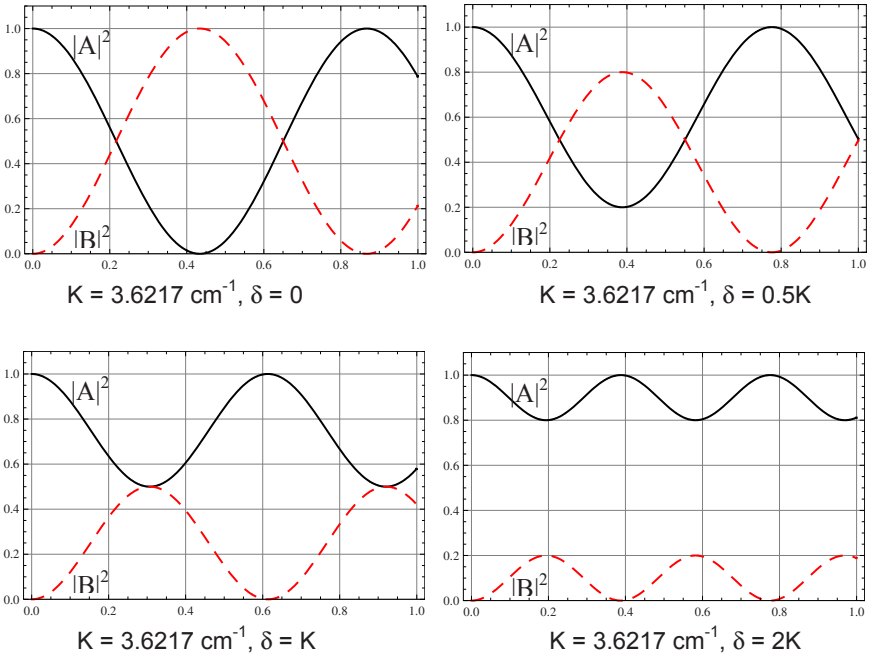


Figure 6.13 *Power coupling between two 10 um wide slab waveguides separated by 4 um. In the symmetrical case (n_f=1.5, n_c=1.499), the power coupling is complete. A finite difference in propagation constant between the two modes leads to less than perfect power exchange.*

When $\delta=0$ Eqs. 6.77 and 6.78 simplify to

$$A = A_0 \cdot \cos[Kz] \tag{6.79}$$

$$B = -jA_0 \sin[Kz] \tag{6.80}$$

These expressions are plotted for various values of δ in Fig. 6.13. We see that when the longitudinal wavevector of the two propagating modes are the same (δ=*0*), then the power is periodically transferred from one guide to the other and then back to the first.

6.4.2 Eigenmodes of the Coupled System

Now let's go back and take a closer look at the coupled mode equations

$$\frac{dA}{dz} = -jKBe^{-j2\delta \cdot z} \tag{6.81}$$

$$\frac{dB}{dz} = -jKAe^{j2\delta \cdot z} \tag{6.82}$$

The field on the coupled waveguide system can be described in terms of the (modified) modes on the two separate guides, so

$$E(z) = \begin{vmatrix} E_1(z) \\ E_2(z) \end{vmatrix} = \begin{vmatrix} A(z)e^{-j\beta_a z} \\ B(z)e^{-j\beta_b z} \end{vmatrix} \tag{6.83}$$

The coupled-mode equations are then in matrix form

$$\frac{dE(z)}{dz} = \frac{d}{dz}\begin{vmatrix} E_1(z) \\ E_2(z) \end{vmatrix} = \frac{d}{dz}\begin{vmatrix} A(z)e^{-j\beta_a z} \\ B(z)e^{-j\beta_b z} \end{vmatrix} = \begin{vmatrix} e^{-j\beta_a z}\dfrac{d}{dz}A(z) - j\beta_a A(z)e^{-j\beta_a z} \\ e^{-j\beta_b z}\dfrac{d}{dz}B(z) - j\beta_b B(z)e^{-j\beta_b z} \end{vmatrix} \tag{6.84}$$

$$= \begin{vmatrix} -je^{-j\beta_a z}KBe^{-j2\delta \cdot z} - j\beta_a A(z)e^{-j\beta_a z} \\ -je^{-j\beta_b z}KAe^{j2\delta \cdot z} - j\beta_b B(z)e^{-j\beta_b z} \end{vmatrix} = \begin{vmatrix} -j\beta_a & -jK \\ -jK & -j\beta_b \end{vmatrix} \begin{vmatrix} A(z)e^{-j\beta_a z} \\ B(z)e^{-j\beta_b z} \end{vmatrix} \Rightarrow$$

$$\frac{d}{dz}\begin{vmatrix} E_1(z) \\ E_2(z) \end{vmatrix} = \begin{vmatrix} -j\beta_a & -jK \\ -jK & -j\beta_b \end{vmatrix} \begin{vmatrix} E_1(z) \\ E_2(z) \end{vmatrix} \tag{6.85}$$

We now postulate that this coupled waveguide structure has guided modes with a well-defined propagation constant

$$\begin{vmatrix} E_1(z) \\ E_2(z) \end{vmatrix} = \begin{vmatrix} E_1(0) \\ E_2(0) \end{vmatrix} e^{j\beta \cdot z} \tag{6.86}$$

Substituting the guided wave solution into the above matrix equation gives

$$j\beta \begin{vmatrix} E_1(z) \\ E_2(z) \end{vmatrix} = \begin{vmatrix} -j\beta_a & -jK \\ -jK & -j\beta_b \end{vmatrix} \begin{vmatrix} E_1(z) \\ E_2(z) \end{vmatrix} \tag{6.87}$$

so

$$-j(\beta_a + \beta)E_1 - jKE_2 = 0 \tag{6.88}$$

$$-jKE_1 - j(\beta_b + \beta)E_2 = 0 \tag{6.89}$$

Non-trivial solutions require a zero determinant

$$\begin{aligned} &-(\beta_a + \beta)E_1(\beta_b + \beta)E_2 + K^2 E_2 E_1 = 0 \Rightarrow \\ &-\beta^2 - \beta(\beta_a + \beta_b) - \beta_a\beta_b + K^2 = 0 \Rightarrow \end{aligned} \tag{6.90}$$

$$\begin{aligned} &\beta = \frac{\beta_a + \beta_b}{2} \pm \frac{1}{2}\sqrt{(\beta_a + \beta_b)^2 - 4\beta_a\beta_b + 4K^2} \\ &\Rightarrow \beta = \frac{\beta_a + \beta_b}{2} \pm \frac{1}{2}\sqrt{(\beta_a - \beta_b)^2 + 4K^2} \end{aligned} \tag{6.91}$$

The corresponding eigenvectors are

$$\begin{vmatrix} E_1(z) \\ E_2(z) \end{vmatrix} = \begin{vmatrix} -K \\ \frac{1}{2}(\beta_a - \beta_b) + \frac{1}{2}\sqrt{(\beta_a - \beta_b)^2 + 4K^2} \\ 1 \end{vmatrix} \cdot e^{j\beta_+ \cdot z} \tag{6.92}$$

and

$$\begin{vmatrix} E_1(z) \\ E_2(z) \end{vmatrix} = \begin{vmatrix} -K \\ \frac{1}{2}(\beta_a - \beta_b) - \frac{1}{2}\sqrt{(\beta_a - \beta_b)^2 + 4K^2} \\ 1 \end{vmatrix} \cdot e^{j\beta_- \cdot z} \tag{6.93}$$

In the degenerate case, $\beta_a = \beta_b$, the eigenvectors simplify to

$$\begin{vmatrix} E_1(z) \\ E_2(z) \end{vmatrix} = \begin{vmatrix} 1 \\ 1 \end{vmatrix} \cdot e^{j\beta_- \cdot z} \tag{6.94}$$

and

$$\begin{vmatrix} E_1(z) \\ E_2(z) \end{vmatrix} = \begin{vmatrix} -1 \\ 1 \end{vmatrix} \cdot e^{j\beta_+ \cdot z} \tag{6.95}$$

We see that the symmetric eigenmode has a lower longitudinal wave vector than the anti-symmetric eigenmode[b]. This means that the effective index is lower for the symmetric mode, which is reasonable when we consider the shape of the eigenmodes. The symmetric mode has a higher field in the area between the guides (the fields of the symmetric mode add in this region, while the fields of the antisymmetric mode subtract). The index in the region between the guides is lower than the index in the cores, so the symmetric mode effectively sees a lower index than the antisymmetric mode.

In the non-degenerate case, the eigenmodes carry different amount of energy in the individual guided modes, i.e. the eigenmodes are not symmetric and antisymmetric. In the extreme case of vanishing coupling, the eigenmodes of the total systems simply equal the modes of the individual guides as we would expect. An interesting situation arises when the two coupled waveguide modes are degenerate at a specific wavelength, but has different dispersion characteristics.

In the non-degenerate case with differences in wave vector that are small compared to the coupling coefficient, we can write

$$\beta = \frac{\beta_a + \beta_b}{2} \pm \frac{1}{2}\sqrt{(\beta_a - \beta_b)^2 + 4K^2} \approx \frac{\beta_a + \beta_b}{2} \pm K \tag{6.96}$$

The eigenvectors are

$$\begin{vmatrix} E_1(z) \\ E_2(z) \end{vmatrix} = \begin{vmatrix} -K \\ \frac{1}{2}(\beta_a - \beta_b) + K \\ 1 \end{vmatrix} \cdot e^{j\beta_+ \cdot z} = \begin{vmatrix} -1 \\ 1 + \frac{\beta_a - \beta_b}{2K} \end{vmatrix} \cdot e^{j\beta_+ \cdot z} \tag{6.97}$$

and

$$\begin{vmatrix} E_1(z) \\ E_2(z) \end{vmatrix} = \begin{vmatrix} 1 \\ 1 - \frac{\beta_a - \beta_b}{2K} \end{vmatrix} \cdot e^{j\beta_- \cdot z} \tag{6.98}$$

We see that for small wave vector differences the eigenmodes are no longer symmetric and antisymmetric, but have unequal power in each guide.

[b] This splitting of the values of the longitudinal wave vectors due to coupling is observed in any set of coupled oscillators, including coupled mechanical oscillators and molecules (coupled atoms).

6.4.3 Conceptual Description of Directional Couplers Based on Eigen Modes

The eigenmodes we have found for the coupling sections of directional couplers provide a simple description of the operation of directional couplers. Consider the Directional Coupler of Fig. 6.14. The modes of the waveguides that make up the coupler interact in the coupling section and create two eigenmodes. The amplitudes of the even and odd eigen modes can be expressed in terms of the mode of the single-mode waveguides

$$A_e(x) = A_{sm}(x-d) + A_{sm}(x+d) = e^{-\frac{(x-d)^2}{\omega^2}} + e^{-\frac{(x+d)^2}{\omega^2}} \tag{6.99}$$

$$A_o(x) = A_{sm}(x-d) - A_{sm}(x+d) = e^{-\frac{(x-d)^2}{\omega^2}} - e^{-\frac{(x+d)^2}{\omega^2}} \tag{6.100}$$

These expressions are approximate. The even and odd eigenmodes will not be exactly equal to the sum and difference of the waveguide modes, but for weakly coupled waveguides this approximation is very good. The waveguide modes can then be expressed in terms of the eigenmodes

$$A_{sm}(x-d) = \frac{1}{2}\cdot(A_e + A_o) = e^{-\frac{(x-d)^2}{\omega^2}} \tag{6.101}$$

$$A_{sm}(x+d) = \frac{1}{2}\cdot(A_e - A_o) = e^{-\frac{(x+d)^2}{\omega^2}} \tag{6.102}$$

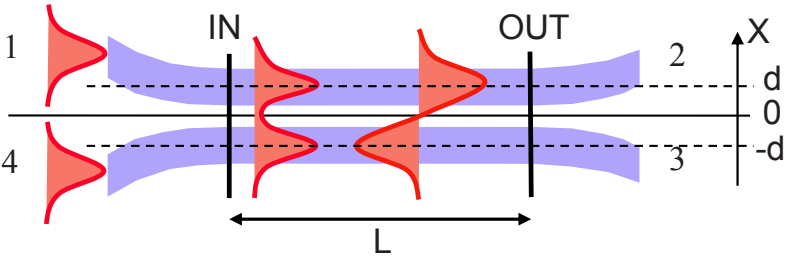


Figure 6.14. In the directional coupler, the modes of two fibers are brought close enough that they interact to create two eigenmodes of the coupling section. One of the modes is even and one is odd.

Now we assume that we have an input field that can be written

$$E_{in}(x) = E_{in0}\left[C_1 A_e(x) + C_2 A_o(x)\right] \tag{6.103}$$

The output field consists of the same two eigenmodes (the definition of a mode is that it propagates without changing), but the phase difference between the two eigenmodes has changed due to the fact that they have different propagation constants on the waveguide.

Let's assume that the propagation-constant difference is $\Delta\beta$. The output amplitude is then

$$E_{out}(x) = E_{in0}\left[C_1 A_e(x) + C_2 A_o(x) e^{j\Delta\beta L}\right] \tag{6.104}$$

or in terms of the waveguide modes

$$E_{out}(x) = E_{in0}\left[A_{sm}(x-d)\left\{C_1 + C_2 e^{j\Delta\beta L}\right\} + A_{sm}(x+d)\left\{C_1 - C_2 e^{j\Delta\beta L}\right\}\right] \tag{6.105}$$

The output on port 2 is approximately equal to the amplitude in the upper wave guide ($x=d$), so we can write

$$E_2 \approx E_{in0} \cdot A_{sm}(x-d)\left(C_1 + C_2 e^{j\Delta\beta L}\right) \tag{6.106}$$

The power on port 2 is then

$$P_2 \cong \left|E_{in0}\right|^2 \cdot \left(C_1^2 + C_2^2 + C_1 C_2 \cos(\Delta\beta L)\right) \tag{6.107}$$

Likewise we find the amplitude and power on port 3

$$E_3 \approx E_{in0} \cdot A_{sm}(x+d)\left(C_1 - C_2 e^{j\Delta\beta L}\right) \tag{6.108}$$

$$P_3 \cong \left|E_{in0}\right|^2 \cdot \left(C_1^2 + C_2^2 - C_1 C_2 \cos(\Delta\beta L)\right) \tag{6.109}$$

We see that these expressions are corresponding to the formulas we found earlier through detailed calculations. This shows that the simple and conceptually powerful eigenmode picture of the directional coupler is capable of explaining all aspects of this device.

6.5 Optical Devices Based on Directional Couplers

As mentioned in the introduction, the directional coupler is used in many, if not most, fiber-optical systems. In this section we describe a few such applications. The list is by no means exhaustive, but rather a selected set that demonstrates the power of the directional coupler.

6.5.1 Modulators and Switches Based on Directional Couplers

Directional couplers can be used as light modulators and switches. Consider the symmetric directional coupler of Fig. 6.15. Its symmetry guarantees that the waveguides are degenerate when no voltage is applied. The length is chosen for complete power transfer, i.e. $KL=\pi/2$. An optical signal on guide A will then be transferred to guide B. The signal can be switched back into guide A by changing the propagation constant in either or both guides. If the switch is made in an electrooptic material like a ferro-electric or electrooptic polymer, then the effective index of the propagating modes can be modulated by applying a voltage to either or both guides.

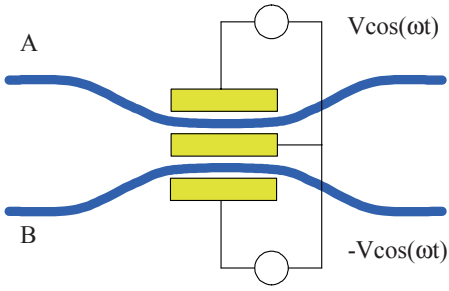


Figure 6.15. Directional coupler as modulator. The modulator is designed so that without a signal applied, the two waveguides are degenerate and the coupling length is chosen for complete power transfer. With a signal applied, the degeneracy is broken, and the power stays in the input guide.

From the expressions for the amplitudes in the two guides when $A(0)=A_0$ and $B(0)=0$,

$$A = A_0 \cdot e^{j\delta \cdot z}\left\{\cos\left[\sqrt{K^2+\delta^2}\,z\right] - j\frac{\delta}{\sqrt{K^2+\delta^2}}\sin\left[\sqrt{K^2+\delta^2}\,z\right]\right\} \tag{6.110}$$

$$B = -jA_0 e^{-j\delta \cdot z}\frac{K}{\sqrt{K^2+\delta^2}}\sin\left[\sqrt{K^2+\delta^2}\,z\right] \tag{6.111}$$

we see that we need a shift of

$$\sin\left[\sqrt{K^2+\delta^2}\,L\right] = 0 \Rightarrow \sqrt{K^2+\delta^2}\,L = \pi \Rightarrow \left(K^2+\delta^2\right)L^2 = \pi^2 \Rightarrow \tag{6.112}$$

$$\delta \cdot L = \sqrt{\pi^2 - \frac{\pi^2}{4}} = \frac{1}{2}(\beta_a - \beta_b)L = \frac{\sqrt{3}}{2}\pi \tag{6.113}$$

to switch the optical signal back to guide *A*. If the propagation difference is increased further, some of the power is again in waveguide *B*, but the power transfer is not complete. This is illustrated in Fig. 6.16.

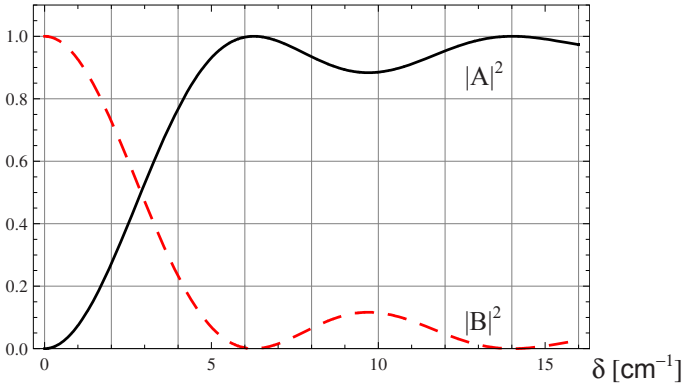


Figure 6.16. Directional-coupler modulator. The plot shows the power in the two output guides as a function of δ, the difference in propagation constant between the two modes. The length of the coupler is chosen for complete power transfer when δ=0. With no voltage applied, the optical power is completely transferred between guides. Applying a voltage decreases the transfer.

The change in effective index of the modes is approximately equal to the shift in refractive index caused by the electrooptic effect, so

$$\delta \approx \frac{2\pi}{\lambda_0}\Delta n \approx \frac{2\pi}{\lambda_0} n^3 r E \tag{6.114}$$

where r is the appropriate electrooptic coefficient for the modulator geometry.

Now let's use these expressions to find approximately how long an interaction region will be required for a high-contrast electrooptic directional-coupler switch. For any integrated optics device, the size is an important figure of merit. It tells us how many devices can be integrated on a single substrate, and ultimately determines the complexity of the systems we can build.

The minimum length of the directional-coupler is

$$L = \frac{\pi\sqrt{3}}{2\delta} \approx \frac{\pi\sqrt{3}}{2}\frac{\lambda_0}{2\pi}\frac{1}{\Delta n} \approx \frac{\sqrt{3}}{4}\frac{\lambda_0}{n^3 r E} \tag{6.115}$$

Assuming the following parameters; λ_0*=1.55 μm, n=3.5, R=20 pm/V*, and *E=1.0 V/μm*, we find L_{min}*=780 μm*. This is a relatively long interaction length, and it complicates integration of large switches based on directional couplers.

6.5.2. Power Combiners and Filters Based on Directional Couplers

Directional couplers are used in a number of passive optical devices including sensors, power dividers, power splitters, and filters. In these devices we exploit the wavelength dependencies of the directional coupler, which are explicit in the formulas for the coupling coefficient and the correction to the propagation constant of the guides.

$$K_{ab,ba} = \frac{\omega \cdot \varepsilon_0}{4} \int_{-\infty}^{\infty} u_b(x) u_a(x) \left(n_c^2 - n_{a,b}^2\right) \cdot dx \tag{6.116}$$

$$M_{a,b} = \frac{\omega \cdot \varepsilon_0}{4} \int_{-\infty}^{\infty} u_{a,b}^2(x) \left(n_c^2 - n_{a,b}^2\right) \cdot dx \tag{6.117}$$

To achieve a sharp filter function, we use different waveguides, so that the coupling coefficient, the propagation-constant corrections, as well as the wavelength dependencies of the individual modes all contribute to create non-degenerate coupling for all but a narrow band of frequencies.

If the waveguides are identical, the only wavelength dependence is in the coupling constant, which leads to a relatively wide filter function as shown in Fig. 6.17. This relatively broadband operation is desirable for switching and some filter functions (e.g. power combining). The majority of filter applications require more narrow-band operation than what is shown in Fig. 6.17.

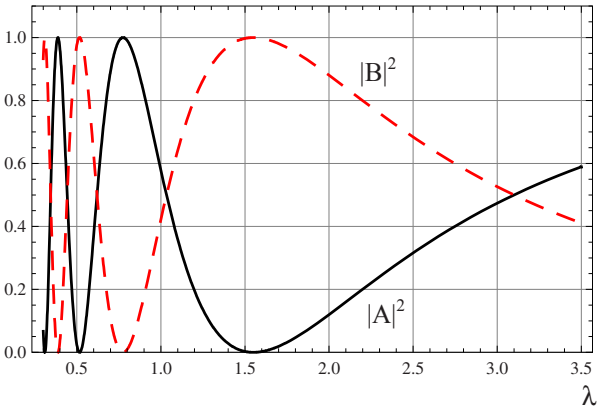


Figure 6.17. Directional coupler as filter. The powers in the two output guides are plotted as a function of wavelength. The length of the coupler is chosen for complete power transfer at λ=1.55um. The relatively weak wavelength dependence of the symmetric directional coupler can be used for power combining in optical amplifiers.

6.6 Periodic Waveguides – Bragg Filters

We will now apply coupled mode theory to counter-propagating waves in a single-mode waveguide with a periodic corrugation as shown in Fig. 6.18.

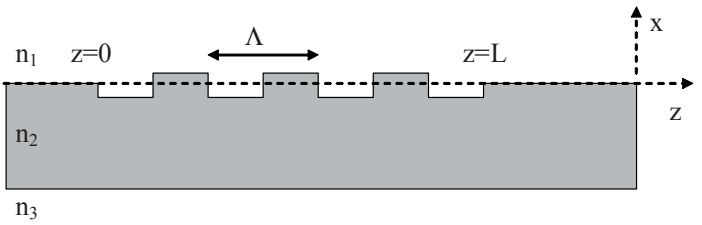


Figure 6.18. Waveguide with periodic corrugation in one of the core-cladding interfaces. The waveguide is single mode, and we make the assumption that the only significant coupling is between counter-propagating guided modes (i.e. radiation modes are unimportant).

The corrugation is scalar and we don't expect coupling between TE and TM modes, so in the following treatment we'll consider TE modes. We start by describing the field in the corrugated structure as a sum of the forward and backward propagating modes

$$E_y = A(z)u(x)\exp\left[j(\omega\cdot t-\beta\cdot z)\right]+B(z)u(x)\exp\left[j(\omega\cdot t+\beta\cdot z)\right] \tag{6.118}$$

where A and B are the amplitudes of the forward and backward propagating waves, and $u(x)$ is the mode profile.

The perturbation in the corrugated region is

$$\vec{P}_{pert} = \Delta n(x,z)^2\varepsilon_0\vec{E} \tag{6.119}$$

We substitute the expression for the field into this expression to get

$$\begin{aligned} &P_{pert} = \frac{1}{2}\Delta n^2\varepsilon_0\left\{Au(x)\exp\left[j(\omega\cdot t-\beta z)\right]+B(z)u(x)\exp\left[j(\omega\cdot t+\beta z)\right]\right\} \\ &\Rightarrow P_{pert} = \frac{1}{2}\Delta n^2\varepsilon_0 e^{j\omega\cdot t}e^{-j\beta\cdot z}\left\{A+Be^{-j2\beta\cdot z}\right\}\cdot u(x) \end{aligned} \tag{6.120}$$

Recall the fundamental coupled mode equation (Eq. 6.47)

$$\begin{aligned} &-\frac{dA_i^+}{dz}\cdot\exp\left[-j(\beta_i z-\omega\cdot t)\right]+\frac{dA_i^-}{dz}\cdot\exp\left[j(\beta_i z+\omega\cdot t)\right]+c.c \\ &=\frac{-j}{2\omega}\frac{\partial^2}{\partial t^2}\int_{-\infty}^{\infty}P_{pert}(x,z,t)\cdot u_i(x)dx \end{aligned} \tag{6.121}$$

which simplifies to

$$-\frac{dA}{dz}+\frac{dB}{dz}\cdot e^{j2\beta\cdot z}=\frac{-j\omega\cdot\varepsilon_0}{4}\left\{A+Be^{j2\beta\cdot z}\right\}\int_{-\infty}^{\infty}\Delta n^2u^2(x)dx \tag{6.122}$$

We will now assume that the corrugation has a square-wave shape as indicated in Fig. 6.18. The general conclusions are not dependent on the exact shape, so the following treatment, with appropriate adjustments, is valid also for non-square corrugations. The square-wave corrugations can be expressed as a series in the following form

$$\Delta n^2(x,z)=\Delta n^2\sum_m c_m e^{j\frac{2m\pi\cdot z}{\Lambda}} \tag{6.123}$$

$$c_m=\begin{cases}\dfrac{-j}{m\pi} & m \text{ odd}\\ 0 & m \text{ even}\end{cases} \tag{6.124}^{c}$$

By comparing this expression to the above coupled mode equations, we realize that only modes that are close to phase matched will experience significant coupling. In other words, we need only keep terms of the same periodicity. In a range of wave vectors around $m\dfrac{2\pi}{\Lambda}$ the equations can be simplified to

$$\frac{dA}{dz}=\frac{j\omega\cdot\varepsilon_0}{4}Be^{j2\beta\cdot z}c_m e^{-j\frac{2m\pi\cdot z}{\Lambda}}\int_{-\infty}^{\infty}\Delta n^2u^2(x)dx \tag{6.125}$$

and

$$\frac{dB}{dz}=\frac{j\omega\cdot\varepsilon_0}{4}Ae^{-j2\beta\cdot z}c_m e^{j\frac{2m\pi\cdot z}{\Lambda}}\int_{-\infty}^{\infty}\Delta n^2u^2(x)dx \tag{6.126}$$

These equations are on a form similar to the ones describing the directional coupler

---

[c] Here we are making the implicit assumption that the lengths of the high and low index regions (L_H and L_L) are the same. It might seem intuitive that the high-index region should be shorter ($L_H<L_L$), but if we insist on matching both physical ($\Lambda=L_H+L_L$) and optical lengths ($n\cdot\Lambda=(n+\Delta n)L_H+(n-\Delta n)L_L$), then it follows that the two regions must have the same length ($L_H = L_L$).

$$\frac{dA}{dz} = K^* B e^{j2\Delta\beta \cdot z} \tag{6.127}$$

$$\frac{dB}{dz} = KA e^{-j2\Delta\beta \cdot z} \tag{6.128}$$

where

$$K = \frac{j\omega \cdot \varepsilon_0}{4} c_m \int_{-\infty}^{\infty} \Delta n^2 \cdot u^2(x) \cdot dx \tag{6.129}$$

$$\Delta\beta = \beta - \frac{m\pi}{\Lambda} \tag{6.130}$$

6.6.1 Energy Conservation in Counter Propagating Waves

Let us check energy conservation in the systems of equations we have found for modes in a Bragg grating. We start by deriving expression for the energies in the forward and backward propagating waves. Based on Eqs. 6.127 and 6.128 we can write

$$\begin{aligned} &\frac{d}{dz}|A|^2 = \frac{d}{dz} AA^* = A\frac{d}{dz}A^* + A^*\frac{d}{dz}A \\ &= AB^* \cdot Ke^{-j2\Delta\beta \cdot z} + A^* B \cdot K^* e^{j2\Delta\beta \cdot z} \end{aligned} \tag{6.131}$$

and

$$\begin{aligned} &\frac{d}{dz}|B|^2 = \frac{d}{dz} BB^* = B\frac{d}{dz}B^* + B^*\frac{d}{dz}B \\ &= BA^* \cdot K^* e^{j2\Delta\beta \cdot z} + B^* A \cdot K_{ba} e^{-j2\Delta\beta \cdot z} \end{aligned} \tag{6.132}$$

The difference between the rate of change in the forward-propagating and backward-propagating energy is then

$$\begin{aligned} &\frac{d}{dz}|A|^2 - \frac{d}{dz}|B|^2 = AB^* \cdot Ke^{-j2\Delta\beta \cdot z} + A^* B \cdot K^* e^{j2\Delta\beta \cdot z} = \\ &BA^* \cdot K^* e^{j2\Delta\beta \cdot z} + B^* A \cdot Ke^{-j2\Delta\beta \cdot z} = 0 \end{aligned} \tag{6.133}$$

We see that the rate of change in forward-propagating energy is exactly balanced by the rate of change in backward-propagating energy, which is the correct result for loss-less, counter-propagating waves.

6.6.2 Modes of the Bragg Grating

The set of equations describing the modes of the Bragg Grating (Eqs. 127-130) can now be solved. Assuming that the forward propagating mode has an amplitude A_0 at $z=0$, and that the backward propagating wave is zero at $z=L$, we find

$$A = A_0 \cdot e^{j\Delta\beta \cdot z} \frac{-\Delta\beta \cdot \sinh[S(z-L)] + jS\cosh[S(z-L)]}{-\Delta\beta \cdot \sinh[SL] + jS\cosh[SL]} \tag{6.134}$$

$$B = A_0 \cdot jK \cdot e^{-j\Delta\beta \cdot z} \frac{\sinh[S(z-L)]}{-\Delta\beta \cdot \sinh[SL] + jS\cosh[SL]} \tag{6.135}$$

where

$$S = \sqrt{K^2 - \Delta\beta^2} \tag{6.136}$$

When $\Delta\beta=0$ this simplifies to

$$A = A_0 \cdot \frac{\cosh[K(z-L)]}{\cosh[KL]} \tag{6.137}$$

$$B = A_0 \cdot \frac{\sinh[K(z-L)]}{\cosh[KL]} \tag{6.138}$$

These expressions are plotted in Fig. 6.19 for two different lengths of the corrugated region.

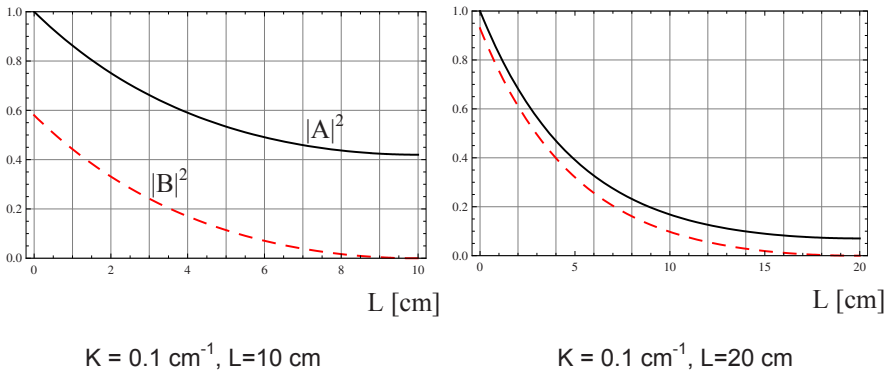


Figure 6.19. Power coupling between counter propagating modes in a corrugated waveguide. The powers in the two waveguides are plotted as a function of length along the waveguide in the corrugated region.

Figure 6.19 shows that as the corrugated region gets longer, more of the power is coupled into the reflected wave. This is what we would expect; a longer grating leads to smaller transmission and larger reflection, which means that the forward and backward propagating fields are closer in magnitude. In the extreme case of an infinite grating, the forward and backward propagating waves are equal at all points in the grating.

6.6.3 One-Dimensional Photonic Bandgaps

The amplitudes we have found for the Bragg reflector (Eqs. 6.134 and 6.135) shows that the general solution (Eq. 6.118) consist of forward and backward propagating waves with the following propagation constants

$$\begin{aligned} \beta_{bragg} &= \beta - \Delta\beta \pm S = \beta - \beta + \beta_0 \pm S \Rightarrow \\ \beta_{bragg} &= \frac{m\pi}{\Lambda} \pm j\sqrt{K^2 - \left(\beta - \frac{m\pi}{\Lambda}\right)^2} \end{aligned} \tag{6.139}$$

We see that when the unperturbed propagation constant, β, is far from being resonant with the fundamental or higher order variations of the Bragg grating $\left(\left|\beta - \frac{m\pi}{\Lambda}\right| >> K\right)$, then the expression simplifies to $\beta_{bragg} = \pm\beta$. In other words, off resonance the solutions are forward and backward propagating harmonic waves. Close to resonance $\left(\left|\beta - \frac{m\pi}{\Lambda}\right| \approx K\right)$ the solutions are no longer simple harmonic waves, but have complex-valued propagation constants.

To better understand the implications of the complex-valued propagation constant, we consider the special case of *m=1*. Equation 6.139 then becomes

$$\frac{\beta_{bragg}\Lambda}{\pi} = 1 \pm j\sqrt{\left(\frac{K\cdot\Lambda}{\pi}\right)^2 - \left(\frac{\beta\cdot\Lambda}{\pi} - 1\right)^2} \tag{6.140}$$

The real and imaginary parts of this expression are plotted in Fig. 6.20 for $\frac{K\cdot\Lambda}{\pi} = 0.1$. Not all solutions to Eq. 6.140 give realizable field distributions. The sign in front of the square root must be picked correctly in the different ranges of β values to yield forward and backward propagating waves as shown in Fig. 6.20.

The plots show that when the wave vector of the unperturbed guide is close to that of the periodic grating, the propagating fields have a *z*-dependence with an imaginary part. This means that the fields are not oscillating, but are exponentially damped. In these resonant regions we therefore cannot have an undamped propa-

gating mode. We say that the waveguide structure has a bandgap. This concept will be further developed in Chapters 14 and 15.

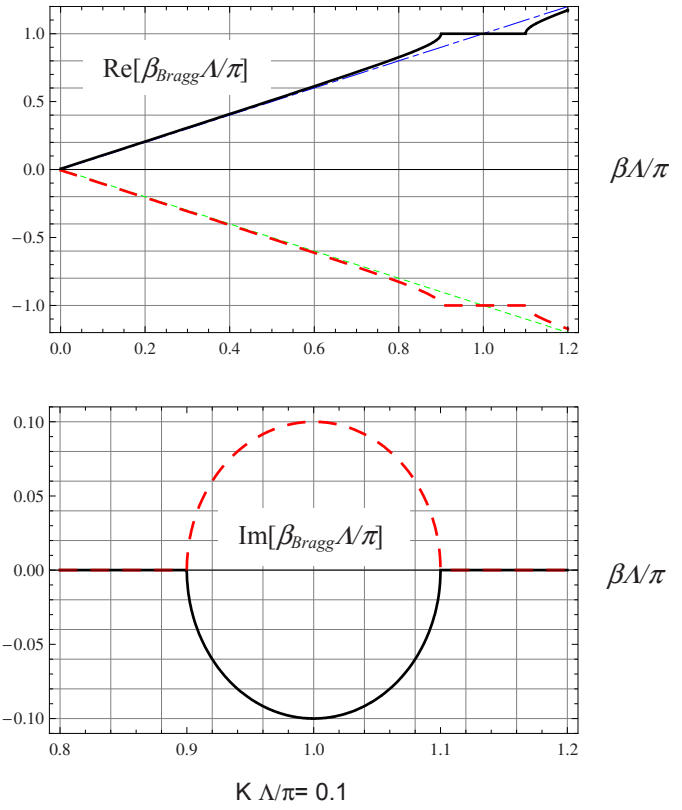


Figure 6.20. Normalized real and imaginary parts of the longitudinal wavevector of forward (solid) and backward (dashed) propagating modes in a periodically corrugated waveguide vs. longitudinal wavevector of the unperturbed modes (dot-dashed and dotted lines in upper graph). In the region close to the Bragg vector, the longitudinal wavevector has an imaginary part, and the mode is evanescent. We call this region the bandgap of the waveguide structure.

6.6.4 Bragg Filters

The expressions we have found for the field amplitudes in the periodically corrugated waveguide allow us to calculate the reflection and transmission spectra of the Bragg grating. For example, the field reflection is simply the ratio of the forward propagating and backward propagating wave at the input to the Bragg section:

$$r = \frac{B(0)}{A(0)} = \frac{A_0 \cdot jK \cdot \dfrac{\sinh[SL]}{-\Delta\beta \cdot \sinh[SL] + jS\cosh[SL]}}{A_0 \cdot \dfrac{-\Delta\beta \cdot \sinh[SL] + jS\cosh[SL]}{-\Delta\beta \cdot \sinh[SL] + jS\cosh[SL]}} \Rightarrow$$
$$r = \frac{jK \cdot \sinh[SL]}{-\Delta\beta \cdot \sinh[SL] + jS\cosh[SL]} \tag{6.141}$$

The power reflectance corresponding to this field reflectivity is plotted in Fig. 6.21a for a Bragg grating with a coupling constant of $K=0.1cm^{-1}$ and a length of $L=10cm$. The reflectance is plotted on a logarithmic scale to clearly show the structure of the side bands. Figure 6.21b shows the phase of the reflections with the power reflectance overlaid for reference.

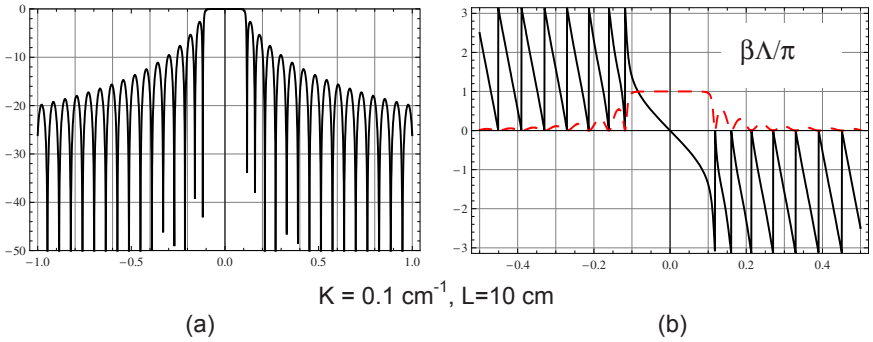


Figure 6.21 *(a) Reflectance of a Bragg grating as a function of detuning. The flat pass band is one of the unique characteristics of the Bragg filter. (b) Reflectance and phase-shift of a Bragg filter as a function of detuning. Note that the phase shift is close to linear throughout the pass band.*

Figure 6.21 shows that Bragg reflector has several desirable features. The pass band has a flat amplitude response and a linear phase variation that avoids distortion of reflected signals. The side band rejection can be improved by using weaker coupling or fewer periods, such that the product *KL* is reduced.

Bragg reflectors with the characteristics reflectance of Fig. 6.21 are implemented as dielectric stacks or as waveguide gratings. Fiber-Bragg grating are fabricated by illuminating standard fiber with near-UV light in a standing wave pattern (set up by interference between two laser beams). These fibers can therefore be made relatively simply with varying Bragg frequencies. The flat pass band is another strong advantage of fiber-Bragg filters. For these reasons, the fiber-Bragg filters are among the leading candidates for implementation of multiplexers for optical communication systems based on Wavelength Division Multiplexing (WDM).

6.7 Waveguide Modulators

Because the optical field is strongly confined in optical waveguides, they in many ways represent the ideal environment for manipulating light. In particular this is true for propagation of light over long distances, but also for more localized operations like modulation, switching etc. If we compare free-space optical modulators to waveguide modulators, and we will see that the absence of diffraction in the waveguide leads to reduced modulation voltage, reduced size, and increased bandwidth of the modulator.

6.7.1 Mach-Zender Modulators

A very popular waveguide modulator geometry is the Mach-Zender modulator, shown in Fig.6.22 and 6.23. In the Mach-Zender, the incoming optical field is split in two parts, usually by a Y-coupler. The two parts are phase shifted with respect to each other, and then recombined in a second Y-coupler.

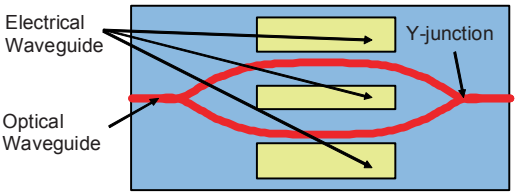


Figure 6.22. Top-view of Mach-Zender modulator. The optical waveguide is split in two to create an interferometer. The modes in the two arms of the interferometer experience a phase modulation of equal magnitude, but opposite sign. The terminations of the electrical waveguide are not shown.

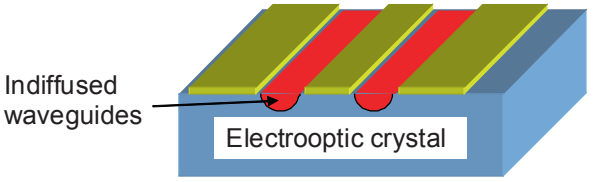


Figure 6.23. Layout of waveguide optical modulator for high frequency operation. The optical waveguides and the electrodes are fabricated on the electrooptic substrate using techniques borrowed from integrated circuit technology.

We assume that assuming that the output Y-coupler is well designed, i.e. incoming power on the single arm will be split equally in the two arms. Simple energy con-

servation considerations then tells us that the mode amplitude at the output single-mode waveguide is

$$A_{out} = \frac{1}{\sqrt{2}}\left(A_1 + A_2 e^{j\phi}\right) = \frac{e^{j\phi/2}}{\sqrt{2}}\left(A_1 e^{-j\phi/2} + A_2 e^{j\phi/2}\right) \tag{6.142}$$

where A_1 and A_2 are the mode amplitudes of the upper and lower waveguides in the center section of the Mach-Zender modulator. The output power is then

$$\begin{aligned} P_{out} &= A_{out} A_{out}^* = \frac{1}{2}\left(A_1 e^{-j\phi/2} + A_2 e^{j\phi/2}\right)\left(A_1^* e^{j\phi/2} + A_2^* e^{-j\phi/2}\right) \\ &= \frac{1}{2}\left(A_1^2 + A_2^2 + A_1 A_2 e^{-j\phi} + A_1 A_2 e^{j\phi}\right) = \frac{1}{2}\left(A_1^2 + A_2^2\right) + A_1 A_2 \cos\phi \end{aligned} \tag{6.143}$$

If the input Y-coupler is also well designed, the amplitudes in the two arms are the same and we find

$$P_{out} = \frac{1}{2} P_{in}\left(1 + \cos\phi\right) \tag{6.144}$$

We see that, as in any interferometer, the output is a harmonic function of the phase difference between the two modes. If the amplitudes in the two arms are different, the modulation waveform has less than perfect contrast. This is illustrated in Fig. 6.24.

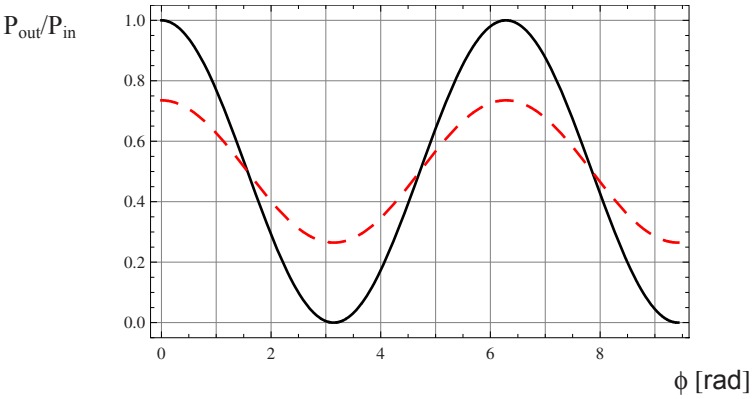


Figure 6.24. Modulation curves for Mach-Zender modulators. The solid curve shows a modulator, in which the power is split evenly between the two arms. The dashed curve is valid for a modulator with a power-splitting ratio of 1:4 in the input Y-coupler and an even split in the output Y-coupler.

This type of waveguide modulator has several advantages over bulk modulators:

1) The fabrication is done with diffusion and lithography on a flat substrate, i.e. the fabrication is compatible with modern integrated-circuit technology.
2) The electrodes can be placed very accurately with respect to the waveguides to optimize the modulating field strength in the waveguides without introducing excessive losses due to absorption in the metal electrodes.
3) The electrodes can be designed to act as an electric waveguide phase-matched to the optical waveguide such that high-speed operation can be obtained.

6.7.2 Figures of Merit for Optical Modulators

The modulation curve for the Mach-Zender illustrates the important parameters of optical modulators. The **modulation index** is defined as the ratio of the maximum change in transmitted power to the maximum transmitted power

$$\eta = \frac{P_{out,\max} - P_{out,\min}}{P_{out,\max}} \tag{6.145}$$

Sometimes you see the following alternative definition

$$\eta = \frac{P_{out,\max} - P_{out,\min}}{\left(P_{out,\max} + P_{out,\min}\right)/2} \tag{6.146}$$

Often this information is expressed as a **contrast ratio** instead

$$Contrast = \frac{P_{out,\min}}{P_{out,\max}} \tag{6.147}$$

which we like to express in decibels if it is small

$$Contrast = 10\log\left[\frac{P_{out,\min}}{P_{out,\max}}\right] \tag{6.148}$$

The **insertion loss** is the ratio of the maximum transmitted power to the input power

$$L = \frac{P_{out,\max}}{P_{in}} \tag{6.149}$$

Other important parameters of optical modulators are discussed briefly below.

The bandwidth of an optical modulator depends ultimately on the physical effect used to create the phase modulation. In practice, however, we find that RC-time constants often are the determining factors. Traveling-wave optical modulators are therefore popular. In these modulators, the modulating field is traveling on an electrical waveguide structure with the same group velocity as the optical waveguide. The bandwidth of a traveling wave modulator is determined by the group-velocity mismatch between the electrical and optical wave.

The modulating **voltage**, required to create a phase shift of π (V_π), should be as low as possible. This is particularly important in high-frequency modulators. A related issue is **power consumption.**

Linearity is important for analog communication formats, e.g. for video signal distribution.

Polarization dependence is important in some applications, but modulators are most often used in a transmitter with a source of controllable polarization. Under these circumstances, polarization dependence is of little consequence.

Wavelength dependence is increasingly important as we start using wider and wider wavelengths bands for fiber optic communications.

Environmental Sensitivity (e.g. temperature sensitivity, shock sensitivity, robustness) is important in any commercial product , but it is particularly significant in telecommunication equipment which must pass very stringent reliability tests.

Compatibility with standard fiber, and other standards is very important both from the point of view of fabrication and sales.

In the final analysis, it is the **size and cost** of the modulator, which will determine its competitiveness in the market place.

6.7.3 Phase Modulation

The phase modulation required to operate the Mach-Zender can be accomplished through a number of physical effects.

Electrooptic effect

The electrooptic or Pockels effect is caused by the fact that the polarizability, and therefore the index of refraction, of some solids can be influenced by an applied electric field

$$\Delta n = c \cdot \vec{E} \tag{6.150}$$

where in general the electrooptic coefficient and the index change are tensors.

A simple symmetry argument shows that materials with inversion symmetry cannot be electrooptic. Assume that a material with inversion symmetry exhibits an electrooptic index change, $\Delta n = c \cdot E$. If the field is inverted, the index change must be the same, i.e. $\Delta n = c \cdot (-E)$, which means that $c=-c=0$.

In a crystal, the index change for a mode of a certain polarization will depend on the relative orientation of the crystal axes, the applied electric field, and the mode polarization. The optical-modulator designer must therefore consider the electrooptic tensor (relatively simple in cubic crystals like GaAs, more complex in ferro-electrics) and design the optical waveguide as well as the electrical electrodes such that the index change is optimized. A typical arrangement is shown in Fig. 6.25.

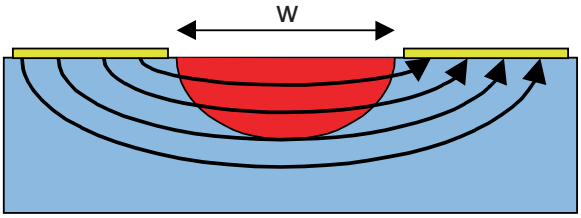


Figure 6.25. Detail of waveguide modulator showing the field distribution in the optical guide. The orientation of the electrooptic crystal is chosen to optimize the index change for a given applied modulating electrical signal.

As in our earlier treatment we make the approximation that both the modulating field and the optical mode are uniform, which leads to

$$\Delta\phi = \frac{2\pi}{\lambda} l \cdot n_0^3 r \cdot E = \frac{2\pi}{\lambda} l \cdot n_0^3 r \cdot \frac{V}{w} \Rightarrow V_\pi = \frac{\lambda}{2n_0^3 r} \frac{w}{l} \tag{6.151}$$

where w is the width of the optical waveguide, and V_π is the voltage required for a phase shift of π radians, which again is the required phase shift for high-contrast modulation.

We see that V_π decreases linearly with the length of the modulator. This should be compared to the square-root dependence on length, which is characteristic for free-space modulators. The difference is that the cross section of the waveguide is independent of its length, while in a free-space modulator, the length and thickness are related.

To compare this to the directional coupler modulator, we solve for the minimum length that will ensure 100% modulation index

$$L_{\min} = \frac{\lambda}{2n_0^3 r}\frac{w}{V_\pi} = \frac{\lambda}{2n_0^3 rE} \tag{6.152}$$

This is a factor of $\frac{2}{\sqrt{3}}$ larger than the minimum length of the directional-coupler modulator (but remember that in directional-coupler modulator, we assumed that all the power saw this phase shift, while in the one-sided M-Z we are only modulating half the power. Assuming the following modulator parameters: λ_0=1.55 μm, n = 3.5, R = 20 pm/V, and E = 1.0 V/ μm, we find $L_{min}=900\ \mu m$, again a relatively long interaction length.

Thermooptic effect

The thermooptic effect can create large index changes, but it is slow and requires relatively large power consumption.

Liquid crystals

Liquid crystals also exhibit large index changes. The speed are orders of magnitude lower than required for optical packet switching (nano second response times), but LC are significantly faster and less power hungry than thermooptical devices.

Band-edge effects

Large index changes and absorption changes can be created in semiconductor materials and waveguides by moving the band edge as a function of applied voltage or injected carriers. These effects are fast and large, but strongly dependent on wavelength.

Displacement

An effective way to change the refractive index is to physically displace materials. That is the approach taken in the Agilent Champagne Switch and in switches and modulators based on micromirror technology that will be discussed in later chapters.

6.7.4 Acoustooptic Modulators

Acoustic waves in bulk (as opposed to waveguides) crystals are widely used for optical modulation. The acoustic wave creates a periodic index variation in the crystal. This index grating can be used to reflect, deflect, disperse, or even frequency-shift (if the acoustic wave is a traveling wave, the optical frequency will be shifted due to the Doppler effect).

In waveguides or fibers, acoustic waves are used to couple modes. A typical example is shown in Fig. 6.26. The periodic force applied to the suspended fiber creates index differences that act as a long period Bragg grating. If the period of the grating matches the difference in wave vector of a guided and unguided mode, then the Bragg grating will provide the necessary phase matching for coupling between the modes, and the guided mode will be attenuated. Unwanted acoustic waves can under special circumstances present problems in optical communication systems (e.g. attenuation of solitons).

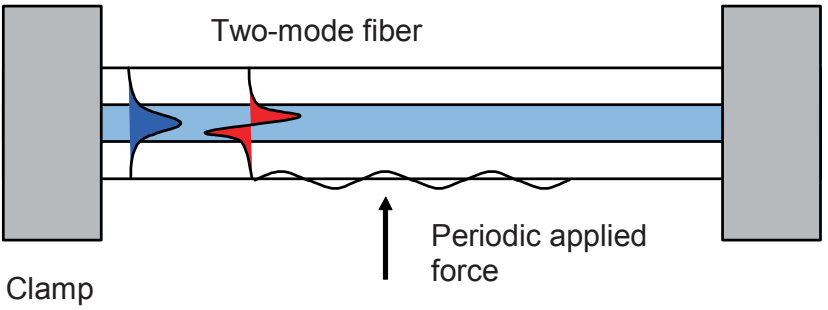


Figure 6.26. Schematic of acoustooptic coupling in waveguides. The periodic force on the waveguide creates an acoustic wave on the fiber that couples modes on the fiber if the periodicity of the acoustic field matches the longitudinal wavevector difference between the modes.

6.7.5 Modified Mach-Zender Modulators

The Y-couplers in the standard Mach-Zender are problematic because they couple light into radiation modes. In many applications, particularly ones that require integration of many modulators or switches (which is what we want to do, because high levels of integration is one of the primary motivations for integrated optics), radiation modes will lead to unwanted and unpredictable coupling between devices. To avoid this problem, directional couplers can be used instead of Y-couplers. This approach, illustrated in Fig.6.27, solves the problem of cross-talk caused by unpredictable radiation modes at the cost of increased wavelength dependence.

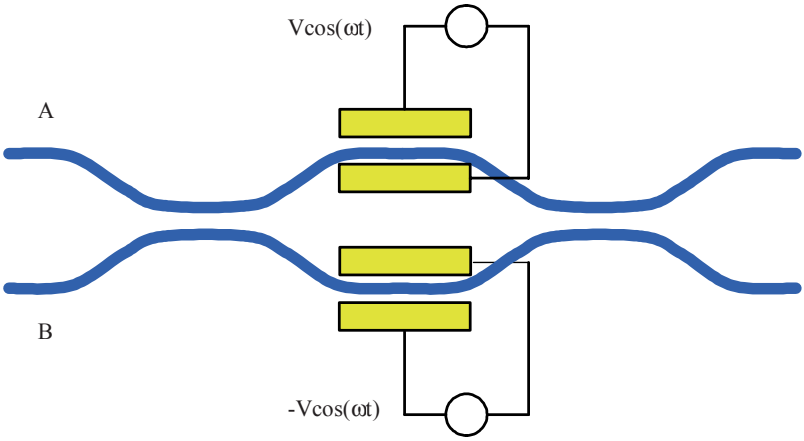


Figure 6.27. Modified Mach-Zender modulator with directional couplers instead of Y-couplers. The directional couplers are designed to split the power on each input guide evenly between the two output guides. Typically the two waveguides are degenerate. This layout has the advantage of acting both as a switch and as a modualtor, but at the cost of larger area requirement.

6.7.6 Directional Coupler Switches

We have already studied the directional-coupler modulator, shown again here in Fig. 6.28. Recall that this device is designed for complete power transfer with no modulating electrical signal applied. Once a voltage is applied, the optical signal on the input guide will not be transferred to the other guide. The signal can therefore be switched back and forth between the outputs by changing the propagation constant in either or both guides.

This modulator is unique because it can also be used as a switch, i.e. the power can be transferred between two well-defined guided modes (as opposed to between a guided mode and radiation modes as in the Mach-Zender). The modified Mach-Zender can also be used for switching, but that is because it incorporates directional couplers.

Note however that the directional coupler is a 1 by 2 switch (or 2 by 1), and cannot be used as a 2 by 2 switch. If there are optical signals at the same wavelength on both inputs on a directional coupler, the distribution of power at the output will depend on the relative phase of the two input signals. Such dependence on the phase of the input signals (as opposed to the relative phase of waves that both are generated by a y-splitter on chip) is of course unacceptable in practical switches.

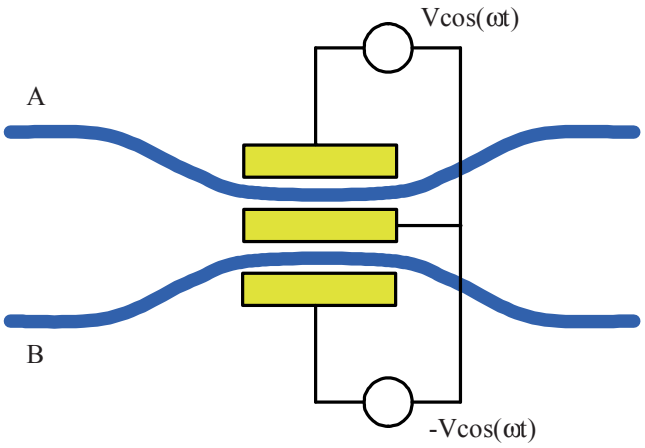


Figure 6.28. Directional coupler modulator. Without an applied electrical signal, the two waveguides are degenerate and the optical power is completely transferred between the guides. Once the degeneracy is broken, the power stays in the input guide.

6.7.7 Fabry-Perot Modulator

We can utilize resonant enhancement in an optical resonator to increase the effect of the relatively small index changes we can obtain by electrooptic means. The Fabry-Perot modulator of Fig. 6.29 is an illustration of this. This type of optical modulator can be implemented in many different ways in optical waveguides. In optical fibers, it is most practical to use fiber Bragg mirrors to create the resonator.

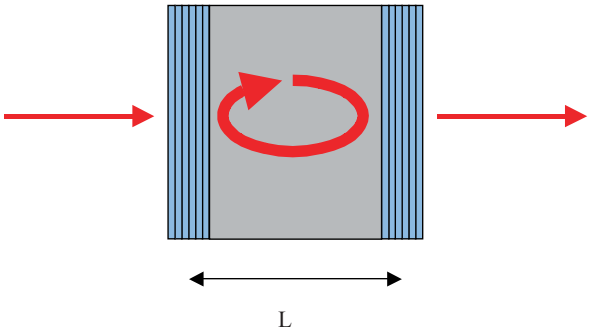


Figure 6.29. Fabry-Perot modulator. The large circulating field in the cavity means that a small change in index is sufficient to break the required phase-matching condition for the transmitted field.

The optical power transmission for a symmetric Fabry-Perot is (see Chapter 12.2.4 for a detailed derivation)

$$T = \frac{1}{1 + \frac{4R}{(1-R)^2}\sin^2\left(\frac{2\pi}{\lambda_0} nL\right)} \tag{6.153}$$

where R is the power reflectivity of the mirrors, n is the optical index of the cavity, L is the length of the cavity, and λ_0 is the vacuum wavelength of the light. This expression is plotted in Fig. 6.30, which shows that the higher the mirror reflectivities, the smaller the change in index required for high-contrast modulation. The limited free-spectral range of the F-P is an issue in many applications.

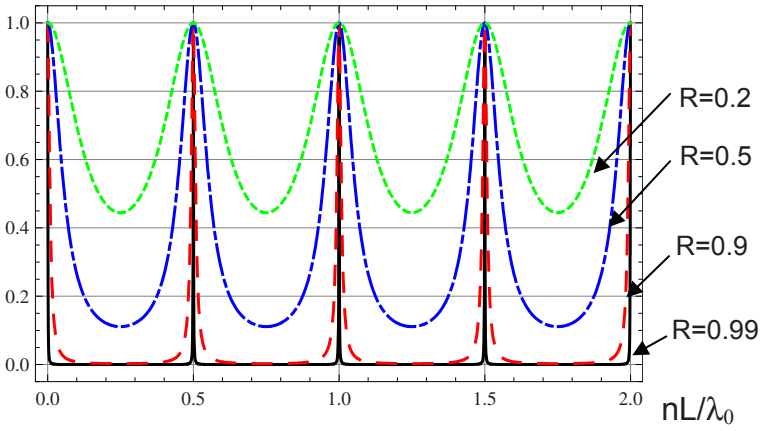


Figure 6.30. Transmission through a Fabry-Perot resonator as a function of normalized length with mirror reflectance as parameter: Solid: R=0.99, Dashed: R=0.9, Dot-dashed: R=0.5, Dotted: R=0.2.

6.7.8 Resonant Waveguide Coupling

Consider a directional coupler with optical power on both inputs. The power in the two output guides will depend on the relative phases of the input fields. This is most easily seen by considering the eigenmodes of the coupler. If the two input modes are of equal amplitude and in phase, the combination of the two modes matches the even eigenmode, and the input field distribution travels unperturbed along coupler. Something very similar happens if the two inputs are of equal amplitude and exactly out of phase (π phase difference). An arbitrary amplitude and phase distribution leads to an output that depends on the interference of the eigenmodes, i.e. it can be modulated by changing the index in the coupling region.

Now we will see how the interactions between two modes in a directional coupler can be used to create the waveguide equivalent of the Fabry-Perot resonator we just studied. Consider the ring-resonator structure with two resonantly-enhanced

directional couplers shown in Fig. 6.31. This device acts like a Fabry-Perot interferometer if we think of *Input1* as the input field of the F-P, *Output1* as the reflected field, and *Output2* as the transmitted field.

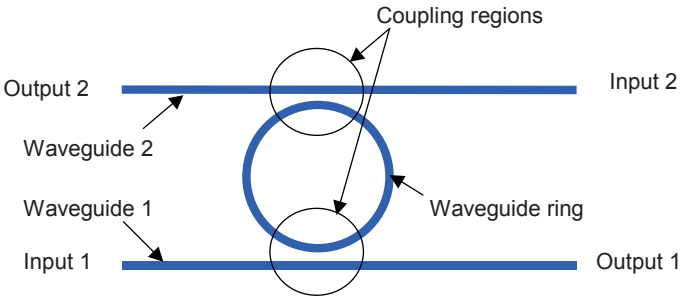


Figure 6.31. Resonant ring filter. The optical field entering the device on Input 1 is coupled (weakly) into the ring resonator. If the circumference of the ring is an integer number of wavelengths, then the field in the ring will build in phase, and the amplitude will build until the coupling from the ring back into waveguide 1 exactly cancels the amplitude in that waveguide. Under these circumstances, all the power on Input 1 is coupled onto Output 2, provided that all the waveguides are loss less.

Before we can derive the response of the ring filter of Fig. 6.31, we must calculate the coupling coefficient of the two directional couplers involved. The index profile is show in Fig. 6.32.

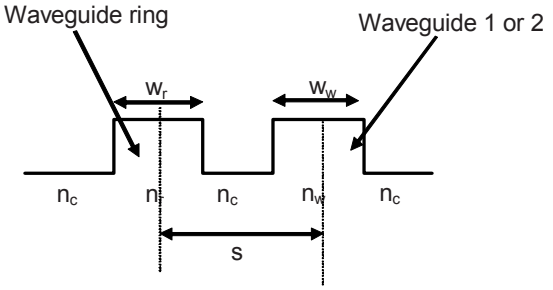


Figure 6.32. Index profile in the guided region of ring resonator. The center-to-center separation, s, between the two waveguides varies because of the curvature of the ring.

Using the same procedure as in our discussion of the directional coupler, we find the following coupling coefficient

$$K(z)=\frac{\omega\cdot\varepsilon_0}{4}\int_{-\infty}^{\infty}u_w(x)u_r^*(x)\left(n_w^2-n_c^2\right)\cdot dx \tag{6.154}$$

which is now z-dependent, because of the curvature of the ring. Recall the coupled mode equations for the directional coupler.

$$\frac{dA_w}{dz}=-jKB_r e^{-j2\delta\cdot z} \tag{6.155}$$

$$\frac{dB_r}{dz}=-jKA_w e^{j2\delta\cdot z} \tag{6.156}$$

where

$$2\delta=\left(\beta_r+M_r\right)-\left(\beta_w+M_w\right) \tag{6.157}$$

Here β_r and β_w are the longitudinal wavevectors of the ring and straight waveguide respectively.

In the ring coupler, we may consider the amplitudes (A and B) constant in the coupling region (correct in the weak-coupling limit). We can then integrate the equations to get

$$\frac{\Delta A}{B}=-j\cdot t=-j\cdot\int_{-\infty}^{\infty}K(z)e^{-j2\delta\cdot z}dz \tag{6.158}$$

and similarly

$$\Delta B=-j\cdot tA \tag{6.159}$$

We will now model the curvature as parabolic in the interaction region. This isn't exact, but a good approximation, if the ring is circular, and it simplified the analysis considerably. The z-dependent ring-waveguide separation can then be expressed

$$s(z)=s_0+\frac{z^2}{R} \tag{6.160}$$

where s_0 is the smallest distance between the ring and the straight waveguide, and R is the radius of the ring (or the geometric mean of the two radii if both waveguides are circular). The coupling coefficient can now be found in closed form [2]

$$t = \frac{\omega \cdot \varepsilon_0 \cdot \cos(\kappa_{xr}\, w_r/2)}{2\sqrt{P_1 P_2}\left(\kappa_{xw}^2 + \alpha_r^2\right)}\left(n_w^2 - n_c^2\right) \cdot \sqrt{\frac{\pi \cdot R}{\alpha_r}} e^{\alpha_r (w_r/2 - s_0)} \cdot \\ \left[\alpha_r \cos(\kappa_{xw}\, w_w/2)\sinh(\alpha_r\, w_w/2) + k_{xw} \sin(\kappa_{xw}\, w_w/2)\cosh(\alpha_r\, w_w/2)\right] \tag{6.161}$$

Here P_i is the normalized mode power, α_i is the decay constant in the cladding, and κ_{xi} is the transverse propagation constant in the core. These parameters can be expressed as

$$P_i = \frac{\beta_i}{2\omega \cdot \mu_0}\left(\frac{w_i}{2} + \frac{1}{\alpha_i}\right) \tag{6.162}$$

$$\alpha_i = \sqrt{\beta_i^2 - n_c^2 k_0^2} \cdot \tag{6.163}$$

$$\kappa_{iw} = \sqrt{n_i^2 k_0^2 - \beta_i^2} \tag{6.164}$$

where k_0 is the free-space wave vector.

Once the coupling between the straight waveguide and the ring is established, we can calculate transfer and reflection characteristics of the ring filter. The treatment is similar to that of the Fabry-Perot interferometer. The circulating amplitude in the ring is given by

$$B_r = B_r \cdot g_{rt} - jt_1 A_{in1} \Rightarrow B_r = \frac{-jt_1}{1 - g_{rt}} A_{in1} \tag{6.165}$$

where t_1 is the coupling coefficient between guide 1 and the ring, and g_{rt} is the roundtrip gain of the ring. It can be expressed as

$$g_{rt} = r_1 r_2 e^{(-\alpha - j\beta_r)2\pi \cdot R} \tag{6.166}$$

where α is the loss in the ring, R is the radius of the ring, and r_1 and r_2 are the attenuation or "reflection" coefficients in the two coupling regions.

The coupling is taking place at a point of the waveguides, so there is no phase shift on the light that is transmitted straight through, only attenuation. The attenuation coefficients are then

$$r_i = \sqrt{1 - (-jt_i)(-jt_i)^*} = \sqrt{1 - t_i t_i^*} \tag{6.167}$$

The transmitted amplitude is

$$A_{out2} = -jt_2 B_r \cdot e^{(-\alpha - j\beta_r)\pi \cdot R} = -jt_2 \cdot \frac{-jt_1}{1-g_{rt}} \cdot e^{(-\alpha - j\beta_r)\pi \cdot R} A_{in1}$$
$$= \frac{-t_1 t_2}{1 - r_1 r_2 e^{(-\alpha - j\beta_r)2\pi \cdot R}} \cdot e^{(-\alpha - j\beta_r)\pi \cdot R} A_{in1} \tag{6.168}$$

The transmittance through the filter is given by

$$T = \left(\frac{A_{out2}}{A_{in1}}\right) \cdot \left(\frac{A_{out2}}{A_{in1}}\right)^*$$
$$= \frac{-t_1 \cdot t_2}{\left\{1 - r_1 r_2 e^{-2\alpha \cdot \pi \cdot R}\left[\cos(\beta_r \cdot 2\pi \cdot R) + j\sin(\beta_r \cdot 2\pi \cdot R)\right]\right\}} \cdot$$
$$\frac{-t_1^* t_2^*}{\left\{1 - r_1 r_2 e^{-2\alpha \cdot \pi \cdot R}\left[\cos(\beta_r \cdot 2\pi \cdot R) - j\sin(\beta_r \cdot 2\pi \cdot R)\right]\right\}} \tag{6.169}$$
$$= \frac{t_1 t_1^* \cdot t_2 t_2^*}{\left\{1 - r_1 r_2 e^{-2\alpha \cdot \pi \cdot R}\cos(\beta_r \cdot 2\pi \cdot R)\right\}^2 + r_1^2 r_2^2 e^{-4\alpha \cdot \pi \cdot R}\sin^2(\beta_r \cdot 2\pi \cdot R)}$$
$$= \frac{t_1 t_1^* \cdot t_2 t_2^*}{1 + r_1^2 r_2^2 e^{-4\alpha \cdot \pi \cdot R} - 2r_1 r_2 e^{-2\alpha \cdot \pi \cdot R}\cos(\beta_r \cdot 2\pi \cdot R)}$$

When the ring losses are negligible, this simplifies to

$$T = \frac{t_1 t_1^* \cdot t_2 t_2^*}{1 + r_1^2 r_2^2 - 2r_1 r_2 \cos(\beta_r \cdot 2\pi \cdot R)}$$
$$= \frac{\left(1 - r_1^2\right) \cdot \left(1 - r_2^2\right)}{1 + r_1^2 r_2^2 - 2r_1 r_2\left[1 - 2\sin^2(\beta_r \cdot \pi \cdot R)\right]} \tag{6.170}$$

In a symmetric ring (r_1=r_2=r) filter this further simplifies to

$$T = \frac{\left(1 - r^2\right)^2}{1 + r^4 - 2r^2\left[1 - 2\sin^2(\beta_r \cdot \pi \cdot R)\right]}$$
$$= \frac{\left(1 - r^2\right)^2}{1 + r^4 - 2r^2 + 4r^2\sin^2(\beta_r \cdot \pi \cdot R)} \tag{6.171}$$
$$= \frac{1}{1 + \frac{4r^2}{\left(1 - r^2\right)^2}\sin^2(\beta_r \cdot \pi \cdot R)}$$

Comparing this to the transmission through a Fabry-Perot filter,

$$T = \frac{1}{1 + \frac{4R}{(1-R)^2}\sin^2\left(\frac{2\pi}{\lambda_0} nL\right)} \tag{6.172}$$

we see that the expressions are identical assuming the following substitution

$$\beta_r \pi \cdot R \to \frac{2\pi}{\lambda_0} nL \tag{6.173}$$

that simply says that the optical circumference of the ring corresponds to twice the optical thickness of the etalon.

Just as for Fabry-Perot interferometers, we usually want the losses of ring filters to be as low as possible. In practical situation, the losses are often dominated by scattering from roughness on the side walls of the ring waveguide. In addition to these kinds of technology dependent losses, there is also a fundamental (i.e. unavoidable) loss mechanism in any curved waveguide.

To see how curved waveguides are fundamentally lossy, consider the basic definition of a waveguide mode

$$\vec{E}(x,y,z) = \vec{E}(x,y,0)e^{-j\beta \cdot z} \tag{6.174}$$

where β is the longitudinal wave vector of the mode, and z is the coordinate along the waveguide axis.. We see that if the waveguide is curved, this equation cannot be fulfilled, because it would mean that some point sufficiently far from the waveguide core, the field must propagate at a velocity higher than the speed of light as illustrated in Fig. 6.33.

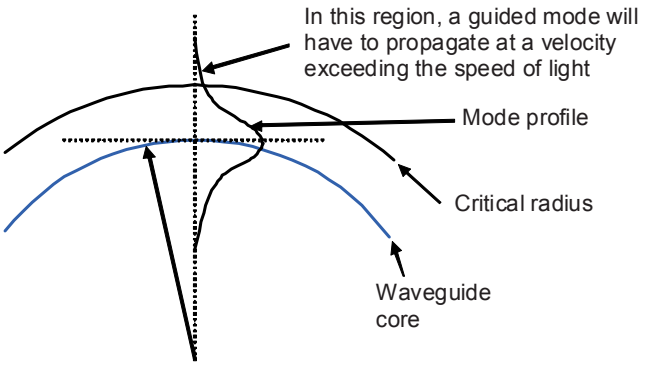


Figure 6.33 There can be no loss-less guided modes on simple curved waveguides, because it would mean that part of the mode have to propagate at a velocity beyond the speed of light.

The curved waveguide can be modeled as a straight waveguide with a linearly varying index, which is increasing on the outside of the curve. (The higher index makes it harder for the field in this region to keep up with the rest of the mode, just like the longer path length makes it harder for the mode on the outside of the curve to keep up.) This is illustrated in Fig. 6.34. Again, we see that a curved waveguide cannot support completely guided modes.

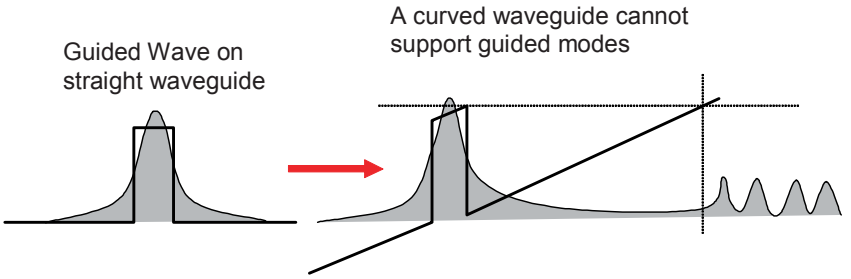


Figure 6.34. Bending loss. Bending of a waveguide can be modeled as a linear variation of the refractive index perpendicular to the waveguide axis. This means that at some distance from the core, the cladding index will exceed the core index, and the mode will not be truly guided.

The losses cased by bending can be found analytically for some simple cases. For rectangular single-mode waveguides, it can be shown that the loss is given by [3]

$$\alpha = \frac{k_3^2 a^2}{8}\left(\frac{n_1^2}{n_3^2}-1\right)^2 \cdot \begin{cases} \left(\frac{n_3}{n_1}\right)^4 \exp\left\{-\frac{k_3^4 a^3 R}{12}\left(\frac{n_3}{n_1}\right)^6\left(\frac{n_1^2}{n_3^2}-1\right)^3\left[1-\frac{1}{2}\right]\left(\frac{k_y}{k_3}\right)^2\right\} & \text{for } E_{11}^x \\ \exp\left\{-\frac{k_3^4 a^3 R}{12}\left(\frac{n_1^2}{n_3^2}-1\right)^3\left[1-\frac{1}{2}\right]\left(\frac{k_y}{k_3}\right)^2\right\} & \text{for } E_{11}^y \end{cases} \tag{6.175}$$

where R is the bend radius, $k_3 = k_0 n_3$ is the wavevector in region 3, k_y is the transversal wavevector of the mode in the y direction, and the other parameters are given in Fig. 6.35.

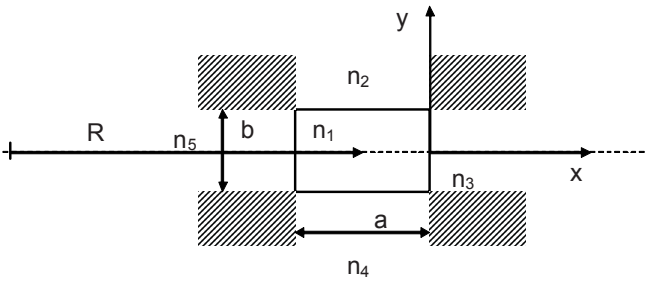


Figure 6.35. Geometry and parameters of bent rectangular waveguide.

The attenuation in curved step-index, single-mode optical fibers is given by

$$\alpha = \frac{1}{2}\left(\frac{\pi}{aV^3}\right)^{\frac{1}{2}}\left[\frac{\kappa \cdot a}{\gamma \cdot aK_1(\gamma \cdot a)}\right]^2 R^{-\frac{1}{2}} e^{-UR} \tag{6.176}$$

where $\kappa\left(\kappa^2 = k_0^2 n_{core}^2 - \beta^2\right)$ is the transversal wavevector, a is the core radius, K_1 is the modified Bessel function of the second kind of order 1, V is the V-number of the fiber, and R is the radius of curvature. The parameters γ and U are given by

$$\gamma^2 = \beta^2 - k_0^2 n_{clad}^2 \tag{6.177}$$

$$U = \frac{4\Delta(\gamma \cdot a)^3}{3aV^2 n_{clad}} \tag{6.178}$$

Together with the formulas we have found earlier, this expression for the loss allows us to calculate the transmittance and reflectance from a ring filter.

As is the case with F-P filters, we can cascade ring filters to achieve faster roll-off and flatter pass-band. Combinations of two or more rings also give us the flexibility to design filters with wider Free-Spectral-Ranges (FSR), and favorable waveguide orientation (the power isn't necessarily propagating back towards the source as in Fig. 6.31.)

6.8 Summary of Fiber and Waveguide Devices

In this Chapter we study fiber-optic devices that are important in Optical MEMS and Nanophotonic. The first class of devices that is described simply provides a means for coupling to optical fibers and waveguides. We use Gaussian Beam theory to derive closed-form analytical expressions to calculate the effects of mis-

alignment on fiber-to-device coupling efficiency, and we describe a set of practical implementations that are well suited for miniaturized optical devices.

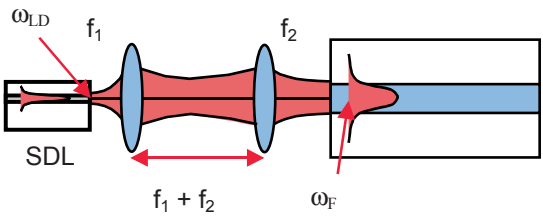
We then develop a perturbation theory that simplified modeling of devices with interacting modes. This coupled-mode theory and the adjunct eigen-mode model are applied to two very basic fiber-optic devices; the directional coupler and the Bragg reflector (which we also studied in Chapter 3). These two structures hold a special place in optical communications for several reasons: (1) They are used in their elementary form in many optical systems, (2) they form the basis for a large number of advanced optical devices, and (3) their descriptions build intuitive understanding of central optics concepts, including photonic bandgap structures.

The last part of the chapter is focused on optical modulators, starting with a general description of waveguide modulators and the physical effects they utilize to create optical signals. Finally we describe the modulator designs that are best suited for miniaturization. These designs include acousto-optic modulators, Mach-Zender modulators, directional-coupler modulators, Fabry-Perot modulators, and coupled-resonator modulators.

Exercises

Problem 6.1 - Coupling to Single-Mode Fibers

We are designing a fiber for single mode operation at *1.55 μm* wavelength. The core index is *1.45*, the cladding index is *1.446*, and the core radius is chosen such that the mode radius is *5 um* at *1.55 um* wavelength. We want to couple light from a semiconductor laser diode (SDL) at *1.55 μm* wavelength into the fiber using the set-up shown below. The SDL has a circular fundamental Gaussian mode with a *1.0 μm* mode radius. The focal length of the first lens (f_1) is *5 mm*.



Coupling from a Semiconductor Laser Diode (SDL) to a single mode fiber.

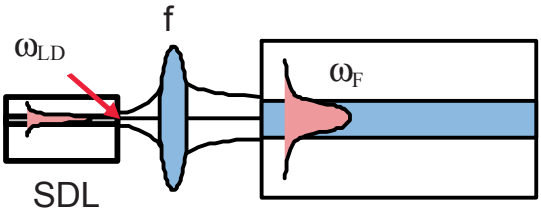
a. What should the focal length of the second lens (f_2) be to optimize the coupling?

In addition to the fundamental circular Gaussian mode with the field distribution $E_1 \propto E_0 \cdot e^{-\left(x^2+y^2\right)/\omega_{LD}^2}$, the laser also supports a second mode with a field distribution given by $E_2 \propto E_0 \cdot 2x \cdot e^{-\left(x^2+y^2\right)/\omega_{LD}^2}$.

b. In the case of perfect coupling of the first mode to the fiber, how much of the second mode will be coupled? (Explain).
c. How can you modify the coupling set-up such that you can switch between coupling all of laser mode number 1, all (or at least most) of laser mode number 2, or a combination of the two? (Explain qualitatively). This could be useful if the two modes are at slightly different wavelengths, but we will not consider wavelength differences of the two modes here.

Problem 6.2 –Coupling to Altered Fibers

We use a single lens to couple light from a semiconductor laser diode (SDL) at *1.55 um* wavelength into a single-mode fiber as shown below. The SDL has a circular Gaussian mode with a *1.0 um* mode radius, and the fiber has a mode radius of *5 μm* at *1.55 μm wavelength*. The core index of the fiber is n_{core}=*1.45*, and we will use this index to calculate reflections from the fiber.

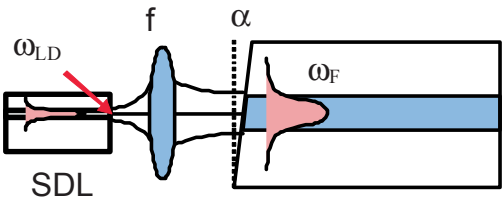


Coupling from a Semiconductor Laser Diode (SDL) to a single-mode fiber.

a. In the case of perfect alignment and mode matching, what fraction of the laser power gets coupled into the fiber, and what fraction gets reflected back into the laser? (Assume no reflections from the lens.)

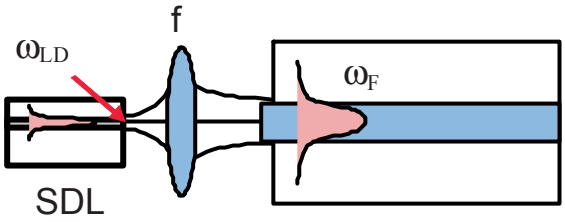
We have the same coupling set up as in a), only the fiber is angle polished as shown below. The angle of the facet with respect to the optical axis is: α=*5 degrees*. Except for the angled fiber facet, the alignment and mode matching is perfect (i.e. the alignment and mode matching would have been perfect if the fiber facet was perpendicular to the optical axis).

b. Under these circumstances, what fraction of the laser power gets coupled into the fiber? (Assume no reflections from the lens.)



Coupling to angle-polished single-mode fiber.

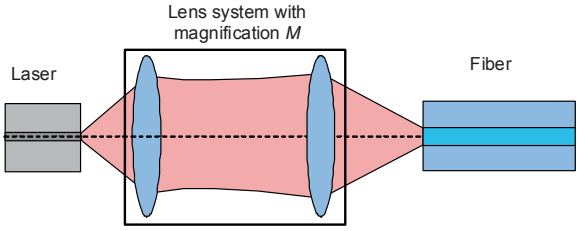
c. Explain qualitatively how the coupling would change if the fiber core was protruding as shown below. (This could happen if the fiber was etched in Hydrofluoric acid, which etches pure SiO_2 faster than the Germanium-doped SiO_2 of the core).



Coupling to single-mode fiber with protruding core.

Problem 6.3 – Coupling to Misaligned Fiber

A lens system with a magnification M, is used to couple light from a laser with a circularly Gaussian mode of radius $\omega_l=1\mu m$ to a single mode fiber with a mode radius $\omega_f=5\mu m$. Assume that the effective index of the fiber mode is *1.5*.



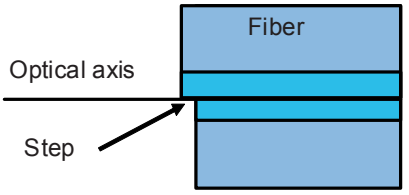
Optical system for coupling a Gaussian beam from a laser into a fiber.

a. What is the maximum coupling that can be achieved, and what is the corresponding magnification?

b. What is the maximum coupling and corresponding magnification if the fiber is offset by *1 um* from the optical axis of the optical system? (Assume that the laser is perfectly aligned with the optical system).

The fiber is damaged such that its facet becomes stepped as shown below. Except for the step, the facet is perfect (i.e. flat and perpendicular to the optical axis). The step divides the fiber facet exactly in half (i.e. the step is at the optical axis), and the step is exactly one wavelength.

c. Explain how this changes the coupling from the laser to the fiber. (Assume that the optical system is perfectly aligned, i.e. all components are exactly on the optical axis.)

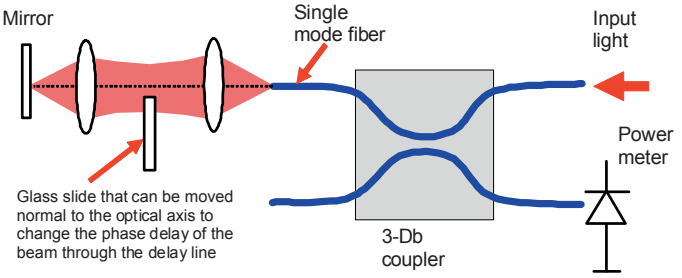


Damaged fiber facet.

Problem 6.4 - Wavefront of Fiber Modes

Consider the experiment shown in the figure below. The light is reflected from a mirror at the position of the output fiber in the original experiment. The amount of light that is coupled back into the fiber is measured using a 3-dB coupler and a power meter. Assume that the glass plate is perpendicular to the optical axis.

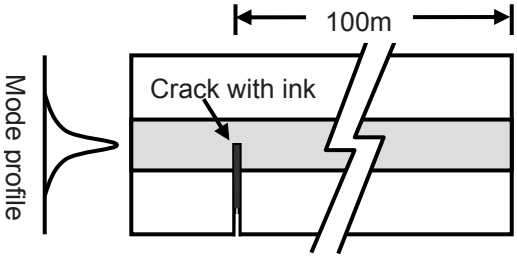
What is the position and minimum thickness of the glass slide that result in zero back coupled light? (Express your answer in terms of x/ω, where ω is the size of the Gaussian beam radius at the waist where the glass slide is placed.)



Modified experimental setup for studying the phase front of optical fiber modes. The optical beam is reflected back into the fiber, and the back coupled light is measured using a 3-dB coupler and a power meter.

Problem 6.5 – Cracked Fiber

We have a single mode fiber with a mode profile as shown. The fiber has a narrow crack that is filled with absorbing ink. The crack covers exactly half the fiber cross section, and even though it is only a few wavelengths wide, all the light in this half of the fiber is absorbed. At the end of the fiber, *100 m* from the crack, the output power is approximately *25%* of the input power.

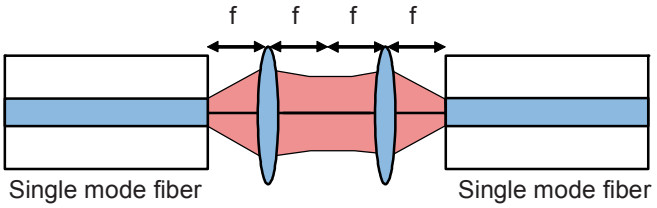


Single mode optical fiber with crack.

a. Sketch the mode profile at the output of the fiber 100m from the crack. (Only the shape is important. We know that the power is 25% of the input power)
b. We rinse the fiber and replace the ink with water that does not absorb significant amounts of light over the narrow width of the crack. The rinsing is done carefully so that the physical dimensions of the crack do not change. Now we observe that optical output power is much less than before, i.e. much less than 25% of the input power. Explain how this can happen.

Problem 6.6 – Fiber-to-Fiber Coupling

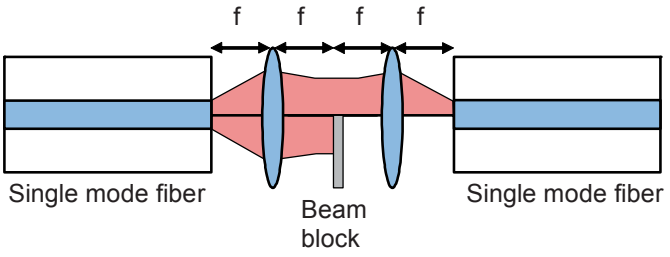
Two single mode fibers are connected through a lens system as shown below. The light propagates from left to right. The fibers are identical with modes that can be approximated as Gaussians with beam radii of *5um* at *1.55um* wavelength. The focal lengths of the lenses are both *5mm*. Assume that the alignment of the system is perfect and that there are no reflections, so that *100 %* of the light is coupled from the fiber on the left to the one on the right.



Coupling between single-mode fibers.

a. What is the Gaussian beam radius at the center point between the two lenses?

A beam block is inserted into the lens system exactly half way between the lenses as shown below. The block intercepts exactly half of the optical beam.



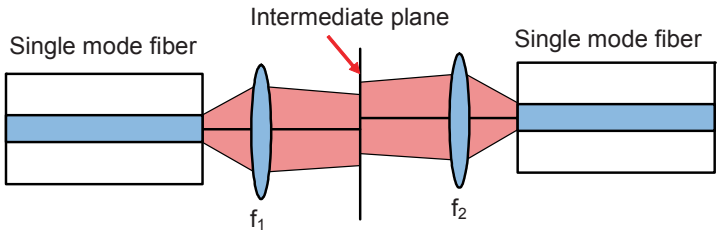
Partially blocked coupling between single-mode fibers.

b) Is the amount of light coupled between fibers more than 50%, about 50% or less than 50%? (Explain your answer)

Problem 6.7 – Misaligned Fiber-to-Fiber Coupling

Two single mode fibers are connected through a lens system as shown below. We know nothing about the fibers except that they are single mode. All we know about the lens system is that at an intermediate plane between the two fiber ends, both fiber outputs have a waist (focus) with the same Gaussian beam radius, ω. The two beams are both perpendicular to the intermediate plane, and their beam centers are separated by 0.1ω.

What is the coupling loss from one fiber to the other? Explain.

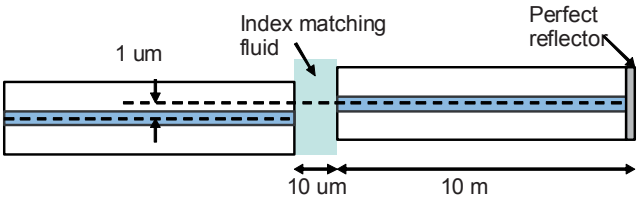


Coupling between single-mode fibers.

Problem 6.8 – Direct Fiber-to-Fiber Coupling

Two identical step-index fibers are offset axially and laterally as shown below. The fiber on the right is terminated in a perfect reflector, i.e. light that is propagat-

ing to the right on this fiber will be 100% reflected at the end. The gap between the fibers is filled with index matching fluid, so there are no reflections or scattering from the gap. Fiber parameters: Core radius: $a=4\mu m$, core index: $n_{core}=1.45$, cladding index: $n_{clad}=1.446$, wavelength: $\lambda=1.55\mu m$.



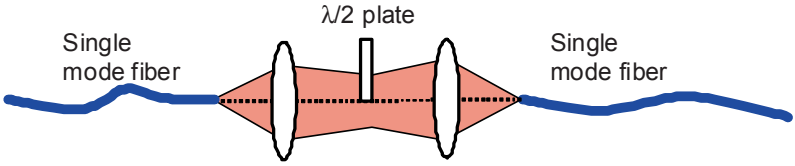
Identical fibers that are offset both axially and laterally.

a. Are these fibers single-mode fibers?
b. What is the transmission loss going from the leftmost to the rightmost fiber?
c. What is the total reflection for light that is traveling to the right on the leftmost fiber, i.e. how much of the light will couple from the left fiber to the one on the right, be reflected from the end, and then couple back into the left fiber again?
d. How does the transmission loss from one fiber to the other change as the core radius, a, is decreased? (Explain qualitatively)

Problem 6.9 - Polarization of Fiber Modes

What is the coupling from the fiber on the right to the fiber on the left in the set-up shown below?

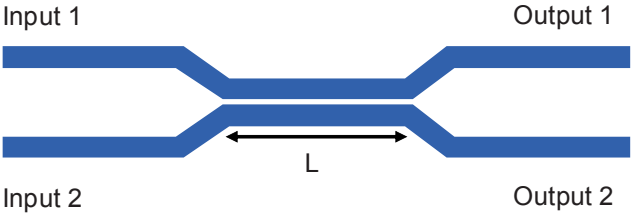
Assume that the light laving the fiber on the right is linearly polarized aligned to the $\lambda/2$ plate such that the polarization is orthogonal after passing though the plate. The plate covers exactly half the beam.



Experimental setup for studying the polarization sensitivity of optical fibers.

Problem 6.10 – Directional Coupler

Consider the symmetric directional coupler shown schematically below. At *1.55 μm* wavelength, we have that the coupling coefficient and the effective index in the coupling region are $K=10\ cm^{-1}$, and $n_{eff}=1.5$, respectively.



Directional Coupler.

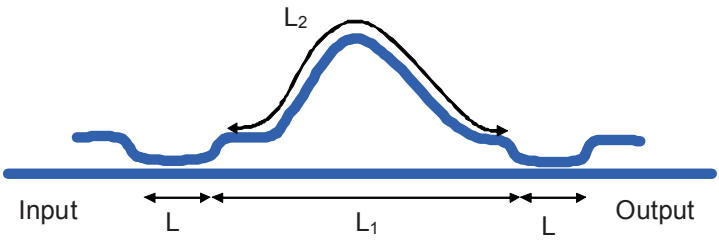
The length, *L*, of the interaction region is chosen such that the directional coupler acts as a *10%* power tap, (i.e. if there is non-zero input only on *Input 1*, then *10%* of the power goes to *Output 2* and *90%* goes to *Output 1*), at *1.55 um* wavelength.

a. What is the shortest coupler length, L, required to make a 10% power tap at 1.55 μm wavelength?
b. For the power tap described in a), what is the distribution of output powers if the only non-zero input is on *Input 2* (*1.55 um* wavelength)?

Problem 6.11 – Fiber Interferometer

Consider the waveguide device consisting of two directional couplers shown in the figure below. The lengths of the couplers (L) are the shortest length that evenly splits the power at *1.55 um* wavelength, and they are much shorter than the length between the couplers (L_1 for the straight waveguide, and L_2 for the curved waveguide). Assume that the directional couplers split a single input evenly between the outputs over the whole wavelength range of interest.

a. Write an expression for the power transmission through the device as a function of wavelength.



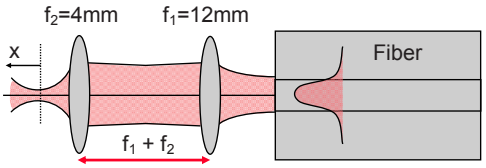
Fiber interferometer.

Consider what happens when this coupler is excited with two different wavelengths. *Input 1* has power, *P*, at exactly *1.55 um* wavelength. *Input 2* has the same power, but the frequency is shifted up by *10 MHz* compared to *Input 1*.

b. Find an expression for the power on *Output 1* as a function of time.

Problem 6.12 - Resolution of Confocal Microscopes

a. Use the formulas on coupling of Gaussian Beams into single mode fiber to calculate the axial resolution of the confocal microscope shown in the figure below (calculate the amount of light that would be coupled back into the fiber from a mirror at a distance *x* from the focus of the beam, and define the axial resolution as the Full-Width-at-Half-maximum (FWHM) of the resulting function of *x*). Express your answer in terms of the Gausain beam mode radius of the fiber and the wavelength.



Fiber Optic Confocal Microscope.

b. Rewrite your answer in terms of the NA of the objective lens (lens 2) and compare to the standard formula for the axial ($\Delta d_z = 0.90\lambda/NA^2$) resolution of a confocal microscope. (Don't expect perfect correspondence.)
c. Verify the standard formula for the transversal resolution ($\Delta d_{x,y} = 0.37\lambda/NA$) of a single-axis confocal microscope.

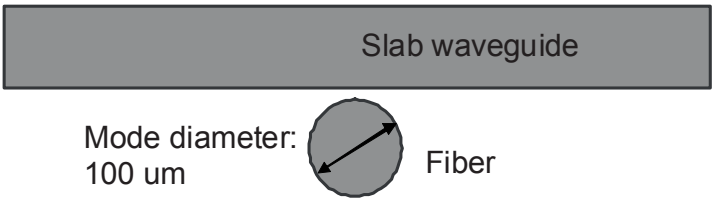
Problem 6.13 – Design for Packaging

Packaging of fiber modulators and other waveguide devices usually includes attachment of fiber pigtails, and this is often the most costly part of the manufacturing process. It therefore make sense to design waveguide devices and fibers such that the required accuracy of the pigtail attachment is minimized. Assume that you have a packaging technology, in which the standard deviation of the lateral placement of the fibers is 1 μm, and the angular standard deviation is $4 \cdot 10^{-3}$.

What is the optimum fiber-mode size for telecommunication devices (1.55 um wavelength) that are packaged using this technology? Assume spherically-symmetric Gaussian modes for waveguides and fibers.

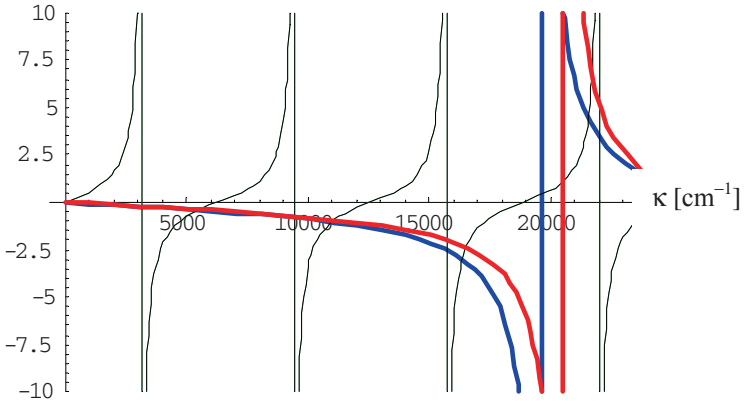
Problem 6.14 – Evanescent Coupling

A directional coupler is made by bringing the core of a single-mode fiber with a mode diameter of *100 um* in proximity to a slab waveguide as shown below. The effective index of the fiber mode (in the presence of the slab guide) is *1.4826*.



Cross-section of directional coupler consisting of a single-mode fiber and a slab waveguide. .

The graphical solutions for the transverse wave vectors of the modes in the slab waveguide in the presence of the fiber are shown in the graphs below. (Use this graph to find values for the transverse wave vectors to use in your calculations.)



Graphical solutions for the TE (left) and TM (right) guided waves. The core index of the slab waveguide is n_f=1.5.

a. The fiber mode is exactly mode matched to one *TE* mode of the slab waveguide at *1.3 um* wavelength. Find which *TE* mode in the slab that is phasematched, and sketch (no detailed calculations) its field profile.

A coupler of the type described in a) has a coupling coefficient $K=100\ cm^{-1}$ and a length $L=5/K$. At the input to the coupler there is no power in the slab waveguide, and the fiber carries a power *P*. Assume that the polarization of the field in the fiber corresponds to the *TE* mode on the slab.

b. What is the power in the fiber at the output? (Hint: Check to see if the effects of diffraction in the slab are significant.)
c. Explain qualitatively how this will change if the coupling coefficient is reduced to $1.0\ m^{-1}$, and the length is increased such that we still have $L=5/K$?

Problem 6.15 – Fiber Polarizer

How can you combine what you have learned about directional couplers and surface plasmons to design a fiber polarizer?

Problem 6.16 – Eigenmodes of Bragg Filters

Find the eigenmodes of the Bragg filter based on the coupled-mode equations we derived in chapter 6.6.

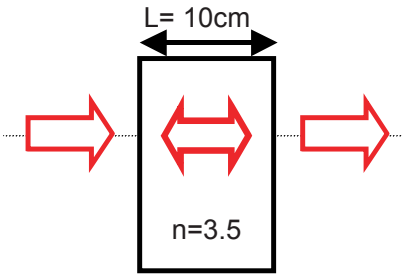
Problem 6.17 - TE to TM Coupling

We have a waveguide with the following parameters: Thickness=*5 um*, n_c=*1.4*, n_f=*1.5*, n_s=*1.45*, wavelength=*1 um*. The characteristic equations for the TE (blue) and TM (red) modes are shown graphically in the figure in the preceding problem.

What is the Bragg-grating period required to couple the TE_3 and TM_3 modes in this waveguide?

Problem 6.18 – Fabry-Perot

A loss less, dielectric slab with an index of *3.5* as shown below is used as a Fabry-Perot interferometer.



a. Calculate the Free-Spectral-Range (axial mode spacing) and bandwidth of the interferometer?
b. How does the Free-Spectral-Range (axial mode spacing) and bandwidth of the interferometer change if the surrounding medium is replaced by one of index n>1? (Explain)

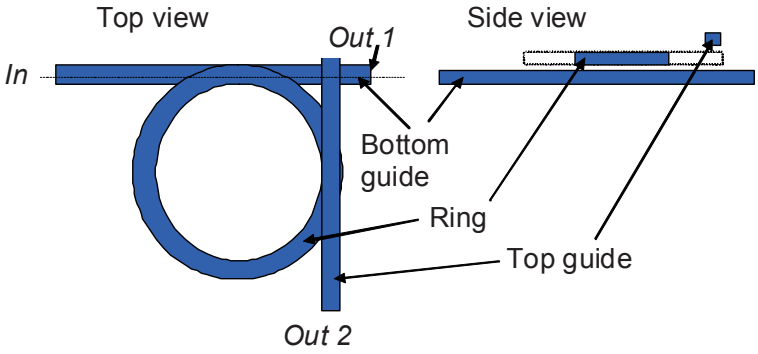
Problem 6.19 – Ring Filter

(This is a design problem, so some parts might be under specified. In those cases you should make reasonable assumptions.)

A Wavelength-Division-Multiplexed (WDM) fiber optical communication system uses 32 wavelength channels, centered at *1.55 um* wavelength and separated by *100 GHz*. Each channel has a signal bandwidth of *10 GHz*.

a. Design a resonant directional coupler like the one in the figure below. The coupler should couple ***every other*** (i.e. either the even or odd channels) wavelength channel on port *In* to port *Out 2*. The non-coupled channels should appaer on *Out 1* and be suppressed by *40 dB* on *Out 2*. The design involves choosing the length of the ring and the coupling coefficient between the straight waveguides and the ring.

Assume that the coupling sections are short (i.e. the power in either the ring or the waveguide does not change appreciably through coupling region), that the coupling coefficient is wavelength independent, that the effective index of the waveguide and ring modes is *1.5*, and that the ring has negligible bending loss.



Resonant-ring cross point.

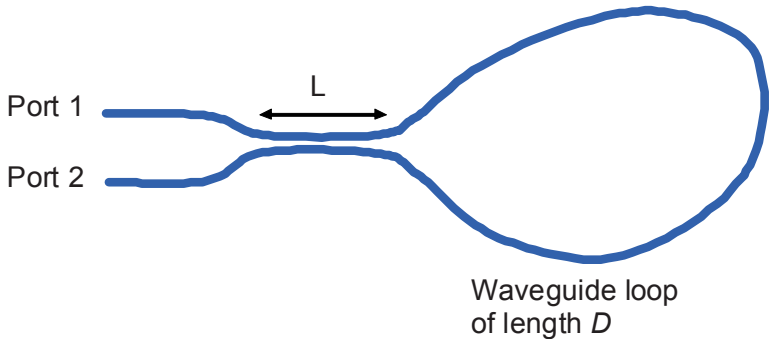
Clearly you can switch all the wavelength channels between the two guides if you use a longer ring resonator.

b. How long should the ring be and what should the coupling coefficient be?
c. How much would you have to change the index of the ring to turn the cross-point from its *ON* state (all channels coupled to *Out 2*) to its *OFF* state (all channels coupled to *Out 1*)?
d. Discuss the advantages and disadvantages of this switch design compared to the the champagne switch anf the MEMS switches of Chapter 8.
e. Can you think of a better implementation of this type of switch than the one shown in the above figure?

Problem 6.20 – Fiber-Loop Mirror

A directional coupler that splits power at wavelength λ_0 on one input evenly between the two outputs is used in the configuration shown in below. We define the reflectance of the structure as the ratio of the power out of *Port 1* to the power into *Port 1*, with no power into, but possibly out of, *Port 2*. Assume that the only significant wavelength dependence of the coupling coefficient is the one that is explicit in the formula.

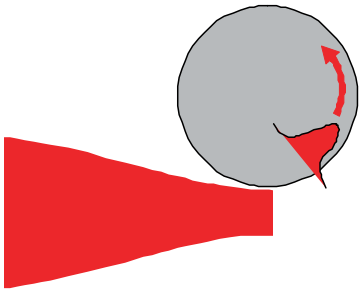
Find an expression for the reflectance as a function of wavelength, coupling coefficient at λ_0, and loop length, D.



Waveguide loop mirror.

Problem 6.21 – Evanescent Coupling to Ring Resonator

Consider the following question: Is it possible to use a perfect ring (or ball) oscillator to couple power out of a laser beam propagating in free space as shown in below?



Will the power in the laser be coupled to the high-Q optical oscillator as shown, or is this a fallacy?

References:

1 D. Marcuse, “Loss Analysis of Single-Mode Fiber Splices”, The Bell System Technical Journal, vol. 56, no. 5, May-June 1977, pp. 703-718.
2 B.E. Little S.T. Chu, H.A. Haus, J.-P. Laine, “Microring Resonator Channel Dropping Filters”, Journal of Lightwave Technology, vol. 15, no. 6, June 1997, pp. 998-1005.
3 E.A.J. Marcatili, “Bends in Optical Waveguides”, Bell Systems Technical Journal, vol. 48, September 1969, pp. 2103-2132.

7: Optical MEMS Scanners

7.1 Introduction to MEMS Scanners

Using mirrors to deflect and position optical beams is a trick that is as old as it is important. Light houses might have been the first "killer app" (or more appropriately "savior app" for someone growing up on the treacherous coast of Norway). With the invention and development of laser, this age-old technology has become ubiquitous! Optical scanners are the enabling components in systems covering an astonishing range of applications, including such important areas as imaging, microscopy, communications, printing, displays, retail, light shows, fiber switches, security, remote sensing, metrology, surveillance, laser machining, and laser surgery.

This variety of uses has of course led to a similarly large variety of implementations. We will limit ourselves to systems using miniaturized scanners based on MEMS technology, but even with this restriction, the field is way too large for a comprehensive description of the different technologies that are in use. Instead we will focus on the fundamental characteristics of optical scanners, and, since we are interested in miniaturized systems, on their scaling.

One consequence of our focus on miniaturization is that we will exclusively consider spatially coherent light, which allows the scanning optics to be significantly smaller than the systems needed to scan spatially incoherent light. Spatial coherence will often, but not always, mean light from single-spatial-mode lasers. The important exception for Optical MEMS is large arrays of microscanner that can be illuminated by a collimated beam from a traditional light sources of small area such that the angular spread of the illumination is small compared to the diffraction angle from each microscanner in the array. Each microscanner is then effectively illuminated by spatially coherent light, even though the light source itself is spatially incoherent[a]. This is the typical illumination scheme for TI's DLP tech-

[a] In this case, spatial coherence is established by spatial filtering the light from the traditional light source. In most cases this is an inefficient process involving sending the light through a pin hole, but the magnification of the beam, which

nology. Our focus on spatially-coherent light simplifies our treatment, because it allows us to use the Gaussian beam theory developed in Chapter 4 to model the performance of scanners.

In the first part of the chapter, we describe the resolution of optical scanners. The resolution can be quantified as the number of pixels, or number of resolvable spots, that the scanner can support. The number of resolvable spots is a fundamental property of a scanner [1]. The optical system can reduce the number of resolvable spot by introducing loss, but no linear, lossless optical system can increase the number of resolvable spots established by the scanner. This insight is very useful when designing scanning systems. By casting the application requirements in terms of a number of resolvable spots, the scanner can be specified, and then the optical system can be designed to fit the scanner. Some iteration might be necessary, but nevertheless, starting with the number-of-resolvable-spots greatly simplifies the design process.

In the second part of the Chapter, we consider effects that can limit or reduce scanner resolution. These include mirror aperture, surface roughness, and static and dynamic mirror curvature. MEMS mirrors are typically coated with a thin metal film to enhance reflectivity. We use the formulae for Fresnel reflections, derived in Chapter 3, to clarify material choices and film thicknesses needed to achieve good mirror performance.

The focus of this book is optical design, but mechanical design is so important for scanners that we devote a section to highlight the most significant issues. The mechanical design is due to the high frequency operation, the low available forces, and the under-damped characteristics typically encountered in most MEMS designs. It is further complicated by the desire to keep fabrication simple and compatible with MEMS parallel processing. We discuss these issues and how they influence the implementations of single-axis and dual-axes scanners.

The last section of the Chapter is devoted to several examples of successful MEMS scanner designs. The examples are chosen to illustrate a range of mechanical designs, including gimbals and universal joints used to implement high-resolution, 2-axes scanners. A wide variety of actuators have been used to implement MEMS scanners, but our focus is on electrostatic actuation due to its prevalence and its material compatibility and relative simplicity of integration with IC manufacturing processes. An important adjunct to this Chapter is therefore Appendix B on Electrostatic actuators.

---

reduces its angular content, combined with the small size of the microscanner, allows the spatial filtering to be efficient.

7.2 Scanner Resolution

The energy-conservation arguments of Chapter 2 tell us that the overlap between two optical fields does not change as the fields propagate through lossless, linear, passive optical systems. For scanners this means that the number of points that can be resolved is an inherent property of the scanner, and cannot be improved by clever system design. This is illustrated in Fig. 7.1 that shows a schematic of a simple scanning system. The lens systems that are used to relay the optical field from the laser to the scanning mirror and then to project the scanned optical beam onto the screen do not change the resolution that is established by the scanning mirror itself. It is of course possible to reduce the resolution by introducing loss, but that is rarely a useful practice. This insight simplifies the design process, because the system requirements on resolution or resolvable pixels translate directly into specifications for the scanning element.

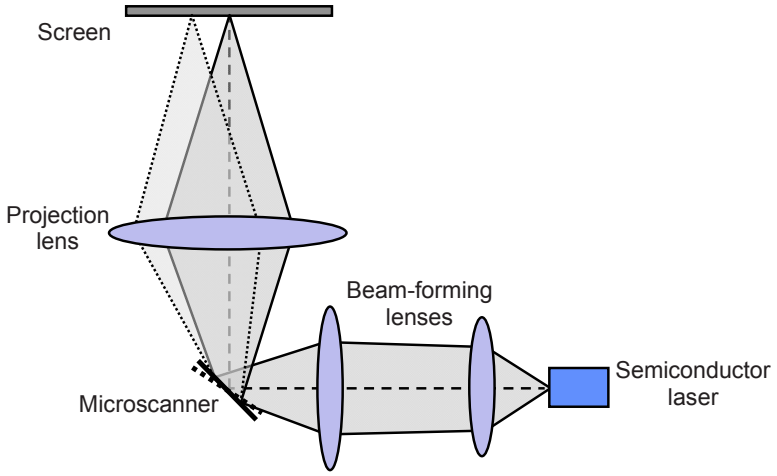


Figure 7.1. Schematic of microoptical scanning system showing the scanner in two positions (solid and dashed lines). The overlap between the scanned beams created by the two settings of the scanner is constant.

When using energy-conservation arguments we must keep in mind that the invariant overlap integrals are taken over all dimensions of the optical field, while in scanning systems we are primarily concerned with spatial separation of intensity distributions. As the two fields corresponding to the two different setting of the scanning mirror propagate through the optical projection system their overlap integrals will be constant, but their distinguishing features will vary, i.e. in some parts of the system the fields will overlap spatially, but be separated angularly, while in other parts it will be the other way around. Almost all scanning systems are designed so that at the output, the fields are only distinguishable in one charac-

teristic, typically spatial location, so that their difference in this dimension is maximized.

The design of a scanning system can therefore be broken down into two parts: First the scanner itself is designed to meet the requirements of the application, and then the lens system is designed to project the scanned output with the correct separation and magnification. The lens systems are as varied as the applications they are designed for. Such systems are the subject of numerous text books, and several sophisticated computer programs for lens design and lens-system design are available. Here we will concentrate on the resolution and implementation of the microscanners themselves.

7.2.1 Resolution of an Ideal Scanner

To design a scanner to a set of specifications, we need to understand the relationship between the physical characteristics of the scanner and its resolution. We will start by considering an optical beam reflecting from an ideal flat, infinite mirror as shown in Fig. 7.2. For simplicity we assume that the beam waist is on the scanning mirror, and that the screen is in the far field, i.e. the screen is sufficiently far away from the scanner that the beam radius increases linearly with distance. In other word, we are sufficiently far away from the scanner that the beam has a well-defined diffraction angle.

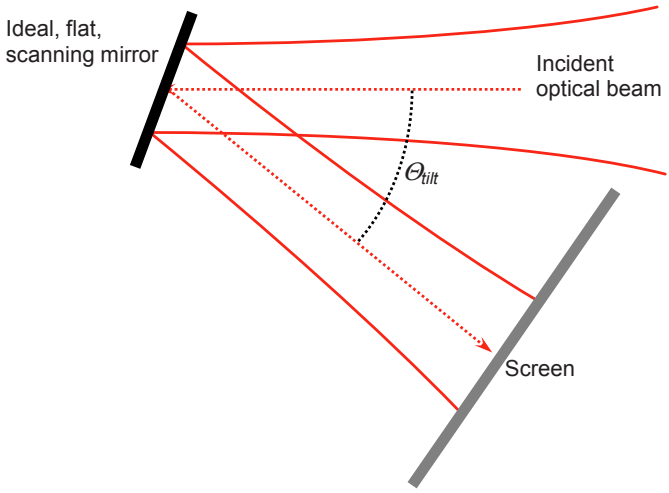


Figure 7.2. The resolution of the scanning mirror is determined by the total range of scanning angle and by the diffraction of the beam coming off the mirror.

The number of resolvable spots (pixels) that the mirror can form on the screen in a one-dimensional scan is then

$$N = \frac{\Delta\theta_{tilt}}{\theta_{diff}} + 1 \tag{7.1}$$

where $\Delta\theta_{tilt}$ is the maximum change in tilt angle that the scanning mirror can achieve and θ_{diff} is the diffraction angle of the beam. The tilt angle is an inherent property of the scanner, while the diffraction angle depend on the beam size on the scanner and the wavelength. In Chapter 4 we gave the half-angle of a Gaussian beam as

$$\theta = \lim_{z \to \infty} \frac{\omega(z)}{z} = \frac{\lambda}{\pi \cdot \omega_0} \tag{7.2}$$

where $\omega(z)$ is the beam radius, and ω_0 is the beam radius at the waist.

This particular definition is arbitrary in that we have simply chosen the 1/e radius ($1/e^2$ for intensity) as a convenient measure for the beam size. In any real scanning system we should chose the beam size that is appropriate for that application. Let's consider scanning displays as an example. Viewer-perception studies of quality of Cathode-Ray-Tube (CRT) images have shown that the optimum pixel separation is the full-width-at-half-maximum (FWHM) of the scanned beam on the screen [2].

This result is applicable to Gaussian beam scanning, because the electron beam in a CRT creates an approximately Gaussian light distribution on the screen. Given this resolution criterion, the diffraction angle (FWHM) becomes

$$\begin{aligned} & e^{-\frac{r_{FWHM}^2}{2\omega^2}} = \frac{1}{2} \Rightarrow r_{FWHM} = \omega\sqrt{0.5\ln 2} \approx 0.589 \cdot \omega \Rightarrow \\ & \theta_{diff} = \lim_{z \to \infty} \frac{2r_{FWHM}}{z} = \lim_{z \to \infty} \frac{1.18 \cdot \omega(z)}{z} = 1.18\frac{\lambda}{\pi \cdot \omega_0} \end{aligned} \tag{7.3}$$

For scanning display we can then express the number of resolvable spots as

$$N = \frac{\Delta\theta_{tilt}}{\theta_{diff}} + 1 \approx \Delta\theta_{tilt} \cdot \frac{\pi}{1.18} \cdot \frac{\omega_0}{\lambda} + 1 \approx \Delta\theta_{tilt} \cdot 2.67 \cdot \frac{\omega_0}{\lambda} + 1 \tag{7.4}$$

If we assume that the scanning system can support angles up to 0.75 radians (limited either by the scanner itself, or by the supporting lens system), we find the following expression for the number of resolvable spots

$$N \approx 2.0 \cdot \frac{\omega_0}{\lambda} + 1 \tag{7.5}$$

In the visible (λ=500nm), we see that a beam radius of 250 micron is sufficient for HDTV resolution! A micromirror of a diameter of about 750 micron can support this size beam. This demonstrates the basic property that makes scanning micromirrors so useful in so many applications; it doesn't take a very large mirror to resolve a large number of spots.

The simple equation above does not tell the whole story. We used a specific resolution criterion (pixels separated by their FWHM) and assumed a rather large scan angle of 0.75 radians. For most systems we find that neither of these two assumptions is valid. For general calculations of scanning Gaussian beams, we therefore us the expression:

$$N = \frac{\Delta\theta_{tilt}}{\theta_{diff}} + 1 = \frac{\Delta\theta_{tilt}}{k \cdot \lambda / \pi \cdot \omega_0} + 1 = \frac{\Delta\theta_{tilt} \cdot \pi \cdot \omega_0}{k \cdot \lambda} + 1 \tag{7.6}$$

where k is a constant that is set by the relevant resolution criterion for the application in question. For a scanning display we have $k=1.18$ as we have seen. If we use the full diffraction angle out to the $1/e^2$ intensity of the Gaussian beam, then $k=2$. Fiber-optic switches based on scanning mirrors require substantially larger separation to achieve acceptable cross talk, so typically we use k-values on the order of 3 or more.

In the next section we will see that the equation we have found for the number of resolvable spots is valid for a much wider range of situations than covered by Fig. 7.2 where the scanned optical beam has its waist on the scanner.

7.2.2 Optimum Resolution of a Scanned Gaussian Beam

In the discussion of resolution criteria and resolvable pixels in Chapter 7.1, we assumed that the beam waist was on the scanning mirror. To get a better understanding of scanning systems in general, we relax that assumption, and consider the case shown in Fig. 7.3. Here we have a converging Gaussian beam that is reflected from a scanning mirror onto a projection screen positioned a distance, d, away from the scanner. The Gaussian beam converges to a waist somewhere after the scanning mirror, and then starts diverging towards the screen. This is a generalized description that incorporates the situation of having the beam waist on the scanner as a special case ($z_0=0$). The way we set up the problem also allow us to give z_0 negative values, which means that beams that are diverging at the scanning mirror are also covered by the formalism.

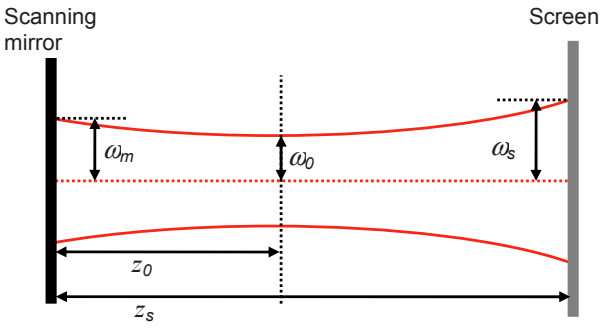


Figure 7.3. Profile of a Gaussian beam that is converging on a scanning mirror and reflected onto a display screen.

In the general case of Fig. 7.3, we rewrite the equation for resolvable spots in terms of the propagation distance from the scanner to the screen and the beam radius on the screen:

$$N = \frac{\Delta\theta_{tilt} \cdot z_s}{k \cdot \omega(z_s)} + 1 \tag{7.7}$$

where the beam radius is given by

$$\omega(z_s) = \omega_0 \sqrt{1 + \left(\frac{\lambda \cdot (z_s - z_0)}{\pi \cdot \omega_0^2} \right)^2} \tag{7.8}$$

We now maximize the number of resolvable spots under variations of the scanner-screen distance

$$\frac{dN}{dz_s} = 0 \Rightarrow z_s = \frac{\pi^2 \cdot \omega_0^4}{z_0 \lambda^2} + z_0 \tag{7.9}$$

Plugging this back into Eq. 7.8 yields the following expressions for the beam waist at the screen, the beam waist at the mirror, and finally the number of resolvable spots.

$$\omega(z_s) = \omega_0 \sqrt{1 + \left(\frac{\pi \cdot \omega_0^2}{z_0 \lambda} \right)^2} \tag{7.10}$$

$$\omega(-z_0) = \omega_0 \sqrt{1 + \left(\frac{z_0 \lambda}{\pi \cdot \omega_0^2} \right)^2} \tag{7.11}$$

$$N = \frac{\Delta\theta_{tilt} \cdot \pi \cdot \omega(z_0)}{k \cdot \lambda} + 1 \tag{7.12}$$

This last expression is the same as the one we found for the resolvable pixels of a scanner with the beam waist on the mirror! The case of a beam focused on the scanner is recreated by setting $z_0=0$, resulting in $\omega(z_0)=\omega_0$, $\omega(z_s)\rightarrow\infty$ (far field), and the same resolution expression. We also notice that the equations work just as well for a diverging beam as they do for the converging beam shown in Fig. 7.3. For a diverging beam the scanner-waist distance, z_0, is negative, which means that the distance, z_s, to the optimum screen position is also negative, i.e. the best resolution is obtained at virtual image plane to the left of the scanner in Fig. 7.3. This virtual plane can be imaged by a lens after the scanner to create a real image plane with optimum resolution.

The conclusion of our treatment is that the resolution of an optical scanning system depends only on the scanner's range of angles, the optical beam radius on the scanner, the wavelength, and the resolution criteria established by the application. Converging, diverging, and focused incident beams give the same resolution proving that the resolution of the scanning mirror does not depend on the incident field. We simply have to adjust the projection system to obtain the maximum resolution.

7.2.3 Scanner Aperture

In Chapter 4 we discussed the effect of truncation on Gaussian beams. We considered the energy loss associated with truncation, i.e. we calculated the amount of energy that was left in the Gaussian beam and showed how much was blocked by the aperture and how much was forward-scattered into higher-order Gaussian fields. We also showed the effect of forward scattering on the beam shape in the far field. Both the energy loss and particularly the changes in beam shape are very important in miniaturized scanning systems. As we will see in later sections on MEMS implementations of optical scanners, it is difficult to fabricate and actuate large-area MEMS scanners. This fact forces us to make the scanning mirrors as small as possible for a given application. The expression we have derived for ideal, infinite mirrors shows that scanner resolution is directly proportional to the beam size on the scanning mirror. Now we have to ask how small we can make the scanning mirror and how the resolution is affected by the truncation caused by a finite size scanning mirror. Intuitively we understand that we cannot make the scanning mirror much smaller than the Gaussian beam size and expect to achieve the same resolution as for an infinite mirror. To create miniaturized scanning systems, we need to quantify the loss of resolution to be able to make optimum design trade offs.

To see how forward-scattered light affect scanner resolution, we go back to our diffraction calculations for truncated Gaussian Beams. Figure 7.4, which is the same as Fig. 4.19 repeated here for convenience, shows the far-field intensity pattern of three Gaussian Beams that have been truncated to various degrees at their beam waist along one axis.

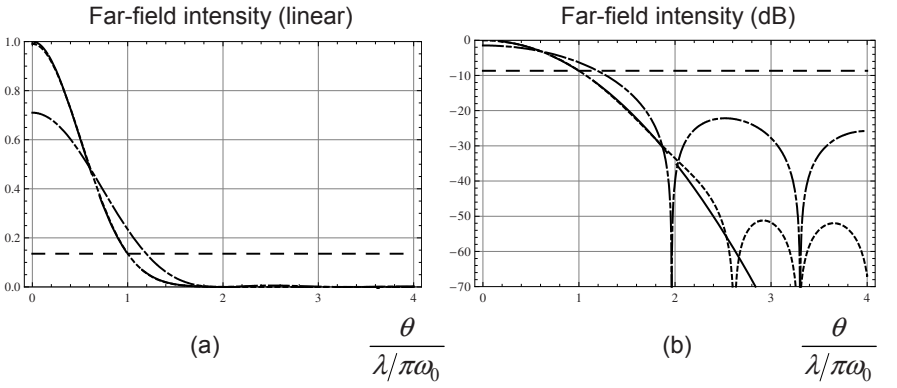


Figure 7.4 *Far-field (angular) profiles of Gaussians truncated in one dimension. The solid lines shows the profile of a Gaussian without truncation, the dashed lines is the profile of a beam truncated at four times its beam radius ($d=4\omega_0$), and the dot-dashed line at twice the beam radius ($d=2\ \omega_0$). (Same as Fig. 4.19. See Chapter 4.7.2 for detailed descriptions)*

The solid line shows an unperturbed Gaussian without truncation, the dashed line shows one has passed through an aperture that equals four times the beam radius, and the dot-dashed line shows one truncated at twice its beam radius. The far fields are calculated using the Fraunhofer Diffraction integral, which is separable in rectangular coordinates, so for rectangular apertures it is sufficient to consider truncation in one dimension. The derivations leading to the graphs of Fig. 7.4 are described in detail in Chapter 4.7.2.

The plots of Fig. 7.4 give us a first answer to the question of how large a scanning mirror should be. A mirror that is four times larger than the beam radius is nearly indistinguishable from the non-truncated beam and has side lobes only on the level of 10^{-5}. A mirror that is only twice as wide as the beam radius on the other hand has a far-field beam shape that is 37% wider than the non-truncated beam and has side lobes at the 1% level. Almost all practical miniaturized scanners have widths that are between these two values.

The effect of truncation on the central lobe of the truncated Gaussian is illustrated in more detail in Fig. 7.5. This figure shows the far-field of Gaussians of different beam radius after being truncated by the same mirror. The graphs gives us an answer to the question of what Gaussian beam size that gives the smallest far field for a given aperture size. We see that the central lobe decreases as we increase the

Gaussian beam radius from one fourth to one half of the mirror width. Increasing the beam radius to equal the mirror width further improves the far field, but the improvement is no longer as large and come at the cost of significant side lobes. Increasing the beam radius beyond that point yields only insignificant reduction of the central-lobe width, while the costs in terms of side lobes and on-axis attenuation are severe.

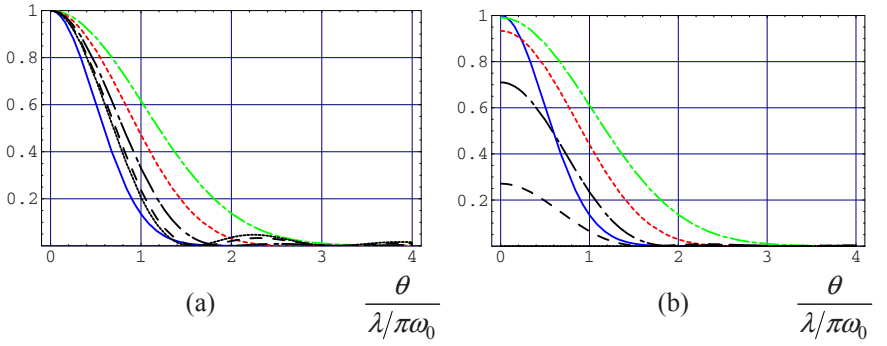


Figure 7.5 *Far-field (angular) profiles of Gaussians of different beam radius truncated in one dimension by the same mirror. The ratios of mirror width to beam radius are 4 (dot-dashed), 3 (dashed), 2 (dot-long-dashed), and 1 (long-dashed). The far-field pattern of a uniform distribution filling the mirror (dotted) and non-truncated Gaussian (solid) with a beam radius equal to half the mirror width are shown for comparison. The graphs in (a) are normalized to the on-axis intensity, emphasizing central-lobe width, while the graphs in (b) are normalized to the on-axis intensity of the non-truncated far-fields to emphasize energy loss.*

The conclusion we draw from Figs. 7.4 and 7.5 is that miniaturized scanners should be designed with apertures that are close to twice the beam radius. Systems that have relatively low contrast requirements, and therefore can tolerate significant side lobes in the far-field pattern, can use slightly smaller mirrors, but reducing the mirror size to equal the beam radius can only be done in the most crude systems. Systems that require high contrast of better than 20 dB need mirrors that are larger than twice the beam radius, but only in extreme cases is it necessary to employ mirrors that are larger than 3 times the beam radius. When calculating the resolution of mirrors that are smaller than about 3 times the beam radius we must take into consideration the increased width of the central lobe of the far field. For the case of a mirror that is twice the beam radius, this increase is 37%.

7.2.4 Surface Roughness, Curvature, and Bending of Micro Mirrors

Surface imperfections will typically not reduce the fundamental resolution of a scanning mirror. Although phase variations caused by surface imperfections can

change the overlap integral, defined in Chapter 2 as $\int_{surface}\left(\vec{E}_1\times\vec{H}^*_2+\vec{E}_2\times\vec{H}^*_1\right)\cdot d\vec{S}$, between two fields corresponding to two different settings of the scanner, these changes are negligible for typical surface imperfections seen in microscanners. In principle it is therefore possible to create high-resolution scanning systems using reflectors that deviate severely from the ideal flat surface. In practice, however, surface imperfections beyond a certain acceptable limit create insurmountable problems for the system design.

The acceptable level of surface imperfections depends among other things on the nature of the imperfection. Mirror imperfection can take many forms, but we will focus on three problems that are common to MEMS scanners; surface roughness, static curvature, and dynamic bending.

Surface Roughness

The effect of surface roughness can be understood by using a simple grating model to quantify reflection. We assume that a mirror surface can be modeled as a shallow grating as shown in Fig. 7.6. The reflected field from the surface can be modeled as the sum of two phasors with a relative phase given by the extra propagation experienced by the fields that are reflected from the bottom parts of the surface grating. The reflected light can then be calculated using the same procedure we applied in our modeling of the Michelson Interferometer in Chapter 2.3. (A more complete grating analysis will be presented in Chapter 10 on diffractive optical MEMS).

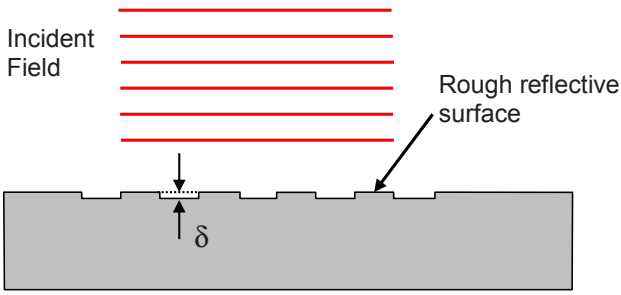


Figure 7.6 Surface roughness modeled as a shallow phase grating.

The part of the incident field that is reflected from the bottom of the shallow surface grating will in reflection have traveled a total distance of 2δ farther than the field that is reflected from the top. The relative phase of the two is therefore $\phi=\frac{2\pi\cdot 2\delta}{\lambda}$, where λ is the wavelength of the incident light. Following the treatment of the Michelson Interferometer in Chapter 2.3, we may write the following expression for the reflected intensity from the surface

$$I_r = I_i \cdot \cos^2 \frac{\phi}{2} = \frac{I_i}{2} \cdot (1 + \cos\phi) = \frac{I_i}{2} \cdot \left(1 + \cos\frac{2\pi \cdot 2\delta}{\lambda}\right) \tag{7.13}$$

where I_r and I_i are the reflected and incident optical intensities. We are analyzing optically flat surfaces with only shallow imperfections, so we expand the cosine function around $\delta=0$. The reflection from the surface then becomes

$$R = \frac{I_r}{I_i} = \frac{1}{2}\left(1+1-\frac{1}{2}\left(\frac{2\pi \cdot 2\delta}{\lambda}\right)^2\right) = 1-\left(\frac{2\pi \cdot \delta}{\lambda}\right)^2 \approx 1-40 \cdot \left(\frac{\delta}{\lambda}\right)^2 \tag{7.14}$$

This formula shows that even very minor surface roughness can cause significant problems. A standard surface-quality specification of $\lambda/12$ gives a reflection of only 73% as shown in the graph of Fig. 7.7. The two-level grating model we have used here represent a worst-case scenario, but surface roughness on the order of $\lambda/12$ is nevertheless too much for most applications. Many systems can tolerate $\lambda/20$, but the most critical systems require roughness better than $\lambda/100$. At this level the roughness is of minor consequence as can be seen from Fig. 7.7.

The good news is that surface roughness less than 25 nm Root-Mean-Square (RMS) is readily obtained for most substrates and thin films used in optical Microsystems. With some effort this can be reduced to less than 10 nm RMS. This corresponds to $\lambda/60$ and $\lambda/150$ for fiber-optic communications wavelengths (~1.5 um), and $\lambda/20$ and $\lambda/50$ for visible wavelengths. The conclusion is that surface roughness is an important problem to consider in the fabrication process, but it can be controlled and reduced to acceptable levels without undue difficulty and expense.

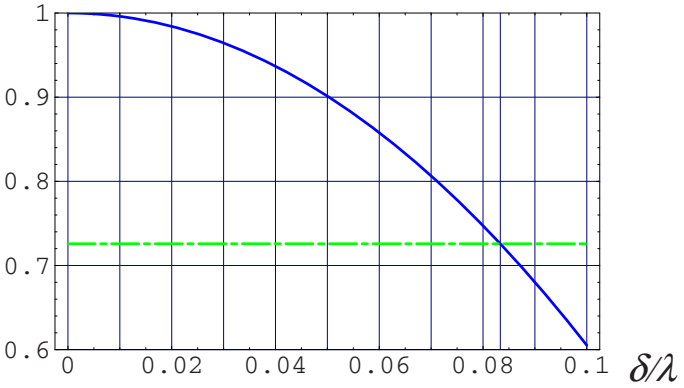


Figure 7.7 *Reflection from a shallow grating as a function of grating amplitude normalized to wavelength.*

Mirror Curvature

We emphasized in the introduction to this section that curvature of a mirror does not fundamentally reduce its resolution. The optical systems just have to be designed to compensate for the curvature of the mirror. The problem with curvature is therefore one of variability; if different mirrors have different curvature, then each mirror must be characterized and its optical systems must be tailored to its specifications. This is tedious, time consuming and costly, so it is important to understand how much curvature variations that can be tolerated. In the following discussion we will assume that the scanning mirror is nominally flat, and that any curvature is due to an imperfect fabrication process. The treatment works equally well for the case of a mirror that is designed to have a certain radius of curvature, as long as we read "radius of curvature" as "deviation from the nominal radius of curvature".[b]

Consider the comparison of a flat and a curved mirror in Fig. 7.8. The figure shows two overlaid beam profiles; one that is reflected off a curved mirror (beam profile shown as solid lines) and one that is reflected of a flat mirror(dashed lines). The beam coming off the flat mirror has its beam waist on the mirror, while the beam coming off the curved mirror is re-focused to a beam waist and then diverges to a larger far-field than the beam that reflects from a flat mirror. If the curved mirror had the opposite curvature, then the beam would have a virtual waist in front of the mirror, and the decrease in resolution would be the same as for a focusing mirror with a radius of curvature of the same magnitude.

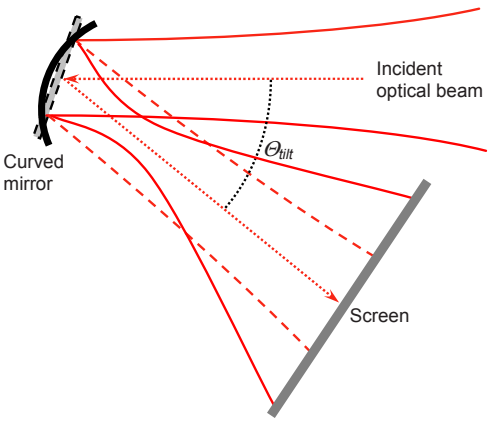


Figure 7.8. *The incident beam on the curved mirror is focused to a waist and then diffracts to create a wider far-field pattern (solid lines) than a beam reflected off a flat mirror (dashed lines).*

[b] Note that the radius of curvature of a flat mirror is infinite, so if a nominally-curved mirror meets its specifications on curvature, then its "deviation from the nominal radius of curvature" is also infinite.

The curved mirror acts as a lens with a focal length of $f=R/2$, where R is the radius of curvature of the mirror. From the formulas for focusing of Gaussian Beams that we derived in Chapter 4.3 we know that the ratio of the beam radius at the waist created by the curved mirror to the beam waist on the flat mirror is given by

$$\frac{\omega_2}{\omega_1} = \frac{f/z_1}{\sqrt{1+f^2/z_1^2}} = \frac{R}{\sqrt{\left(\frac{2\pi\cdot\omega_1^2}{\lambda}\right)^2 + R^2}} \tag{7.15}$$

where ω_1 is the beam radius of the waist on the mirror, ω_2 is the re-focused beam radius, z_1 is the Rayleigh length of the waist on the mirror, and λ is the wavelength. The angular far-field spread of a Gaussian beam is inversely proportional to the beam waist, so the ratio of the far-field diffraction angles, or equivalently the ratio of the beam radii on the projection screen, is the inverse of this expression

$$\frac{\theta_{curved}}{\theta_{flat}} = \sqrt{1+\left(\frac{2\pi\omega_1^2}{\lambda R}\right)^2} \tag{7.16}$$

To facilitate comparison between the effects of mirror curvature and surface roughness it is convenient to re-write this expression in terms of the mirror bow and diameter as defined in Fig. 7.9.

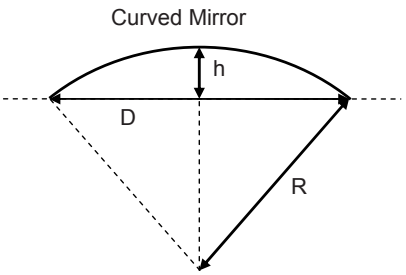


Figure 7.9. Definition of mirror bow.

The mirror bow can be written

$$h = R\left[1-\sqrt{1-\left(\frac{D}{2R}\right)^2}\right] \approx \frac{D^2}{8R} \tag{7.17}$$

where we have assumed that the radius of curvature is large compared to the mirror diameter. This is a very good approximation for all practical mirrors, particu-

larly because we are really considering deviation from a nominal R value, not the value itself. If we further assume that the mirror diameter is equal to 3 times the beam radius, then we have $R \approx \frac{9\omega_1^2}{8h}$ and the expression for the far field diffraction angle ratio becomes

$$\frac{\theta_{curved}}{\theta_{flat}} = \sqrt{1+\left(\frac{16\pi h}{9\lambda}\right)^2} \tag{7.18}$$

The two expressions we have derived for the ratio of the far field diffraction angles are plotted in Fig. 7.10. The plot in Fig. 7.10a demonstrates that a radius of curvature (or more correctly an error in the radius of curvature) equal or smaller than the Rayleigh length leads to a severe increase of the far-field beam radius and a corresponding reduction of the number of resolvable pixels of the scanner. When the radius of curvature is larger than twice the Rayleigh length, the resolution penalty for curvature is acceptably small for most applications.

The optical beams used with microscanners typically have beam radii of less than a few hundred micrometer. If we use ω_1=*400um* as an upper limit, we find that the Rayleigh lengths are equal to or less than *1m* and *33cm* for visible and telecommunication wavelengths, respectively. Creating mirrors with radii of curvature that are substantially larger than these values present a challenge in many optical MEMS technologies. This is particularly true for surface micromachining where stress gradients lead to curvature of free-standing thin films. Optical MEMS based on Silicon-on-Insulator materials, on the other hand, use mirror substrates that typically are tens of microns in thickness, so in these technologies mirror curvature is generally not a problem.

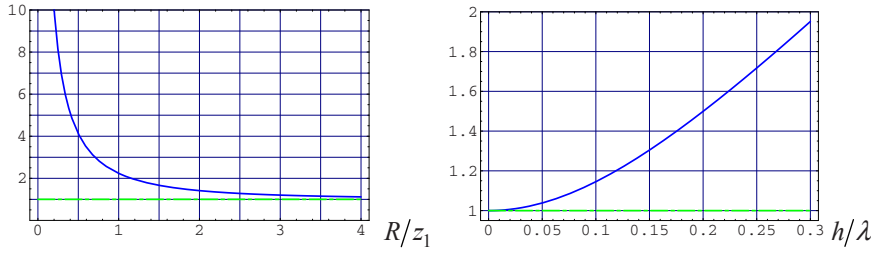


Figure 7.10. Ratio of the far field diffraction angle for a curved mirror to that of a flat mirror. The graphs in (a) shows the ratio plotted vs. the mirror radius of curvature normalized to the Rayleigh length, z_1, while (b) shows the ratio vs. mirror bow normalized to the wavelength.

Figure 7.10b contains the same information as 7.10a, but plotted against mirror bow instead of radius of curvature. Comparing this graph to the one of Fig. 7.7, it

seems that for a given deviation from a flat mirror, high spatial frequency imperfections like surface roughness have larger negative effect on system performance than low spatial frequency imperfections like curvature. We should be careful not to infer too much from this comparison because the two graphs are describing different physical effects, but it is generally true that low spatial frequency errors degrade performance less and are easier to compensate. This is fortuitous for Optical MEMS that tend to use interfaces that are polished as a part of the fabrication process, while it is a problem in integrated optics where etched sidewalls of waveguides may be rough and lead to excessive waveguide loss.

Dynamic bending

The two previous sections cover the most common forms of *static* imperfections in micromirrors; surface roughness caused by inadequate polishing and mirror curvature caused by stress and stress gradients. In addition to these static imperfections, microscanners also experience dynamic bending as a consequence of actuation forces and inertial forces acting on the mirror during operation. In principle we can avoid detrimental effects of static mirror curvature by correcting the radius of curvature of the field coming off the mirror in the beam forming (pre-compensation) or projecting (port-compensation) optics. Such compensation would be practically impossible for dynamic mirror bending, because the shape of the phase irregularities are more complex, and, most importantly, the compensation would have to dynamically change throughout the scanning oscillation cycle to adapt to the changing mirror imperfections[c]. Dynamic mirror bending is therefore an effect that must be understood so that its detrimental effects can be kept to an acceptable level.

Consider a scanning mirror that is driven into harmonic, oscillatory motion around a fixed axis of rotation as illustrated in Fig. 7.11. The mirror will experience actuation forces and inertial forces that bend the mirror to varying degrees throughout the oscillation cycle. At the extreme end of the cycle the mirror surface will have an s-shaped form as shown.

Assuming that the actuation forces are applied to the mirror as a moment at a fixed rotation axis and that the scanner has a rectangular mirror of constant cross section, it will experience an acceleration force that varies harmonically throughout the oscillation cycle and increases linearly with distance from the axis of rotation. The curvature of the mirror will then increase as the cube, and the deviation from flatness will increase as the fifth power of the distance from the axis (plus lower order terms to meet the boundary conditions) [3].

[c] Such dynamic compensation is theoretically conceivable by applying a dynamic phase corrector like a tunable lens or adaptive-optics mirror, but the author is unaware of any practical systems that use this strategy.

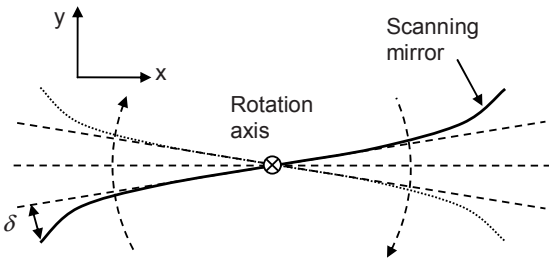


Figure 7.11. A scanning mirror that undergoes harmonic angular oscillations around a fixed axis of rotation will experience actuation forces and inertial forces that bend the mirror.

The assumption that the microactuator that drives the scanner applies a moment at the rotation axis of the mirror is not correct for most microscanners. The more typical situation is that the actuation forces are distributed over the mirror, either as electrostatic forces (as we will see in several examples at the end of this Chapter) or through linkages to the mirror away from the rotation axis. These distributed forces will in general complicate the shape of the mirror, but will not dramatically change the magnitude of the maximum deviation from flatness, shown as δ in Fig. 7.11. This is due to the fact that, in operation, the momentum created by the actuation forces will by definition equal the momentum of the acceleration forces. If the actuation forces are applied to the end of the mirror (worst case) we get roughly a doubling of the bending. In the interest of obtaining general results, we will simply ignore such "minor" effects. We will also ignore any elastic energy storage in the mirror itself. In other words, we assume that the mirror is driven well below the resonance frequency of its first-order mechanical resonance. This is a valid assumption for nearly all practical mirrors; the scanner might be driven at the resonance frequency of the overall actuator-mirror system, but for a well-designed scanner that frequency should be substantially lower than the fundamental resonance frequency of the mirror itself.

With these assumptions we need only consider the acceleration forces when calculating the bending of the mirror. The acceleration in the *y*-direction at a distance *x* from the rotation axis is

$$\ddot{y} = -\theta_{\max} x \cdot \omega^2 \cos(\omega t) \tag{7.19}$$

where ω is the natural frequency and θ_{max} the angular amplitude of the oscillations. The acceleration load per unit length in the x-direction of the mirror is then

$$w = -\theta_{\max} x \cdot \omega^2 \cos(\omega t) \cdot b \cdot h \cdot \rho \tag{7.20}$$

where *b* is the width of the mirror perpendicular to the x-direction, *h* the thickness, and ρ the density of the mirror. The standard reference work by Roark gives the

following formula for the end-point defection for a cantilever with a load per unit length that increases linearly from zero at the base to a maximum value of w at the end[d] [4]

$$\delta_{\max} = \frac{11w\left(\frac{L}{2}\right)^4}{120EI} \tag{7.21}$$

where L is the length of the mirror in the x-direction, E is Young's modulus for the mirror material, and I is the mirror's moment of inertia, given by

$$I = \frac{bh^3}{12} \tag{7.22}$$

Putting it all together we find the following formula for the maximum deviation from flatness

$$\delta_{\max} = \frac{11\rho}{160E} \cdot \theta_{\max} \omega^2 \frac{L^5}{h^2} \tag{7.23}$$

The material constants (ρ and E) are determined by the fabrication technology, and the maximum angle ($\theta_{\max}$), the scanning frequency (ω), and the mirror size (L) are set by the optical system specifications, so the only parameter that we are free to chose is the mirror thickness (h). To determine acceptable values for the thickness we need to know how much defection is acceptable.

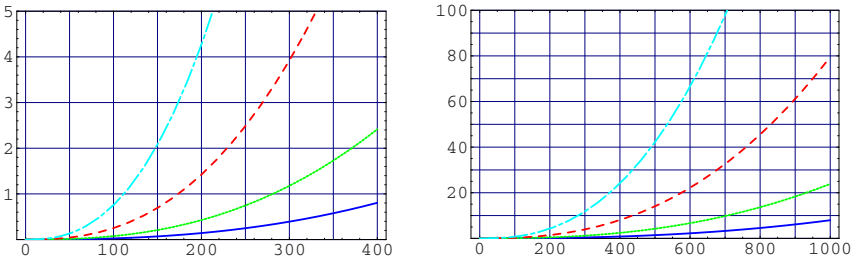
A mirror bent into an s-shape by actuation and acceleration forces will clearly diffract light differently than a flat mirror or a mirror of constant curvature. The exact diffraction pattern in the near and far field can be calculated using the Huygens-Fresnel diffraction integral. The result is a graph similar to Fig. 7.10b that shows that the maximum deviation from flatness must be a small fraction of a wavelength. The exact increase in diffraction angle will depend on the details of the mirror shape, but it is a safe assumption that if we keep the maximum deflection below *1/20* of a wavelength, then the increase will be acceptable. Using this criterion, we find that the mirror thickness must fulfill the inequality

$$\delta_{\max} = \frac{11\rho}{160E} \cdot \theta_{\max} \omega^2 \frac{L^5}{h^2} \le \frac{\lambda}{20} \Rightarrow h \ge \omega \sqrt{\frac{11\rho}{8E} \cdot \theta_{\max} \frac{L^5}{\lambda}} \tag{7.24}$$

[d] At first glance this formula seems inconsistent with our previous conclusion that the maximum deviation should go as the fifth power of the mirror length. Notice, however, that we here express the deflection in terms of the maximum load w. If we instead used the slope of the load, w/L, we would regain the fifth-power dependence.

This minimum mirror thickness is plotted vs. total mirror length in Fig. 7.12 for four different oscillation frequencies of 1.0 KHz, 3.0 KH, 10.0 KHz, and 30.0 KHz. The plots are for a poly-silicon mirror with a density of $\rho=2331\ kg/m^3$ and a Young's modulus of $E=160\ GPa$ [5]. The angular oscillation amplitude is assumed to be *0.12 radians=6.9 degrees*. This means that the mirror scans through a total range of 13.8 degrees, and the angle range of an optical beam reflected from the mirror is 27.6 degrees, which is a typical scan range for high-quality microscanners. The wavelength is set to 1.5 um, so for operation in the visible, the mirror thickness must be increased by $\sqrt{3}$.

The figure on the left (a) shows the required thickness for relatively short mirrors up to 400 um. We see that typical surface micromachining thicknesses on the order of 1 to 2 um cannot support large scanning mirrors. For example, a 2 um thick mirror operated at a modest 3.0 KHz must be less than about 370 um in length to avoid excessive dynamic bending, and if it is operated at 30 KHz, then it must be less than 150 um. A fabrication technology that allows thicker mirrors is required to create larger mirrors that can operate at high frequencies. Silicon-on-Insulator (SOI) MEMS will readily support fabrication of mirrors with thicknesses of a hundred microns or more[e]. Figure 7.12b shows that such thick SOI mirrors can measure up to one mm and still be operated at high scanning frequencies.



(a) (b)

Figure 7.12. Minimum mirror thickness [um] required to keep deviation from flatness to less than λ/20 at λ=1.5 um for a rectangular mirror of uniform thickness rotated around a fixed axis at frequencies of 1.0 KHz (solid line), 3.0 KHz (dotted line), 10 KHz (dashed line), and 30 KHz (dot-dashed line) as a function of totals mirror length [um] perpendicular to the axis of rotation. Figures (a) and (b) show the same data in two different ranges of mirror size.

The conclusion is that mirror flatness is a serious challenge in the design of microscanners. Mirrors fabricated using surface micromachining can only support modest sizes and frequencies. SOI mirrors do better, but the large thicknesses re-

[e] Strictly speaking we should use a direction-dependent Young's modulus when calculating bending of crystalline silicon, but for rough calculations the Young's modulus for polysilicon is a good approximation.

quired for large, fast scanners lead to bulky mirrors that require high-force actuators. Figure 7.11 and the associated discussion show that we can mitigate this problem by using mirrors that, instead of being of uniform thickness, have increasing thickness, and therefore the stiffness, towards the ends of the mirror. (This might seem counter-intuitive because increasing the thickness will also increase the acceleration forces and therefore the bending moments towards the mirror ends. However, the mass and therefore the bending moment increases linearly with thickness, while the stiffness goes as the third power of the thickness, so the net effect is to reduce bending.) A compelling implementation of this stiffer-towards-the-end strategy is to create a "microdrum" as shown schematically in Fig. 7.13.

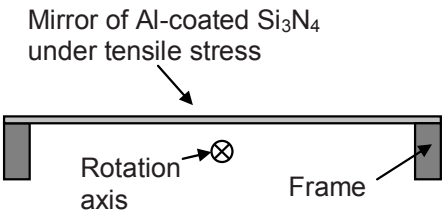


Figure 7.13. A thin film, stiffened by tensile stress and supported by a rigid frame, enables light-weight mirrors than can be operated at high frequencies without deforming to unacceptable levels.

Here the reflective surface is made of a thin film coated with a metal film to enhance reflectivity. The thin film is supported by a stiff frame made in SOI MEMS or some other type of bulk micromaching technology. The frame can be made rigid by increasing its thickness, while the thin film is kept flat by tensile stress. This type of structure is obviously more complicated to manufacture than a simple, uniformly thick mirror, but it has been demonstrated that high-quality mirrors and scanners can be fabricated this way [6,7].

7.3 Reflectivity of Metal Coated Micromirrors

In Chapter 7.2 we studied the influence of mirror shape on the reflected optical field. It is clear from our considerations that it is important to precisely control both surface roughness and mirror curvature to achieve the full resolution potential of microscanners. This makes it difficult to apply multilayer-stack reflectors in Optical MEMS, because of curvature variations caused by thermally induced stress in the multilayers. It has been shown that this challenge can be met by careful stress engineering and thermal control [8,9], but that still leaves the problem of material and process compatibility of multilayer stacks with IC and MEMS technology. Another method for making high-reflectivity surfaces is to use Photonic-

Crystal reflectors as described in Chapter 14, but that technology is still in its infancy. Consequently, most optical microsystems use simple metal coatings to enhance reflectivity. Choosing the materials and thicknesses of metal reflectors is therefore an important part of microscanner design.

Reflection of metal films on micromirror substrates can be calculated using the Fresnel-reflection formulae extended to multilayer structures, as we derived in Chapter 3. The combination of an Aluminum reflective film on a silicon or polysilicon mirror substrate is very common in Optical MEMS. Figure 7.14 shows the reflectivity of that type of layered structure as a function of Aluminum-film thickness and incident angle at visible and telecommunication wavelengths[f]. Other wavelengths and other combinations of reflective-film materials and substrates can straight-forwardly be modeled using the same formulae developed in Chapter 3.

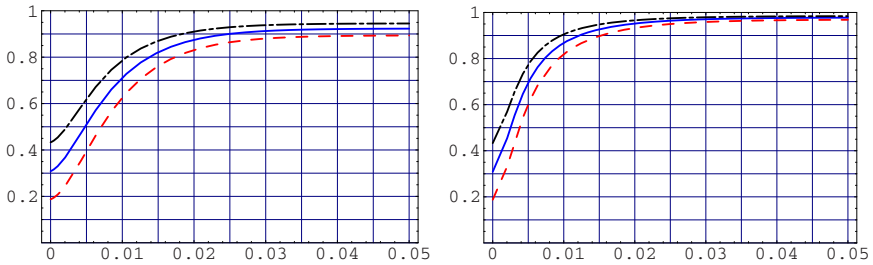


Figure 7.14. Reflectivity of Al on silicon as a function of Al thickness at 0.5 um (a) and 1.55 um (b) wavelength. The solid lines show the reflectivity at normal incidence, while the dashed (TM) and dot-dashed (TE) lines are for 45 degrees incident angle.

The index of refraction values (*n* is the real part, while *k* is the imaginary part that leads to absorption) the that are used to calculate the reflectivities are given in Table 7.1 together with the reflectivity values of pure of Aluminum and gold at the technologically important wavelengths of 500 nm (visible) and 1.550 nm (long-haul fiber optics).

The data of Fig. 7.14 and table 7.1 show that Aluminum is a very good reflector both at visible wavelengths and in the conventional band (c-band) around 1550 nm wavelength used for long-haul fiber-optical communication. For both wavelength ranges we see that the film thicknesses required to achieve close to bulk reflectivity are only on the order of 30 to 40 nm. Gold has slightly higher reflectivity in c-

[f] The reflectivity data of Fig. 7.14 are for light incident in air on an Aluminum film on a silicon substrate. The reflectivity is lower for light incident from the silicon side of the same film. This is not in violation of the reciprocity we demonstrated for loss-less mirrors in Chapter 2, because the Aluminum films absorb optical radiation at these wavelengths.

band as can be seen in Table 7.1, but it is much worse in the visible (~50%). Gold also represents an unacceptable contaminant in Si foundries, so it is much less used.

Incident Angle	n	k	0°	45°-TM	45°-TE
Aluminum at 500 nm	0.769	6.080	0.923	0.894	0.946
Aluminum at 1550 nm	1.440	16.00	0.978	0.969	0.984
Gold at 1550nm	0.550	11.4912	0.984	0.977	0.988

Table 7.1. Index of refraction [10] and reflectivity of Al and Au at selected wavelength wavelengths. The calculated reflection coefficients are valid for idealized, perfect surfaces. In practice, the reflectivity will be reduced by material imperfections and surface imperfections as discussed in preceding sections.

The plots of reflectivity for TE and TM polarized light at 45 degrees incident angle also shows that Aluminum mirrors have reasonable polarization characteristics, particularly at 1550nm wavelength. The TE and TM reflections are not identical as, but the differences are sufficiently small that they rarely represent the dominant polarization effects in practical systems.

The fact that only 30 to 40 nm is sufficient for high reflectivity simplifies the use of Aluminum in delicate mechanical structures. The small required reflector thicknesses and low Young's modulus relative to that of Silicon [6], make Aluminum reflectors very compliant and therefore unable to impart significant bending moments on their underlying substrates. In contrast to dielectric Bragg mirrors, Aluminum films therefore do not contribute to curvature of the mirror surface, in spite of the fact that their thermal expansion coefficient is vastly different from Silicon and most dielectrics used in MEMS technology.

The combination of high reflectivity over a broad range of wavelengths, compatibility with commercial IC process technology, and unproblematic incorporation into fragile MEMS structures, are the reasons that Aluminum is the material of choice for most Optical MEMS reflectors. The problems with high-reflectivity metals are power handling and temperature stability. Metals have significant loss at optical frequencies, which means that the light that is not reflected is absorbed[g]. That leads to heating of the reflectors and catastrophic failure is the incident power is too high. The exact power handling capacity depends strongly on the reflector's thermal resistance, which again depends on the mechanical design and the atmosphere [11,12]. Thermal management is therefore an important part of the design of optical MEMS.

[g] The exception is very thin films (e.g. <30nm of Aluminum) that are designed for partial transmission. Such films also have reduced reflectivity compared to thicker mirrors.

The cause of the low power handling capacity can be partly found in the fact that Aluminum and other high-reflectivity metals have low melting temperatures. Aluminum melts at 500°C and can therefore not tolerate operation or processing temperatures much beyond 480°C. This puts constraints on the post processing of Aluminum reflectors and therefore complicates their fabrication and limits their use. These shortcomings have lead to the development of MEMS-compatible Bragg reflectors and Photonic-Crystal reflectors as mentioned above.

7.4 Lens Scanners

Rotating mirrors are by far the most common implementation of MEMS scanners, but there are other architectures that have advantages for certain applications. The linearly-translating lens scanner, shown schematically in Fig. 7.15, is advantageous for many systems in that it operates in transmission, which allows a straight-forwards in-line systems design and simplifies the optical design. The lens scanner also uses a linear actuator, as opposed to the rotation actuators needed in scanning-mirror architectures.

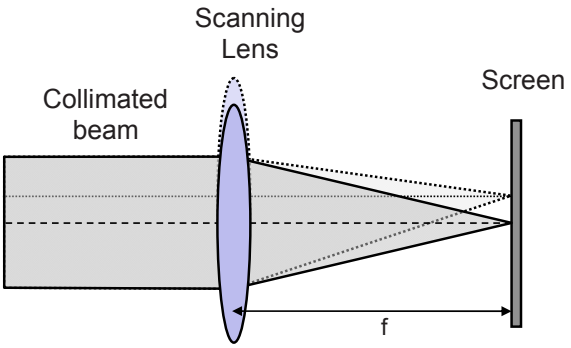


Figure 7.15. Schematic of microoptical scanning system showing the scanner in two positions (solid and dashed lines). The overlap between the scanned beams created by two settings of the scanner is constant.

In the prototypical lens scanner of Fig. 7.15 we assume that the incoming Gaussian beam is "collimated", i.e. its beam waist is on the lens or it is separated from the lens plane by an amount that is a small fraction of its Rayleigh length. Using the formulae we derived in Chapter 4, we can then express the beam radius at the waist on the screen (ω_{image}) in the focal plane of the lens as

$$\omega_{image} = \frac{f\lambda}{\pi \cdot \omega_{lens}} \tag{7.25}$$

where ω_{lens} is the beam radius of the Gaussian beam on the lens. In our simple system, the total range of motion of the image on the screen (Δ) equals the total motion of the scanning lens.

If we use the Full-Width-at-Half-Maximum (FWHM) resolution criterion we discussed in section 7.2.1, we find the following expression for the number of resolvable spots on the screen

$$N_{lens} = \frac{\Delta}{2r_{FWHM}} + 1 = \frac{\Delta}{2\sqrt{0.5 \cdot \ln 2} \cdot \omega_{image}} + 1 = \frac{\pi \cdot \omega_{lens} \cdot \Delta}{2\sqrt{0.5 \cdot \ln 2} \cdot f\lambda} + 1 \tag{7.26}$$

Based on the discussion of scanner apertures in section 7.2.3 we set the lens diameter minus the total motion equal to three times the beam radius

$$N_{lens} = \frac{\Delta}{2r_{FWHM}} + 1 = \frac{\pi \cdot (D - \Delta) \cdot \Delta}{6\sqrt{0.5 \cdot \ln 2} \cdot f\lambda} + 1 \approx 0.9 \cdot \frac{D}{f} \cdot \frac{\Delta}{\lambda} \tag{7.27}$$

We have derived a simple rule of thumb that says that the number of resolvable points of a lens scanner is roughly equal to the total motion of the lens, measured in wavelengths, divided by the *f*-number of the lens.

We can now compare this to the resolving power of a scanning mirror by rewriting the expression we found earlier to emphasize the similarities:

$$N_{mirror} = \frac{\pi}{2\sqrt{0.5 \cdot \ln 2}} \cdot \Delta\theta_{tilt} \cdot \frac{\omega_0}{\lambda} + 1 \approx \cdot 1.8 \cdot \frac{\Delta\theta_{tilt} \cdot 1.5\omega_0}{\lambda} + 1 \tag{7.28}$$

Here $\Delta\theta_{tilt} \cdot 1.5\omega$ represent the maximum motion of the edge of a scanning mirror that rotates around its center. It can be considered equal to the maximum motion, Δ, of the scanning lens, although we should keep in mind that in the mirror case, only the edge of the mirror moves the maximum distance, while in the case of the lens, the whole component moves the same. Clearly the factors in front of the motion-to-wavelength ratio also favor the mirror scanner, except for when lenses of extremely small f-numbers are used. Lastly, we note that lenses, and fast (low f-number) lenses in particular, typically are thicker and more massive than mirrors.

We conclude that lens scanners are less efficient that mirror scanners in terms of the amount of mass that must be moved to achieve a certain number of resolvable spots. For a given actuator technology with a given maximum force, this means that lens scanners have less speed, less resolution (due to less motion), or both. In optical microsystems speed and resolution are always important features, so lens scanners are only used in applications where the advantages of their in-line geometry are particularly significant.

7.5 Mechanical Scanner Design – One Dimensional Scanners

We have seen that the specifications on flatness and size for high-resolution optical scanning can be met by micromirrors fabricated in standard MEMS technology. Now we turn our attention to the mechanical structures that support the scanning mirrors. The mechanics is strongly dependent on the type of actuator that is used to drive the scanner into angular oscillations, so we need to consider the role of the actuator as we study mechanical scanner design. Microactuator principles and design are central to the development of all MEMS, and this topic is treated in numerous texts [13,14,15].

7.5.1 Transformation from Linear Motion to Rotation

One of the biggest differences between macroscopic and microscopic mechanical design is that there are no good linear-motion bearings or rotary bearings of any type in MEMS technology[h]. The fundamental reason for this is that surface forces, including friction that depends on randomly distributed surface roughness, are much larger relative to volume forces (mechanical stress, acceleration forces, electrostatic forces) in microscopic systems than in macroscopic systems. There are strong economical incentives for further miniaturization, so the trend is for microsystems to shrink further so that the ratio of surface to volume forces becomes yet larger. It is therefore unlikely that sliding bearings will play an important role in optical microsystems in the foreseeable future. More complicated bearing types, like roller bearings, jewel bearings, magnetic bearings, and fluid bearings can in principle be made in MEMS technology, but these solutions are getting more expensive as the structures are getting smaller.

Microhinges [16,17] have played an important role in the development of many types of MEMS applications, including optical packaging [18] and fiber switches [19,20]. The accuracy and repeatability of microhinges are sufficient for some applications, but not for all. More importantly, the long-term reliability of sliding micro-hinges has not been proven. The primary use of micro-hinges in practical system is therefore for one-time configuration, not continuous operation.

The lack of reliable, high-quality, sliding bearings makes it difficult to transform linear motion into rotation. Just like in the macroscopic domain, most actuators in the microscopic domain produce inherently linear motion. For example, in macroscopic combustion motors we use a series of axles with rotary bearings to trans-

[h] The Texas Instrument's Digital Light Processing technology relies on making and breaking small-area contacts between moving mirrors and an underlying substrate, but this is different from the precision sliding that must take place in a high-quality linear or rotary bearing.

form the linear motion of the pistons into rotation of the wheels of the vehicle. In microsystems we deploy a variety of linear actuators (parallel-plate electrostatic actuators, combdrives, piezoelectric actuators, thermal actuators) and use a set of flexural and torsional springs to convert linear motion into rotation.

Some common microsystem approaches to linear-motion-to-rotation transformation are shown in Fig. 7.16. Each of these methods have their own challenges and drawbacks. The scanning-cantilever mirror of 7.16a illustrates how the linear motion of piezoelectric or thermal actuators creates rotation. This is a very simple solution, but it has several problems. First, the elongation of the actuator both elongates and curves the cantilever. The elongation of the cantilever does not contribute to rotation, so the energy that goes into elongation is wasted. Second, the rotation is around an axis far from the center of the mirror, creating unwanted translation in addition to rotation. The net effect of these two difficulties is that the rotation of the scanning-cantilever mirror is less than what can be achieved with other means given the same actuator force and range.

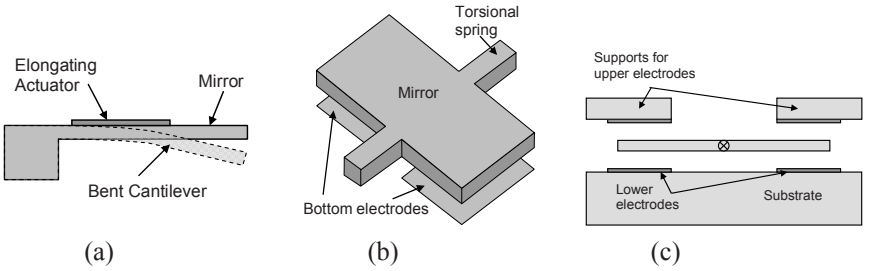


Figure 7.16. Compliant springs allow transformation from linear motion to rotation without the use of sliding joints. The scanning cantilever (a) is simple, but combines large translation with rotation. The rotational mirror (b) minimizes translation by using a suspending spring that is compliant in torsion and stiff under bending. The more complicated structure in (c) allows a pure torque to be applied to the mirror, so translation can be suppressed altogether.

Figure 7.16b shows a scanning mirror that is suspended by torsion bars that are designed to be compliant to twisting, but stiff in bending. When the mirror is subject to a combined torque and linear translation force from the electrostatic fields between the mirror and one of the bottom electrodes, then the resulting motion is predominantly rotation with only negligible bending. As we will see in the next section, it is not always possible in microtechnology to create a spring that is stiff in bending and compliant in rotation. The more complicated scanning mirror of Fig. 7.16c is therefore often the preferred solution. The existence of both upper and lower electrodes enables generation of a pure torque around the axis of rotation, so that translation of the mirror can be avoided altogether. Such pure-torque actuators can conveniently be implemented using vertical combdrives as described in Chapter 7.7.

7.5.2 Torsional Spring Design

In macroscopic machinery we can use complicated mechanical designs to create spring that are compliant in torsion and stiff in bending. In microsystems the limitations of the fabrication technology dictates that only simple structures be used, so we must consider the rotation and bending of simple beams of rectangular cross sections. Consider a uniform, fixed-end cantilever of length *L*, width *b*, and height *h* as shown in Fig. 7.17.

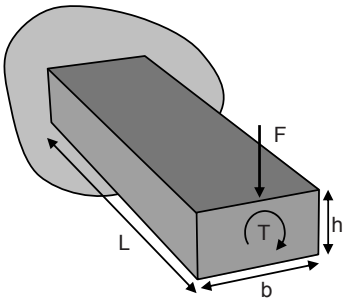


Figure 7.17. A cantilever with a simple rectangular cross section has a ratio of bending to torsional spring constants that is inversely proportional to the square of the length of the cantilever.

With a point load, *F*, applied to its end, the bending of the cantilever is given by [21]

$$y = \frac{L^3}{3EI} \cdot F = \frac{4L^3}{E \cdot bh^3} \cdot F \tag{7.29}$$

where *E* is the Young's modulus, and $I = \frac{bh^3}{12}$ is the moment of inertia of the cantilever. A torque, *T*, applied to the end of the cantilever yields a rotation [22]

$$\phi = \frac{L}{c_2 \cdot G \cdot bh^3} \cdot T \tag{7.30}$$

and maximum stress of

$$\tau_{\max} = \frac{T}{c_1 \cdot bh^2} \tag{7.31}$$

In these equations *G* is the Shear modulus, which can be expressed in terms of Young's modulus (*E*) and the Poison ratio (ν) as $G = \frac{E}{2(1+\nu)}$, and the constants c_1 and c_2 depend on the cross-section of the cantilever as given in Table 7.2.

The ratio of bending stiffness $\left(\frac{F}{y}=\frac{E\cdot bh^3}{4L^3}\right)$ to torsional stiffness $\left(\frac{T}{\phi}=\frac{c_2\cdot G\cdot bh^3}{L}\right)$ of the cantilever is

$$\frac{F/y}{T/\phi}=\frac{1+\nu}{2\cdot c_2\cdot L^2} \tag{7.32}$$

The ratio is inversely proportional to length squared, so we achieve increased bending stiffness over torsional stiffness by simply shortening the cantilever.

h/b	c_1	c_2
1.0	0.208	0.14068
1.2	0.219	0.1661
1.5	0.231	0.1958
2.0	0.246	0.229
2.5	0.258	0.249
3.0	0.267	0.263
4.0	0.282	0.281
5.0	0.291	0.291
10.0	0.312	0.312
∞	0.333	0.333

Table 7.2. *Coefficient for use in formulas for maximum stress and angular deflection of torsion bars with rectangular cross sections [19].*

When applying this approach, we must take care not to let the maximum stress become too large. Consider for simplicity a torsion bar with a square cross section (h=b). The torsional stiffness is then proportional to the fourth power of the cross-sectional side divided by the bar length.

$$\frac{T}{\phi}\propto\frac{h^4}{L} \tag{7.33}$$

This means that to maintain a constant stiffness as we scale the spring to smaller dimensions, we must reduce the thickness of the torsion bar as the fourth root of the length. From the above formulas we also find that the maximum stress is proportional to the angular rotation multiplied by the ratio of height to length

$$\tau_{\max}\propto\phi\cdot\frac{h}{L} \tag{7.34}$$

which means that if we scale the torsion bar linearly ($h\propto L$), then the maximum stress in the bar is given by the rotation. If on the other hand, we keep the torsion-

bar stiffness constant, then the maximum stress is proportional to the length to the power of negative three fourths

$$\tau_{\max}\big|_{L\propto h^4} \propto \phi \cdot L^{-\frac{3}{4}} \tag{7.35}$$

We see from these three proportionalities that torsional bars scale reasonably well to small sizes. As we scale torsion bars to shorter lengths, we maintain the torsional stiffness by scaling the side-dimension as the fourth root of the length, but we can only do that until we have reached the maximum allowed stress. As we scale beyond that point, the side dimension must be reduced linearly with the length, with the consequence that the torsional stiffness goes down as the third power of the length. Under the same conditions the bending stiffness goes down linearly in the length.

This scaling characteristic presents two problems to the MEMS designer. It is always possible to achieve a desired ration of bending stiffness to torsional stiffness, but in some cases it means reducing the torsional stiffness to avoid increasing the maximum stress beyond acceptable levels. The second problem is more a practical one; it is often difficult in a given MEMS technology to define torsion bars with sufficiently small cross sections and sufficiently short lengths to achieve the desired bending to rotation ratio. This is a bigger problem when the actuation forces are small so that very compliant structures are required.

These problems of scaling of torsional springs to obtain a favorable ratio of bending stiffness to torsional stiffness have lead many MEMS designers to the concept of a supported flexural bearing. A typical example is shown in Fig. 7.17. Here a scanning mirror suspended by torsion bars and actuated by parallel-plate electrostatic actuators is supported by a ridge that is in frictional contact with the mirror and prevents the mirror from being linearly translated towards the substrate electrodes.

This type of supported flexural bearing was first demonstrated for micro-optical scanners by Petersen [23], and has later been used in a large number of MEMS and microoptics applications. The accuracy, repeatability, and reliability of this type of structures are, however, still unproven, so, as for the microhinges discussed above, the use of supported flexural bearings have been confined to laboratory demonstrations, and they have not had significant commercial impact.

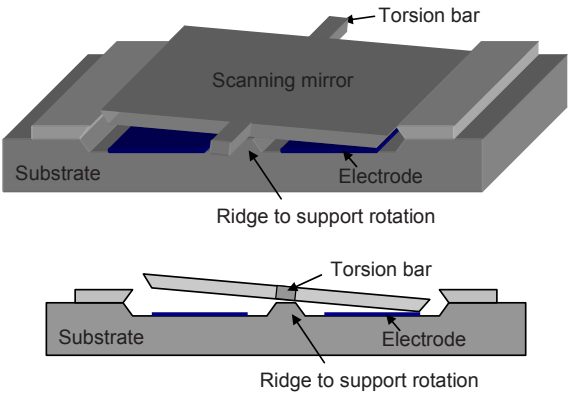


Figure 7.17. Perspective drawing (top) and cross section (bottom) of scanning mirror suspended by torsion-bar flexural bearings with ridge support to avoid translational motion of the mirror towards the electrodes.

7.5.3 Mechanical Resonances

The lack of good sliding joints means that most MEMS actuator systems use some type of mechanical spring, e.g. a torsion bar or bending beam, to suspend the mass to be moved and to provide a restoring force for the actuator. The three scanners of Fig. 7.16 are all of this construction, but the principle is applied to a much wider class of MEMS than just translation-to-rotation transformers. Such spring-mass systems can to first order be described as damped harmonic oscillators like the one shown in Fig. 7.18.

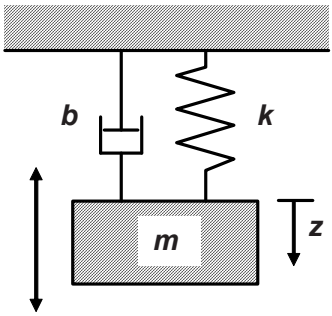


Figure 7.18. Simple harmonic oscillator model of a MEMS actuator with an actuated mass suspended by a mechanical spring. The dashpot represents all losses in the system.

The force-balance equation for this spring-mass system is

$$m\frac{d^2z}{dt^2}+b\frac{dz}{dt}+k\cdot z=F(t) \tag{7.36}$$

where m is the mass, b is the damping coefficient, k is the spring constant, and F is an actuation force that is acting on the mass. We now assume that the actuation force has an harmonic time dependence, $F=F_0\cdot\cos(\omega t)$, i.e. we use the phasor notation introduced in Chapter 2. The equation then takes the form

$$m\cdot(j\omega)^2 z+b\cdot(j\omega)^2 z+k\cdot z=F_0 \tag{7.37}$$

with the solution

$$z=\frac{\frac{F_0}{m}}{-\omega^2+j\omega\cdot\frac{b}{m}+\frac{k}{m}}\Rightarrow\frac{z}{F_0/k}=\frac{1}{-\frac{\omega^2}{\omega_0^2}+j\frac{\omega}{\omega_0}\cdot\frac{1}{Q}+1} \tag{7.38}$$

In the last expression we have used the standard definitions of resonance frequency, $\omega_0=\sqrt{\frac{k}{m}}$, and Quality factor, $Q=\frac{\omega_0 m}{b}$. The logarithm of the amplitude and the phase of the normalized response, $\frac{z}{F_0/k}$, is plotted in Fig. 7.19 for Q-values ranging from 0.2 to 10.

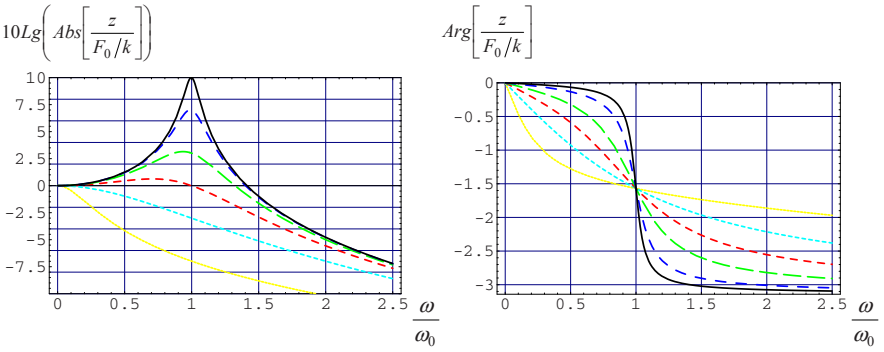


Figure 7.19. Normalized amplitude (a) and phase (b) response of a harmonic oscillator as a function of normalized frequency for Q-values ranging from Q=0.2 (lower dotted line in a, flatter dotted line in b) to Q=10 (upper, solid line in a, steeper, solid line in b).

The plots show that the damping, described by the Quality factor, has a profound effect on the mechanical response of the spring-mass system. Systems with Q-factors below 0.5 are said to be over-damped, and they exhibit substantially lower

bandwidth than one would expect from their resonance frequencies. For example, the system with $Q=0.2$ has a -3dB frequency of $0.36 \cdot \omega_0$. Reduced bandwidth is rarely a desired characteristic, so in practical MEMS design we try to avoid over-damping.

Quality factors above 0.5 lead to under-damping. We see from the graphs that under-damped systems have bandwidth well in excess of the resonance frequency and that the excess bandwidth grows increases with increasing Q-factors. The system with $Q=10$ for example has a -3dB frequency of $1.73 \cdot \omega_0$. This extra bandwidth comes, however, at the cost of a very non-uniform frequency response. The response is strongly resonantly enhanced around the resonance frequency, which will lead to signal distortion, overshoot in the transient response, and in some cases even self oscillations. Under damping is therefore desired in systems that are designed for single frequency operation and that require large motion. A good example is microcantilevers for tapping-mode Atomic Force Microscopy (AFM). AFM cantilevers vibrate at a near constant frequency and require relatively large motion, so the resonant enhancement of an under-damped system is ideal for this application.

When the Q-factor is exactly 0.5 the system is critically damped, meaning it that has the minimum damping that is required to avoid overshoot in the response to a step-function in the applied force. Overshoot, or "ringing", leads to increased settling times after a change in input, so critically damped systems are for many practical purposes faster[i] than any other system with the same resonance frequency, but different damping. Critical damping is therefore a goal in the design of many mechanical systems.

Unfortunately, it is difficult to control mechanical damping in microsystems. Just as in macroscopic machinery, the MEMS designer tries to avoid losses due to friction and material damping, because it leads to positioning errors, wear, and problems with repeatability and reliability. We are then left with fluid flow as the dissipation mechanism that we can attempt to control to achieve the desired damping characteristics in our devices. Fluid flow in optical MEMS typically means gas flow because the combination of high speeds required and the relatively small forces provided by microactuators makes it impossible to operate Optical MEMS in liquids[j].

Control of damping in Optical MEMS in general, and microscanners in particular, therefore comes down to controlling dissipation in the gas flow around the scan-

[i] Exactly what we consider a faster response will of course depend on the application. For example, if we have an application in which ringing in response to a input variation is of no consequence, then an under-damped system will be considered faster than a critically damped one.

[j] The obvious exception is biological applications that require Optical scanners operating in aqueous environments.

ning mirrors and other mechanical parts that are being moved by the microactuators. In certain cases it is indeed possible to create systems with the desired damping characteristics by correct design of the aerodynamics of the moving parts, by carefully selecting the ambient gas species and pressure, and by adding gas-flow channels and pressure chambers. However, this is a complex engineering task that adds complexity and cost to the system implementation. An additional technical difficulty is that damping depends on the mode profile (see the following section, 7.5.4), making it difficult to design air-flow systems that will achieve the desired a damping of all the resonant mechanical modes.

In practical MEMS implementations we therefore most often have to accept the fact that the damping cannot be engineered to have the optimum value, and focus our efforts on designing systems that can tolerate undesirable damping characteristics without unacceptable consequences. For systems that are dominated by squeezed-film damping [24] that typically means having to live with the lower bandwidth of an over-damped response. High-resolution microscanners, on the other hand, are typically not affected by squeezed-film damping, because rotation is less efficient than linear translation in trapping and squeezing a thin film of gas. Instead the dissipation is dominated by air flow around the rotating mirrors; a much weaker damping effect. Almost all reported high-resolution microscanners are therefore significantly under-damped. The high resonant gain of under-damped systems leads to a number of operational difficulties, including undesired motion at the resonance frequency due to non-linearities in the actuators, and coupling to higher-order resonant modes. Examples of such undesired effects and ways of avoiding or mitigating them are described in Section 7.7.

7.5.4 Higher-Order Mechanical Resonances

The ideal harmonic oscillator described in Section 7.5.3 serves as a simple, yet useful, first-order model of many MEMS structures. Its mass-less spring and point mass, that is confined to linear translation in one dimension, allow only one form of potential energy storage (stretching of the spring) and one form of kinetic energy storage (motion of the mass), and therefore a single mechanical resonance. Real mechanical systems have massive springs and distributed masses that can move with all six degrees of freedom of 3-dimensional space. The distributed springs and masses give real machinery a large number of distinct ways to store potential and kinetic energy and a correspondingly large number of mechanical resonances[k].

An example of a scanning mirror with multiple resonances is shown in Fig. 7.20. This mirror has dimensions that are typical for high-resolution scanners made in Silicon-on-insulator (SOI) materials using Deep-Reactive-Ion-Etching (DRIE). The mirror is 500 by 500 um, the torsion bars are 300 um long and 4 um wide, and

[k] In a continuum model, the number of resonances is infinite.

both the mirror and the springs are 21 um thick. (The geometrical dimensions, as well as the material parameters, of the scanner in Fig. 7.20 are listed in Table 7.3 in the next section).

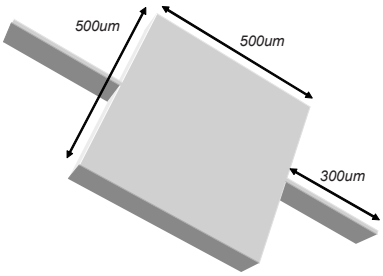


Figure 7.20. 1-dimensional scanning mirror suspended by torsion bars designed to allow the mirror to rotate around the rotation axis. The distributed springs and masses lead to multiple higher-order resonances that must be considered in the design and operation of the scanner.

The distributed mass and the fact that the torsional springs are compliant in all directions (although significantly stiffer axially than in transversal bending) give this scanner multiple resonant modes. The five modes of lowest order (i.e. lowest resonance frequency) are shown in Fig. 7.21 and the characteristics of the resonances are summarized in Table 7.3. The lowest-order resonant mode is rotation around the torsion bars, followed in order by (2) in-plane, linear translation with the springs bent in the horizontal direction, (3) in-plane rotation, (4) out-of-plane linear, translation, and (5) rotation around an axis perpendicular to the torsion bars.

The scanner is well designed in that its lowest resonance frequency belongs to the mode that has the desired motion: rotation around the torsional springs. The ratio of the resonant frequency of the second lowest mode, in-plane translation, to that of the fundamental mode is also sufficient to avoid strong coupling between the modes. This ratio can be extended further by making the cross section of the torsion bars closer to rectangular. This would lower the resonance frequency of the 4^{th} mode, but a reduction of its resonance will not have consequences until it is lower than the resonance frequency of the 2^{nd} order mode.

The modes described in Fig. 7.21 and Table 7.3 illustrate an important point of microscanner design; mechanical systems have complicated responses with multiple resonances. In macroscopic machinery, unwanted motion is damped by inserting dissipative shock absorbers that attenuates undesired modes of operation. In microscanner this is not a practical solution, because it is prohibitively difficult to create damping structures of the correct construction. The saving grace for microscanners is that they are relatively simple so that it is possible to design the overall structure to avoid coupling of energy to unwanted modes of operation. (Imagine what the spectrum of resonances for a complicated piece of machinery like a car

looks like, and how it would behave if all resonances were under-damped!). The conclusion is that microscanners, like all under-damped mechanical structures, must be kept as simple as possible so that they can be designed so that the desired motion is the fundamental resonance and that all higher-order resonance frequencies are substantially higher than the fundamental.

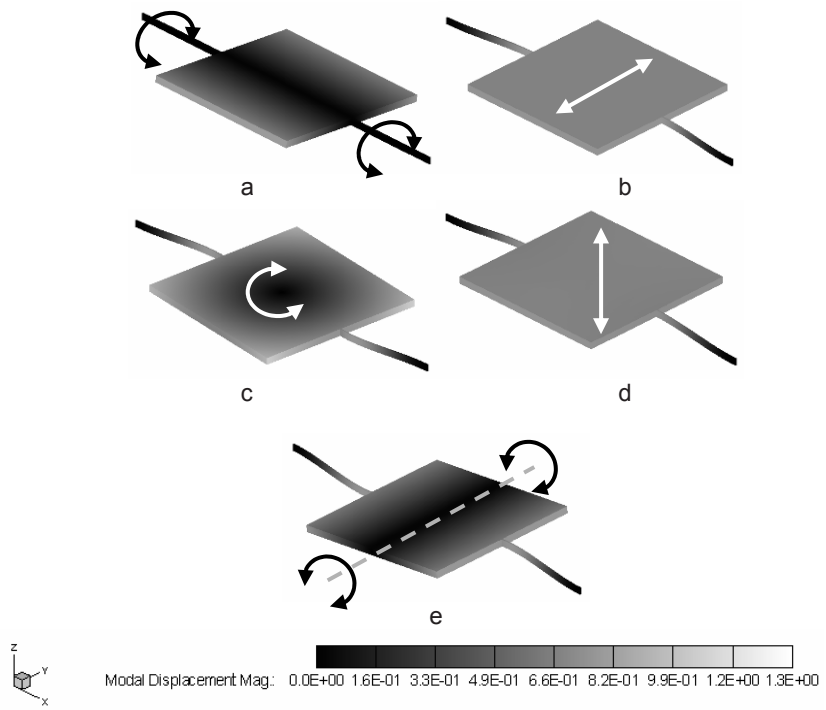


Figure 7.21. Modal analysis of 1-D scanner.

Mode #	Motion	Resonance Frequency
1	Rotation around the torsion bars (7.21a)	4.12 kHz
2	In-plane, side-to-side linear translation (7.21b)	5.73 kHz
3	In-plane rotation around the center of the mirror (7.21c)	11.44 kHz
4	Out-of-plane, up-and-down linear translation (7.21d)	28.05 kHz
5	Rotation around an axis perpendicular to the length direction of the torsion bars (7.21e)	79.74 kHz

Table 7.3. Mode descriptions and resonance frequencies of the scanner shown in Figs. 7.20 and 7.21.

7.6 Two Dimensional Scanners

In our discussion of the mechanics of microscanners, we have so far concentrated on one-axis, or one-dimensional, scanners. Now we extend the discussion to two-axis, or two dimensional, scanning that is required by many applications. There are two quite different ways to rotate a scanning mirror around two orthogonal axes; we can use a gimbal or a universal joint. The gimbal consists of two frames that are connected with bearings so that they can rotate around orthogonal axes as shown in Fig. 7.22. In MEMS gimbals the bearings are flexural and the two frames are typically in the same plane when no actuation force is applied.

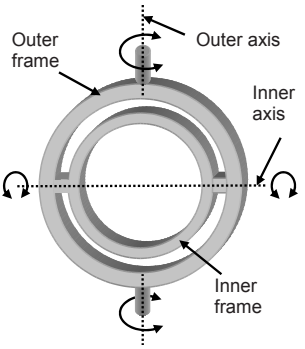


Figure 7.22. Illustration of the gimbal principle. When implemented with sliding, rotary bearings, the gimbal allows full rotation of both frames around their respective rotation axes. MEMS gimbals are typically implemented with flexure bearing, so that oscillatory rotation, but not continuous one-directional rotation, is allowed.

In precision macroscopic machinery we use universal joints that, just like gimbals, are based on rotary bearings. A common architecture is a block with axels in orthogonal directions joining two shafts than can rotate so that the angle between them can take almost any value. This construction is straightforward in the macroscopic domain, but would be prohibitively difficult to create with MEMS technology. Fortunately, the universal joint is particularly simple to implement with flexure bearings; a simple spring that is compliant in two dimensions is in principle all that is needed. In practical MEMS we typically use more than one spring to establish the required two degrees of freedom of rotation. An example of a flexure-bearing universal joint with multiple springs is shown in section 7.7.

Clearly the mechanical structures of MEMS gimbals and universal joints are much more complex than the simple scanning mirror we analyzed in section 7.5.4. Consequently we have to pay close attention to the different resonant modes of the overall structure. As for the 1-axis scanner, the preferred rotations should have the lowest resonance frequencies to avoid parasitic motion in other modes.

As an example of a relatively simple design that fulfills this criterion, consider the gimbal shown in Fig. 7.23. This structure is made from to layers of silicon, each 21 um thick. These two layers are connected mechanically, but isolated electrically, by a silicon dioxide layer that is 0.5 um thick. The outer frame is almost in it entirety constructed of both silicon layers. Only in the area around the bases of the inner springs are there gaps in the lower Si layer so that the two bottom halves of the outer frame can be held at different electrostatic potentials. The inner frame and the inner springs are made of the upper Si layer only. The dimensions and material parameters of the gimbal are given in Table 7.4.

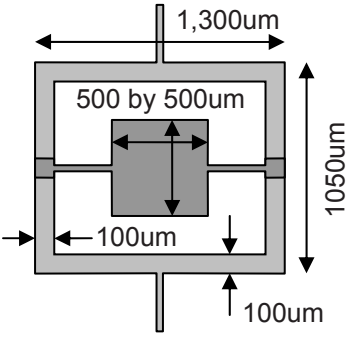


Figure 7.23. Gimbal layout seen from below. The outer frame is constructed from two layers of silicon except in the areas around the bases of the inner springs where the lower layer is removed. The mirror and the inner springs are constructed from only the upper Si layer.

Scanner Dimensions			
	Length	Width	Thickness
Mirror	500 um	500 um	21 um
Inner spring	300 um	4 um	21 um
Outer spring	250 um	4 um	42.5 um
Silicon Dioxide			0.5 um
Silicon Material Parameters			
Density	2500 kg/m^3		
Young's Modulus	169 GPa		
Shear Modulus	65 GPa		
Poison Ratio	0.3		

Table 7.4. Geometrical and materials parameters of the scanner shown in Fig. 7.23.

The lowest order resonant modes of this structure are shown in Fig. 7.24 and Table 7.5. We see that the fundamental mode is the desired rotation around the outer torsion bars. This is not surprising given the large mass loading of this mode. The second mode is the desired rotation around the inner torsion bars. The next three modes are all due to sideways bending of the torsion bars. The resonance frequencies of these three modes are close to that of the inner-axis rotation mode. Care must be taken to ensure that fabrication variations, e.g. over-etching that makes the springs thinner, don't allow these to become one of the dominant modes

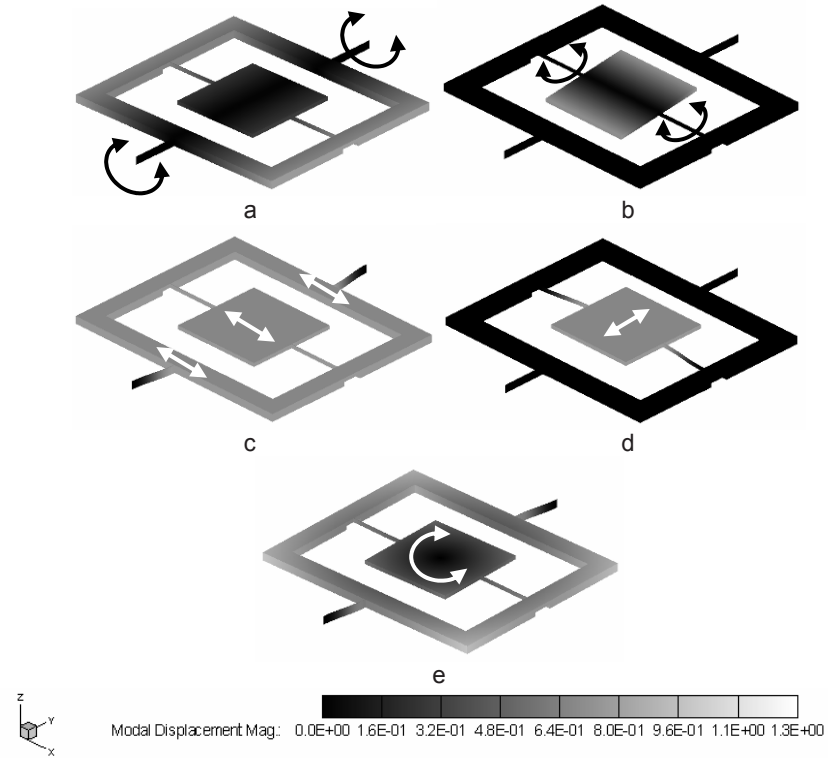


Figure 7.24. Modal analysis of 2-D gimbaled scanner.

Whether the differences between the resonance frequencies of modes 3 through 5 and that of mode 2 are sufficient in this example will depend on the application. If the rotations of the mirror are driven over wide ranges of frequencies up to their resonances, then they will occasionally be excited at sub-harmonics of the higher order modes. The actuators will have non-linearities, so these sub-harmonics will couple to, and set-up motion of, the higher-order modes. Such parasitic motion in the higher orders will lead to high-frequency dither of the scanning beam. These types of effects must be avoided in high-precision scanning applications. If, on the other hand, the rotations of the mirror are at fixed frequencies, as they would

be to set up a fixed raster scanning of Lissajou pattern, then coupling to higher-order modes can be suppressed by correct choices of the driving frequencies.

Mode #	Motion	Resonance Frequency
1	Rotation around the outer torsion bars (7.25a)	1.10 kHz
2	Rotation around the inner torsion bars (7.25b)	4.12 kHz
3	In-plane, side-to-side linear translation due to bending of the outer torsion bars (7.25c)	5.18 kHz
4	In-plane, side-to-side linear translation due to bending of the inner torsion bars (7.25d)	5.73 kHz
5	In-plane rotation around the center of the mirror due to bending of the outer torsion bars (7.25e)	6.05 kHz

Table 7.5. Mode descriptions and resonance frequencies of the 2-axis, gimbaled scanner shown in Figs. 7.23 and 7.24.

The bottom line is that mechanical resonances must be carefully considered in under-damped optical scanners. In principle we would like to design the system to only have the desired modes of operation, but that is in practice impossible, so the best we can do is to make sure that we are not coupling energy into unwanted mechanical modes. That can be achieved through a combination of clever mechanical design and careful operation.

7.7 High Resolution 2-D Scanners – Design Examples

In this section we give three design examples of 2-D scanners. These examples give different design solutions to the related problems of mechanical stability and efficient actuation. They are chosen are chosen to give the reader an appreciation of the vast the design space for MEMS scanners. The examples do by no means span this space; there are numerous other scanner designs, each with their own advantages and drawbacks, reported in the literature. Nor do the examples represent "best" solutions. What's best depend on the application, and it is a moving target with new improvements being introduced at a rapid pace.

7.7.1 Gimbaled Scanner

A simple, but functional, architecture implemented in two SOI layers on a silicon substrate is shown in Fig. 7.25 [25]. The gimbal is driven by electrostatic comb-drive actuators on its two orthogonal axes of rotation. For clarity only a few combs are shown in the schematic. In a practical design, we seek to maximize the number of combs to maximize the force. The upper SOI layer is grounded as shown. An applied voltage to the left electrode (V_1) appears across the inner, left combdrive and rotates the mirror counter-clock wise around the inner springs,

while a voltage applied to the right electrode (V_2) rotates the mirror clockwise around the inner springs. Likewise, voltages applied to the upper electrode (V_3) and the lower electrode (V_4) rotates the mirror around the outer springs.

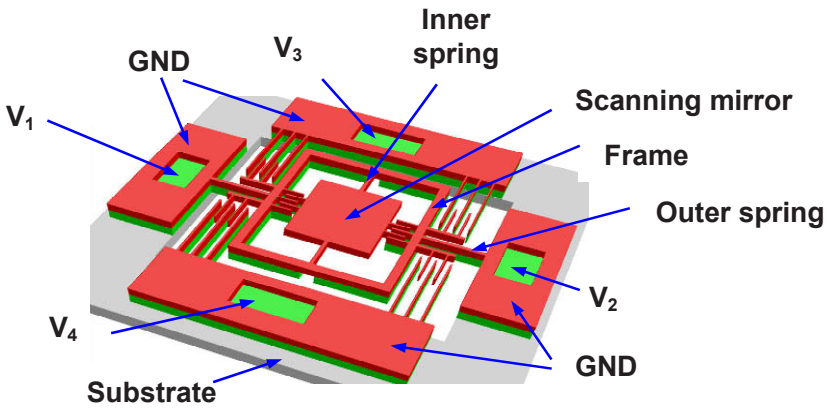


Figure 7.25. Schematic of 2-D gimbaled microscanner implemented in two SOI layers on a Si substrate. The scanning mirror rotates with respect to the frame around the inner springs, while the frame rotates with respect to the substrate around an orthogonal axis defined by the outer springs.

For clarity, the schematic of Fig. 7.25 is shown with only a few comb teeth in each drive. In a practical layout, like the one shown in Fig. 7.26, the emphasis is on maximizing the number of comb teeth to minimize the voltage needed to obtain the required total force.

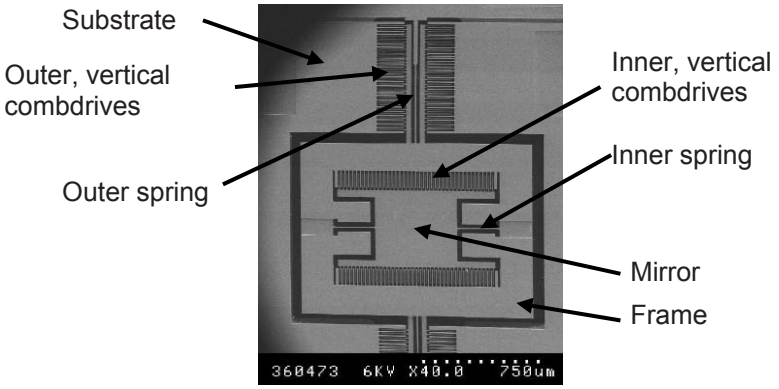


Figure 7.26. Scanning Electron Micrograph of a two-axis gimbaled micromirror. The mirror is driven into rotation around the inner springs by the electrostatic forces of the inner combdrives. The outer springs are made of two insulated silicon layers so that the driving voltages can be delivered to the inner comb actuators.

The relatively complex mechanical structure of the gimbaled microscanner can be fabricated using only five masks as shown in Fig. 7.27. The key processing steps are the double-masking (Masks 2 and 3) of LTO mask layer (steps c and d), and the self-aligned patterning of the lower and upper teeth of the vertical combdrives (step f). The double masking avoids the difficulty of performing lithography on wafers with large vertical height differences. It is a standard MEMS-fabrication trick that is used in numerous devices. Equally important is the process that allows the upper and lower combteeth to be defined by the same mask (self alignment), because it ensured that the combdrives are aligned and therefore can tolerate the largest possible applied voltages to create the maximum force and torque. (see Appendix B)

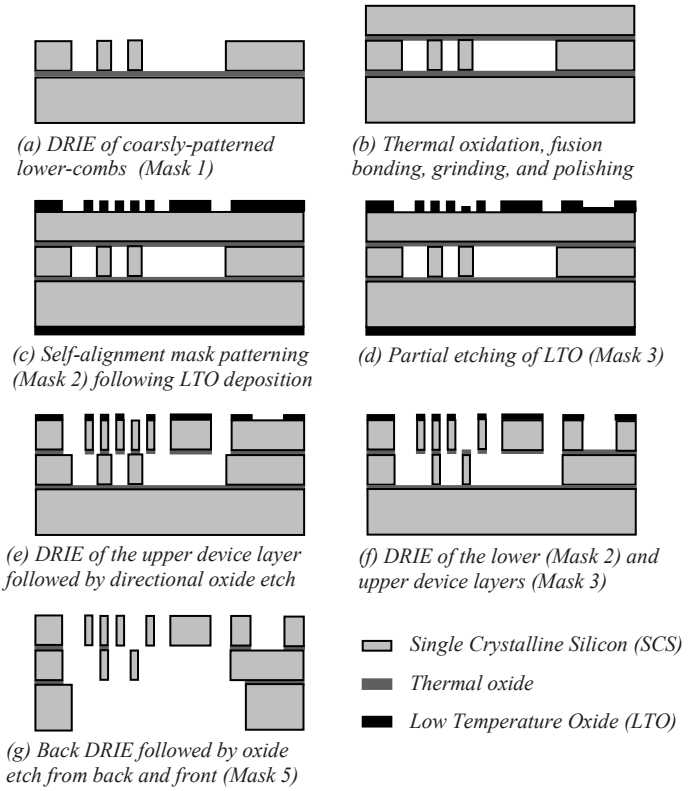


Figure 7.27 *Fabrication process for two-axis gimbaled micromirrors (not to scale). The process requires two SOI layers (Silicon-on-insulator). The lower comb teeth are initially coarsely patterned by the first mask before the second SOI layer is bonded to the wafer. These teeth are later trimmed (step f) to be self-aligned to the upper comb teeth.*

High-performance optical scanners have been demonstrated using the design and fabrication technology depicted in Figs. 7.26 and 7.27. Micromirrors measuring

500×500 μm achieved scan ranges of better than ±7.5° (optical) on both axes using drive voltages of 133V (inner axis) and 200V (outer axis). These results are for quasi-static operation. If the mirrors are driven at their resonance frequencies (3.5 KHz on the inner axis and 908 Hz on the outer axis), then the angular range can be extended, while the required voltages can be reduced by a factor of 5 or more.

7.7.2 Universal Joint Microscanner with "Terraced-Plate" Actuators

Figure 7.28 shows a universal-joint implementation of a 2-D microscanner. The universal joint is created by two connected, orthogonal torsion bars that together allow the mirror to tilt in any direction. The scanning mirror is electrostatically actuated by four substrate electrodes; one in each quadrant of the scanning mirror. The substrate electrodes, together with the mirror itself that acts as the counter electrode, sets up torques to tilt the mirror around the universal joint. The substrate electrodes are tapered or "terraced" to minimize the separation between the substrate and the mirror that acts as the counter electrode.

A very useful feature of this design is that the actuators and the universal joint are all placed underneath the scanning mirror. This saves real estate on the chip and allows high-fill-factor arrays to be implemented. The microscanner of Fig. 7.28 is therefore well suited for fiber switches, described in Chapter 8, and adaptive optics and other array applications described in Chapter 9.

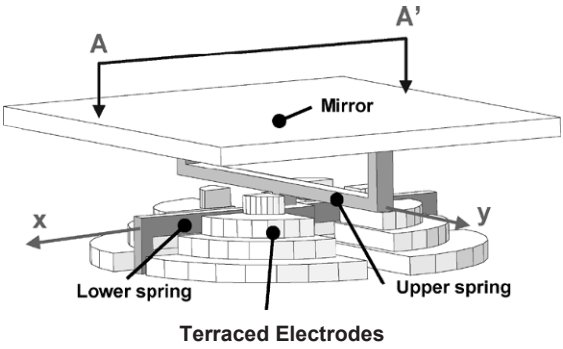


Figure 7.28 *Schematic drawing of a two-dimensional, universal joint microscanner. The universal joint is created by two connected torsion bars in orthogonal directions. The bottom electrodes are terraced to increase the actuation forces. Reprinted from [26] with permission.*

The universal-joint microscanner of Fig. 7.28 is fabricated in the SUMMiT-V™ surface-micromachining process [27] established by Sandia National Laboratories in Albuquerque, New Mexico. The scanner is constructed from five layers of polysilicon that are sequentially deposited and patterned, allowing complicated mechanical designs to be realized. Due to the relatively thin poly-silicon layers

that are used (the top layer that defines the mirror is only 2.25 um thick), the design of Fig. 7.28 is not practical for mirror sizes beyond a few hundred microns on a side. Scanners measuring 100 by 100 um achieved ± 4.4° optical scan angles with a resonance frequency of 20.7 KHz. Increasing the mirror size to 200 by 200 um, reduced the resonance frequency to 1.4 KHz.

7.7.3 Universal Joint Microscanner with Combdrive Actuators

The universal joint of Fig. 7.28 is particularly simple and well suited to the terraced electrostatic actuators that are employed. The design does not scale well to larger mirrors, however, because as the mirror size is increased, the separation between the mirror electrode and the terraced electrodes must also increase to allow the mirror to move through the required angular range. This means that there is less force available to move a bigger and bulkier mirror, which in turn means that the spring constants have to be reduced and/or the voltages increased. In practice this means that for scanning mirrors measuring 500 by 500 um or more, we have to look for alternative actuation technologies. The simplest from a fabrication and materials-compatibility point of view is to stick to electrostatic actuation, but to use combdrives instead of parallel-plate or terraced actuators. (See Appendix B for an in-depth comparison of parallel-plate and comb-drive electrostatic actuators)

Figure 7.29 shows an example of a universal-joint design using vertical combdrives. The combdrives are arranged into three "pure" rotators that each actuates one beam linage that is connected to the mirror. Each rotator can lift or lower its bending beam to impart an upward or downward force on the mirror so that it can be tilted around any axis in the plane of the mirror.

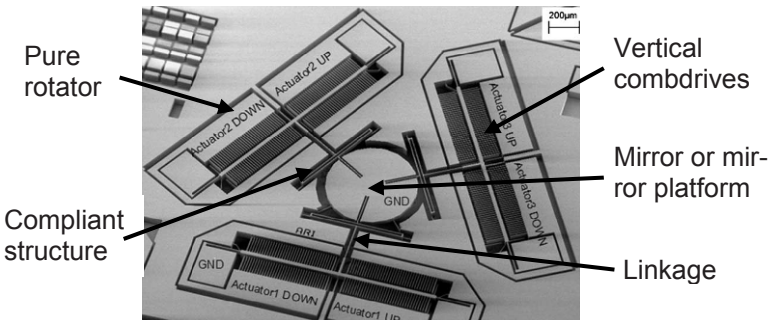


Fig. 7.29. Scanning Electron Micrograph of a universal-joint microscanner fabricated on SOI wafers by DRIE. The three rotators shown is the minimum needed for two-axis scanning. Larger numbers of rotators can of course be used, four being the most typical. Reprinted from [28] with permission.

The beam linkages have a built-in compliant structure that allows the linkage to bend sharply at a point (or more correctly, over a short distance), and thereby take up the angular difference created when the linkage forces the mirror to rotate. The compliant structure also reduces the torque that is transmitted back from the mirror to the combdrives. Together the rotators, the linkages, and the compliant structures create a universal joint that allows the mirror to rotate in two dimensions.

The scanner of Fig. 7.29 is created by DRIE of SOI wafers in a process that has much in common with the process shown in Fig. 7.27. Only a single SOI layer is needed, however, so the process is somewhat simpler than for the gimbaled scanner. The universal-joint design also has the advantage that there is no inner frame that has to be moved. This reduced the inertia on one axis of scanning and gives the universal-joint scanner a performance that is more symmetric on the two axes. Typical performance of the scanner shown in Fig. 7.29 is ± 10° optical scan angles with drive voltages of less than 150 V and a lowest resonance frequency of 1.9 KHz.

The simple fabrication and excellent optical performance of the universal-joint scanner come at a price. There are two issues, both stemming from the complicated mechanical structure. First, the mode with the lowest resonance frequency in structures driven by linkages as shown is piston motion, i.e. linear translation in and out of the plane of the chip. If this motion is wanted, this is not a problem, but if pure scanning is the objective, then the piston motion will represent an unwanted operation that must be suppressed by careful control of the actuation voltages and external influences.

The second problem with the linkage-driven scanner is that to get large angular rotation, it is important that the points of attachment to the mirror of the three linkages are not too far apart. This reduces the unobstructed area of the mirror as can be seen in Fig. 7.29. This can be solved in practice by attaching a separate, larger mirror on a pedestal to the central mirror platform, but this complicates the fabrication process and adds inertia to the scanner.

7.8 Summary of MEMS scanners

Design of optical scanning systems is a complex problem that seems to defy generalization due to the large number of widely different application requirements. The design process can be simplified by realizing that the most pertinent specification on any scanning system is the number of spots that the scanner must be able to resolve. This specification, the number of resolvable spots, is an inherent characteristic of the scanner and cannot be increased by clever optical design. This chapter therefore starts with an in-depth discussion of scanner resolution. We find that the number of resolvable spots of an ideal scanner is given by:

$$N = \frac{\Delta\theta_{tilt} \cdot \pi \cdot \omega_0}{k \cdot \lambda} + 1 \tag{7.39}$$

where $\Delta\Theta_{tilt}$ is the optical angular range of the scanner, ω_0 is the Gaussian beam radius on the scanner, and k is a constant that is set by the relevant resolution criterion for the application in question. For a scanning display we have $k=1.18$, while other applications, e.g. fiber-optic switches, require substantially less cross talk and therefore k-values on the order of 3 or more. This is the most important result of this chapter!

Next we consider the effects of practical limitations on scanner resolution. By modeling the propagation of truncated Gaussians, we find that miniaturized scanners should be designed with apertures that are close to twice the beam radius. Systems with low contrast requirements can use slightly smaller mirrors, while systems that require high contrast need mirrors that are larger than twice the beam radius, but rarely larger than three times the beam radius. We also establish criteria for acceptable surface roughness and static curvature of scanning micromirrors, and found that these criteria are straight-forward to meet in practice. Dynamic mirror bending, on the other hand, is a serious problem that must be solved by proper mechanical design. The discussion of the optics of microscanners is rounded out by a description of the reflectivity of metal coatings used in MEMS technology and by a brief explanation of the characteristics of lens scanners.

The last part of the chapter (sections 7.5, 7.6, 7.7) is dedicated to the mechanical design of microscanners. Due to the enormous variety of mechanical designs that have been proven useful in scanning systems, this treatment is necessarily much less general and relies more on examples. One general statement that can be made about microscanners is that they are underdamped. This means that care has to be taken not to excite unwanted modes of operation, because once excited, these modes will persists for long periods due to the lack of damping. As a rule, the wanted modes of operation should have the lowest resonance frequencies, while unwanted modes should be designed to have significantly higher resonance frequencies. Through numerical simulations, we showed that it is straight-forward to design one-axis scanners such that the preferred motion has the lowest resonance, but this presents a much bigger challenge for two-axes scanners. The chapter was completed by a description of three different examples that illustrate the trade-offs in mechanical design of two axes scanning systems.

Exercises

Problem 7.1 – Laser display

Consider a laser display consisting of a circular raster scanning mirror, and a semiconductor laser producing a Gaussian beam at *600 nm* wavelength. The mirror scans 480 lines, each with 640 resolvable spots. The resolution criterion is that resolvable spots should be separated by their FWHM. The distance from the lens to the screen is *10 m*. Assume that to avoid unwanted diffraction effects, you have to make the diameter of the mirror three times larger than the Gaussian beam radius on the mirror. There are no lenses or other optics between the mirror and the screen.

a. What is the minimum mirror size?
b. What is the radius of curvature of the Gaussian beam at the mirror you found in a)?
c. Could the same set-up be used to make a green-on-black display with the same resolution?
d. How fast would the mirror have to scan? How would the displayed picture look?

Problem 7.2 – Scanning Microscope

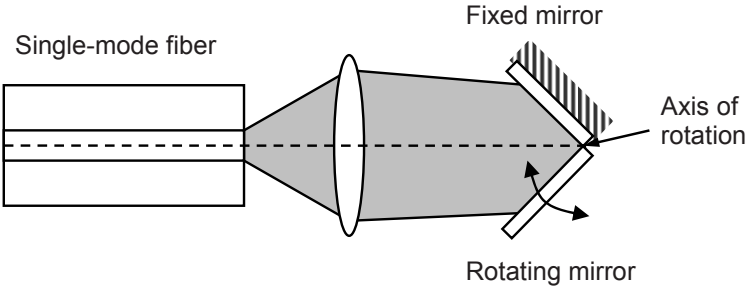
We want to build a scanning microscope that produces a focused laser spot with a Gaussian beam radius of 1 um at a wave length of *500 nm*, and we want to scan this spot over a field of view of *400 by 400 um*.

a. If we use a *500 by 500 um* mirror, then how large must the range of scan angles be to cover the field of view?
b. If we use a mirror with mechanical scan angles of $\pm 5^{o}$ on both axes, then how large must the mirror be to cover the field of view?

Problem 7.3 – Corner-Cube Modulator

a. Prove that two reflecting planes that form a 90º corner will retroreflect (i.e. send back in the direction is came from) an optical beam that is in a plane perpendicular to their intersecting line.
b. Extend the proof to a corner cube, which is a 3-D corner formed by three reflecting planes intersecting at 90º.

Consider a corner-cube modulator as shown in the figure below. One mirror is fixed and one is rotating under the control of a MEMS actuator. The optical beam from the fiber is collimated to a beam radius of *100 um* on the corner cube.



Corner-cube modulator.

c. Why do we not have to worry about the third dimension (perpendicular to the plane of the drawing) in this set up?
d. Express the back coupled light on the fiber in terms of the rotation angle of the mirror.
e. Extend the analysis to a 3-D corner cube with arbitrary angle of incidence.

Problem 7.4 – Al-Air Bragg Mirror

a. Design a Bragg mirror of alternating Al and air layers and optimize its reflectivity. Is it possible to get higher reflectivity than from bulk Al?
b. How can a mirror like this be implemented?

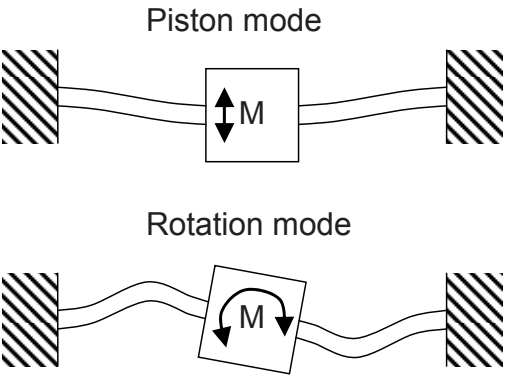
Problem 7.5 – Al-Si Mirror

It is common knowledge that mirrors are reciprocal, i.e. their reflections and transmissions are the same (except for phase) independent of which direction (back-to-font or front-to-back) the light is incident on the mirrors.

a. Test this myth by calculating the reflectance and transmittance of an Aluminum film on silicon in air as a function of aluminum thickness. In other words, calculate the reflectance and transmittance of the Aluminum film on silicon coming from the air side and from the silicon side.
b. Is it possible to use an Al film on silicon as an Anti-Reflection (AR) coating?

Problem 7.6 – Adjusting Resonance Frequencies

Consider the simple mechanical resonator in the figure below.



Vibration modes of a mechanical resonator.

a. Which one of the two resonant modes (piston and rotation) has the lowest resonance frequency? (Hint: Consider both energy storage in the springs and effective mass loading.)
b. How can you change the spring design to increase the resonance frequency of the piston mode relative to the rotation mode?

References

1 P.M. Hagelin, "Micromachined Mirrors for Raster-Scanning Displays and Optical Fiber Switches", PhD dissertation submitted to the Department of Electrical Engineering and Computer Science at the University of California, Davis, 2000.
2 J. Hagerman, "Optimum spot size for raster-scanned monochrome CRT displays", Journal of the Society of Information Displays (SID) vol. 1, no. 3, pp. 367-369, 1993.
3 S.D. Senturia, "Microsystem Design", Kluwer Academic Publishers, 2001, Chapter 9.
4 Warren C. Young, "Roark Formulas for Stress and Strain", Sixth Edition, McGraw-Hill, 1989, p. 102.
5 S.D. Senturia, "Microsystem Design", Kluwer Academic Publishers, 2001, p. 196.
6 J.T. Nee, R.A. Conant, M.R. Hart, R.S. Muller, K.Y. Lau, "Stretched-film micromirrors for improved optical flatness", IEEE Thirteenth Annual Inter-

national Conference on Micro Electro Mechanical Systems, 23-27 Jan. 2000, Miyazaki, Japan; p..704-709.
7 J.T. Nee, R.A. Conant, R.S. Muller, K.Y. Lau, “Lightweight, optically flat micromirrors for fast beam steering”, 2000 IEEE/LEOS International Conference on Optical MEMS, 21-24 Aug. 2000, Kauai, HI, USA; pp. 9-10.
8 W. Liu, J.J. Talghader, “Current-controlled curvature of coated micromirrors”, Optics Letters, 1 June 2003, vol.28, no.11, pp.932-934.
9 W. Liu, J.J. Talghader, “Thermally invariant dielectric coatings for micromirrors”, Applied Optics; 1 June 2002; vol.41, no.16, pp. 3285-3293.
10 E.D. Palik, "Handbook of Optical Constants of Solids", Academic Press, Inc., Orlando, 1985.
11 D.M. Burns, V.M. Bright, “Optical power induced damage to microelectromechanical mirrors”, Sensors and Actuators A (Physical), vol. A70, no. 1-2, 1 October 1998, pp. 6-14.
12 L.M. Phinney, O.B. Spahn, C.C Wong, “Experimental and computational study on laser heating of surface micromachined cantilevers”, Proceedings of the SPIE - The International Society for Optical Engineering, vol.6111, p.611108-1-7, 21 Jan. 2006.
13 C. Liu, “Foundations of MEMS”, Prentice Hall, 2006.
14 G.T.A. Kovacs, “Micromachined Transducers Sourcebook”, McGraw-Hill, 1998.
15 N. Maluf, “An Introduction to Microelectromechanical Systems Engineering”, Artech House, 2000.
16 K.S.J. Pister, M.W. Judy, S.R. Burgett, R.S. Fearing, “Microfabricated hinges”, Sensors and Actuators A (Physical), vol. 33, no. 3, June 1992, pp. 249-256.
17 L.Y. Lin, S.S. Lee, K.S.J. Pister, M.C. Wu, “Micro-machined three-dimensional micro-optics for integrated free-space optical system”, IEEE Photonics Technology Letters, vol.6, no.12, December 1994, pp.1445-1447.
18 M. Daneman, O. Solgaard, N.C. Tien, K.Y. Lau, R.S. Muller, "Laser-to-fiber Coupling Module Using a Micromachined Alignment Mirror”, IEEE Photonics Technology Letters, vol. 8, no. 3, March 1996, pp. 396-398.
19 S.S. Lee, L.Y. Lin, M.C. Wu, “Surface-micromachined free-space fibre-optic switches”, Electronics Letters, vol. 31, no. 17, 17 August 1995, pp.1481-1482.
20 P.M. Hagelin, U. Krishnamoorthy, J.P. Heritage, O. Solgaard, "Scalable Optical Cross-Connect Switch Using Micromachined Mirrors", IEEE Photonics Technology Letters, vol. 12, no. 7, July 2000, pp. 882-885.
21 W.C. Young, “Roark’s Formulas for Stress and Strain”, 6th ed, McGraw-Hill, 1989, p. 100.
22 F.P Beer, E.R. Johnston, J.T. DeWolf, “Mechanics of Materials”, 4th ed., McGraw-Hill, 2005.
23 K.E. Petersen, “Silicon as a Mechanical Material”, Proceedings of the IEEE, May 1982, vol.70, no.5, p.420-457.

24 S.D. Senturia, “Microsystem Design”, Kluwer Academic Publishers, 2001, Chapter 13.
25 D. Lee, O. Solgaard “Two-Axis Gimbaled Microscanner in Double SOI Layers Actuated by Self-Aligned Vertical Electrostatic Combdrive”, Proceedings of the Solid-State Sensor and Actuator Workshop, pp. 352-355, Hilton Head, South Carolina, June 6-10, 2004.
26 J.-C. Tsai, M.C. Wu, “Gimbal-Less MEMS Two-Axis Optical Scanner Array With High Fill-Factor”, IEEE Journal of Selected Topics in Quantum Electronics, vol. 14, no. 6, December 2005, pp. 1323-1328.
27 Online - http://mems.sandia.gov/tech-info/summit-v.html
28 V. Milanovic, G.A. Matus, D.T. McCormick, “Gimbal-Less Monolithic Silicon Actuators for Tip–Tilt–Piston Micromirror Applications”, IEEE Journal of Selected Topics in Quantum Electronics, vol. 10, no. 3, May/June 2004, pp. 462-471.

8: Optical MEMS Fiber Switches

8.1 Introduction to MEMS Fiber Switches

MEMS technology is close to ideal for implementation of optical fiber switches. The most significant reason for this is that MEMS can be scaled to sizes smaller than the typical length scales of the standard Single Mode Fiber (SMF) with a physical diameter of 125 um, a mode diameter of about 10 um, and a collimated mode size that can be orders of magnitude larger in highly functional switches. So MEMS offer the opportunity to create miniaturized systems where the overall size is limited by fundamental principles, e.g. diffraction, rather than by the bulk of the mechanical components. This miniaturization in turn leads to compact systems that are stable, robust, and inexpensive to package, install, and operate.

The parallel-processing fabrication paradigm that MEMS share with ICs is important for fiber switches in two ways; First, fiber optics is ubiquitous and standardized, so there is the potential for large scale production of simple, low-cost components. Second, large multifunctional fiber switches require well-matched and well-aligned optics and switching elements arranged in large arrays that cannot be cost-effectively fabricated and packaged using serial processing. From a fundamental point of view, such large scale switches are enabled by the ability of scanning micromirrors to spatially separate large numbers of channels, as we have seen in the previous chapter.

In this chapter, we will study a hierarchy of four MEMS fiber switches, the 2 by Matrix Switch, the N by N Matrix Switch, the Planar Beam Steering Switch, and finally the 3-D Beam Steering Switch. Miniaturization is important, so for each of these we will ask the questions: How small can we make the switching mirrors? and How small can we make the overall system? To answer these questions we use Gaussian beams theory to model the optical propagation through the switches and thereby find the scaling laws for the different kinds of fiber switches. We will find that the MEMS fiber switches that we are considering offer increasing fiber-port counts, but also increasing complexity, as we progress through the hierarchy.

8.2 Fiber Optical Switches and Cross Connects

Optical fibers are excellent carriers of information over long distances due to their low loss, immunity to disturbances, and fidelity of transfer. These properties are consequences of the confinement of the optical mode to the core of the fiber, i.e. the optical field is buried deep within a protective tube of glass. These same properties make it hard to manipulate guided optical signals while they are still in the waveguide or fiber. Fiber networks therefore consist of a combination of waveguide and free-space optical devices as shown in Fig. 8.1.

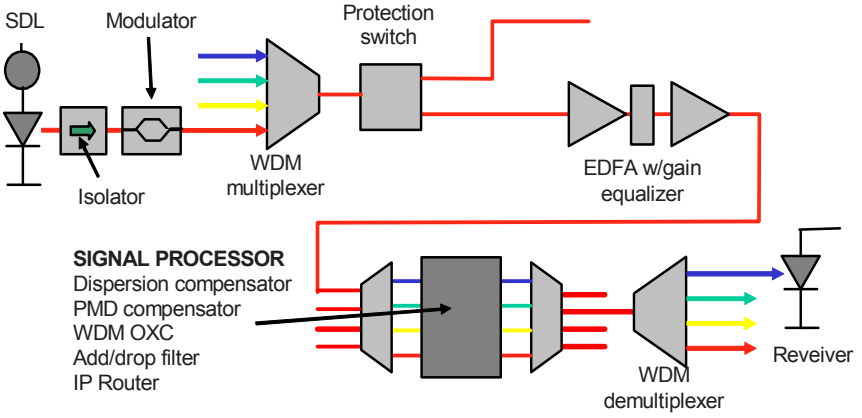


Figure 8.1. *In a typical fiber-optic network, the signals are converted from the electrical to the optical domain by direct modulation of a Semiconductor Diode Laser (SDL) or by an external modulator. The optical signal is then multiplexed with other wavelengths, and transmitted to its destination where it is converted to the electrical domain by a photodetector in a receiver circuit. The transmission typically involves long lengths of optical fibers, several gain stages, dispersion compensators, and switches of various types. We make a distinction between switches, in which the signal stays in the optical domain throughout (transparent switches) and those that involve opto-electronic conversion.*

The traditional way to configure an optical communication network is to use optical waveguides for point-to-point signal transfer, and convert to the electrical domain for switching and signal processing. Figure 8.1 shows a typical signal path through a fiber-optical network. The signal is converted from electronic to optical form in a semiconductor laser or external modulator[a]. Several wavelength chan-

[a] A directly modulated SDL introduces extra phase modulation, or chirp, due to the charge-density induced changes in refractive index in the cavity. This extra bandwidth leads to extra dispersion so direct modulation is typically used in

nels are then multiplexed together on one fiber, and transported over long fiber links, which include Erbium Doped Fiber Amplifiers. Dispersion in the fibers and free-space components necessitates the use of dispersion compensators.

The signal typically goes through several switches before it reaches its destination and is converted back into an electronic signal. There are many types of fiber-switch technologies, but we can generally classify them as either all-optical switches, in which the signal is not converted to electronic form, or electronic switches, in which all switching operations are done on electronic signals. The advantages of electronic switching is that the signal can be fully regenerated so that it is identical to the original signal that was sent (provided that no errors were made in the binary decision circuitry), or changed in some prescribed manner. In the electronic domain the signal can be re-leveled through amplification, re-timed through buffering, re-shaped by signal processing, spatially re-positioned by switching, and spectrally re-positioned by using a laser at a different wavelength in the electro-optic conversion. In other words, the functions indicated in the signal-processing block (dispersion compensation, polarization-mode-dispersion compensation, Wavelength Division Multiplexed cross connection, add/drop filtering, signal routing) can all be done in the electronic domain.

Conversion from the optical domain to the electrical and back again requires a photo receiver and a laser the associated modulator and driving circuitry. The expense of such electro-optic conversion provides powerful motivation for all-optical solutions. Optical re-leveling with Erbium-Doped Fiber Amplifiers revolutionized optical communications by making Wavelength-Division-Multiplexing practical. Similarly all-optical switching allows more flexible and cost effective systems.

The usefulness of all-optical switching has lead to the development of a large number of technologies, all with their own strengths and weaknesses. As we will see in this chapter, Optical MEMS has several unique characteristics, including low-insertion loss, low polarization dependence, high wavelength range, low cross talk, low power consumption, low cost, small size, and superior scaling to high port counts.

The main disadvantage of optical MEMS fiber switches is their speed. MEMS switches are fast enough for provisioning and restoration of optical communication networks, but they are not fast enough for optical packet switching, which require switching speed on the order of nanoseconds to tens of nanoseconds. To operate at these speeds typically electrooptic waveguide switches are needed. This type of switch is expensive and difficult to scale to large port numbers. MEMS switches are therefore the leading candidates for protection switching and circuit

---

lower capacity systems, while external modulation is used when it is important to optimize the bit rate.

switching where switching transitions on the order of tens of microseconds is sufficient.

8.3 MEMS Switch Architectures

Bulk mechanical fiber switches have been around since the 1960's when fiber optic communications first became practical. In these types of devices, the input fiber is moved mechanically into alignment with the correct output fiber. A typical design uses a turntable to accurately move the input fiber as shown in Fig. 8.2. These devices are little more than automated patch-panels, and they are too slow, too fragile and too expensive for application that require the integration of large numbers of switches. These switches are used mostly in fiber management.

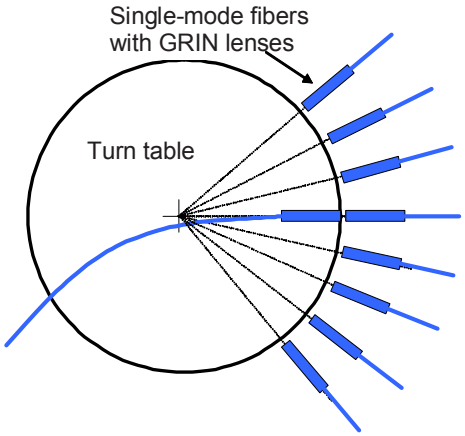


Figure 8.2. *1 by 8 bulk fiber optic switch. The turntable, which is carrying the input fiber, is moved by stepper motors such that the correct output fiber can be selected.*

MEMS technology is the obvious choice for miniaturizing mechanical optical switches, but the turn table of Fig. 8.2 does not lend itself to MEMS implementations. Instead, two types of switch architectures based on micromirrors have emerged; the Matrix Switch and the Beam-Steering Switch. The Matrix Switch is also called the 2-dimensional switch because the input fibers and output fibers are typically (but not necessarily) confined to a plane. Likewise, the Beam Steering Switch is called the 3-dimensional switch because the fibers typically (but not necessarily) fill a volume as opposed to being confined to a plane.

The Matrix Switch is illustrated in Fig. 8.3. In its simplest implementation, it is a 2 by 2 Switch that requires only a single MEMS mirror as shown in Fig. 8.3a. When the mirror is in the quiescent position (shown solid), the light of Input Fiber

1 is coupled to Output Fiber 1, while Input Fiber 2 is coupled to Output Fiber 2. With the mirror in the actuated position, the light of Input Fiber 1 is coupled to Output Fiber 2, and Input Fiber 2 is coupled to Output Fiber 1. This type of 2 by 2 cross-coupling functionality can be implemented in a large number of different technologies (see for example the champagne switch described in Chapter 3).

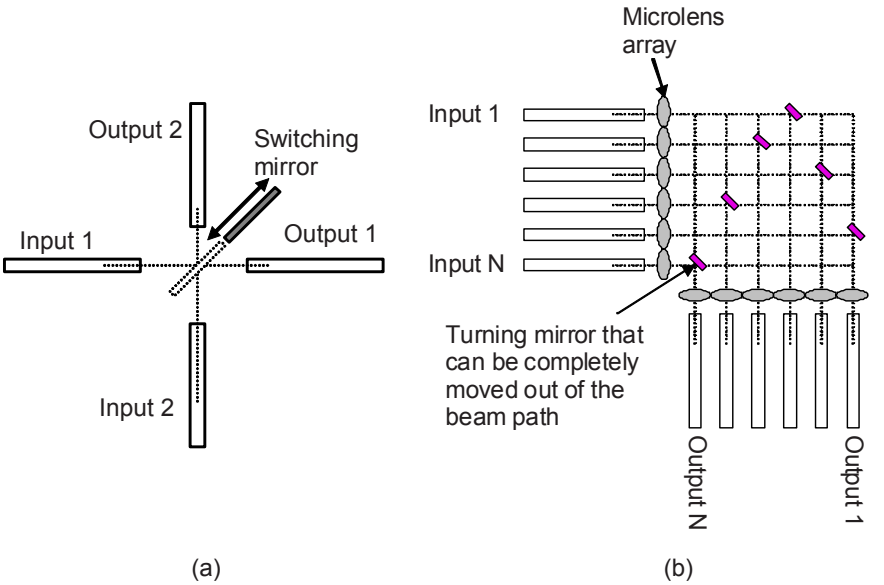


Figure 8.3. *Matrix Switches works by moving mirrors in and out of the optical beams. The 2 by 2 Switch is very compact using only a single mirror, while the N by N switch requires N^2 mirrors and a large footprint due to diffraction.*

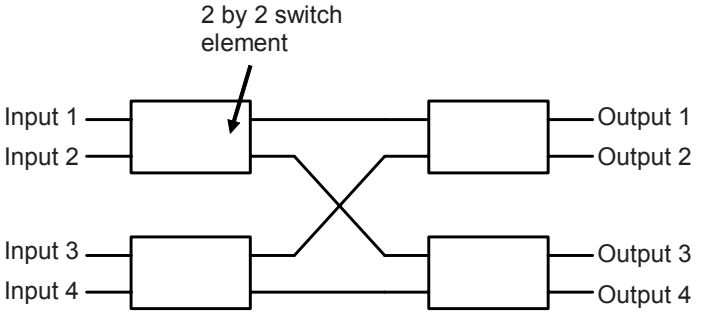


Figure 8.4. *A generic layout for cascading 2 by 2 Switches into larger cross connects. High numbers of 2 by 2 Switch points are needed for large switch matrices.*

Many switch technologies requires that larger scale switches be constructed by cascading 2 by 2 Switches as shown in Fig. 8.4, while the simplicity and low loss of the MEMS implementations allow scaling to *N* by *N* ports by using a total of N^2 mirror as shown in Fig. 8.3b. Here the input beams are collimated and they propagate down a row of turning mirrors, of which only one is activated. Likewise, the output fibers are receiving collimated beams from a column of mirrors, of which only one is activated. The net effect is that each input fiber can be connected to any output fiber, with the only restrictions on connectivity being that two input fibers cannot be coupled to the same output, and that two outputs cannot receive light from the same input.

Except for the *2 by 2 Switch*, which has only a single mirror, an *N by N* Matrix Switch has N^2 mirrors. The reason for the exception is that the *2 by 2* MEMS fiber switch uses both surfaces of the turning mirror to deflect the incoming beams. A similar arrangement using both sides of the mirror can be used for switches with larger port counts ($N>2$) as described in section 8.5, but for port counts higher than *2 by 2*, it doesn't achieve full connectivity.

The N^2 mirrors of a *N* by *N* Matrix Switch has a total of $2N^2$ mirror states that combine to 2^{N^2} switch configurations. Only *N!* are required to make all possible one-to-one connections between *N* inputs and *N* outputs, so the vast majority of the switch configurations are superfluous. The useful configurations are the ones that have *N* activated mirrors such that each row and column have only one active mirror.

The distances between the input and output fibers increase with the ports count in the *N* by *N* switch. This means that the beam diameter in the switch matrix also must grow with *N*. Each of the N^2 mirrors of the *N* by *N* Matrix Switch must have two states; one passive state where the mirror is not intersecting any of the input beams, and one active state where the mirror is positioned at the intersection of the optical axes of one input fiber and one output fiber. The requirement that the switching mirrors must be completely removed from the optical beams in their passive states is a complication, because it means that the mirrors have to be displaced by a distance comparable to the beam diameter.

As we will see in the detail analysis below, the *N* by *N* Matrix Switch requires very large foot prints as *N* increases, and it is also a problem that the required number of mirrors increases as the square of the fiber-port number. The Beam Steering Switch shown in Fig. 8.5 does not suffer from these shortcomings. It requires less space and only *2N* mirrors to support *N* by *N* switching. Each input fiber is connected to a specific output fiber through two mirrors. The first mirror directs the input light to the correct out put mirror, which in turn is position to accept the light from the chosen input and direct it to the output fiber.

The beam-steering switch only requires *2N* mirrors for a *N* by *N* switch, but each mirror must have *N* resolvable positions, for a total of $2N^2$ possible mirror states. These states combine to 2^{N^2} switch configurations, of which only *N!* are useful, just as for the *N* by *N* Matrix Switch. The useful configurations are those in which each input mirror points to a unique output mirror (i.e. no two input mirrors point to the same output mirror), and each output mirror points to a unique input mirror.

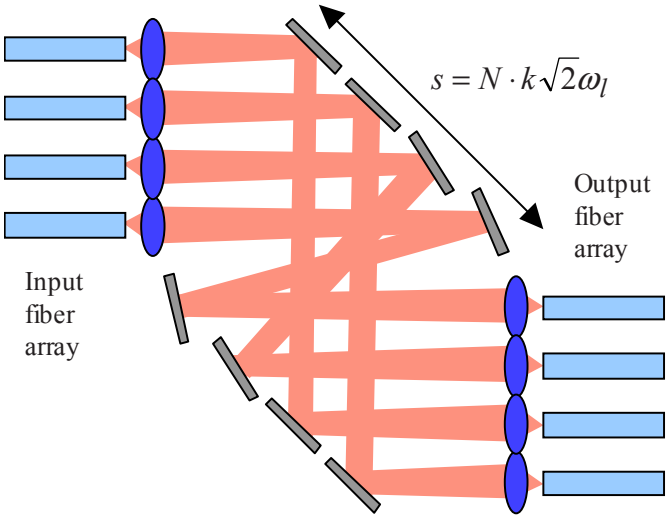


Figure 8.5. *Beam steering optical switch. Each input channel is incident on a mirror of the input array. The input mirrors direct the optical beams to the correct mirror in output array. 2N mirrors are required for N by N switches.*

Figure 8.5 shows all fibers in a single plane, but clearly the input and output fibers can be arranged in 3-dimensional volumes, and each input can still be coupled to each output, provided that each of the steering mirrors can rotate around two axes. One of the advantages of the Beam Steering Switch over the Matrix Switch is that each mirror always stays in the beam path. The required motion is a simple rotation, as opposed to the translation required in the Matrix Switch. Equally important is the fact that the Beam Steering Switch scales better, because it requires fewer mirrors and less distance between the input and output fibers. This translates into a smaller switch footprint as we will see in the analysis below.

Micromirrors that can rotate around two axes are significantly more complex than single axis mirrors as we saw in Chapter 7. A particularly attractive switch architecture is therefore the WDM optical cross connect (WDM OXC) shown in Fig. 8.6. The WDM OXC has *N* inputs, each with *M* WDM channels. (To keep the drawing simple, both N and M are 3 in Fig. 8.6.) The WDM channels are demul-

tiplexed to create a total of NxM spatially separate cannels. Each of the *N* sets of *M* WDM channels at the same center wavelength is then spatially switched into *N* outputs in *M* separate Beam Steering Switches; one for each wavelength. Finally, the *M* WDM channels on output 1 of each beam-steering switch are multiplexed onto Output Fiber 1. Likewise the *M* WDM channels on output 2 of each beam-steering switch are multiplexed onto Output Fiber 2 and so on.

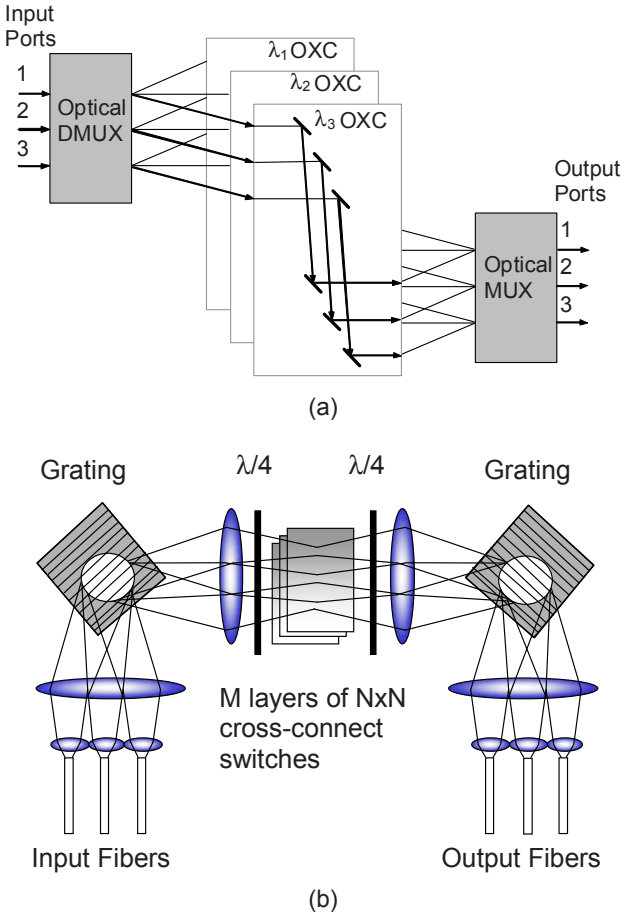


Figure 8.6. *Schematic of WDM Switch Architecture (a). Optical multiplexers (MUX/DEMUX) spatially separate the input optical beams by wavelength. Each wavelength channel is routed to an independent NxN cross-connect (shown are three switches with N = 3) or optical add-drop multiplexer. Figure (b) shows a possible implementation using gratings as free-space optical multiplexers.*

The result is that each input WDM channel can be switched to any output fiber. This is the desired functionality for WDM OXCs, and it is realized by a set of *M* simple *N* by *N* Beam Steering Switches that only require single-axis rotation. The sets of *NxM* input mirrors can be implemented as a mirror array on a silicon chip. The same is true for the output mirrors, so the MEMS mirrors require only two chips. The only switching that is not supported by this compact WDM OXC is conversion of an input wavelength channel to another wavelength on the output. This function requires a wavelength shift, which in practice means electro-optic conversion.

8.4 2 by 2 Matrix Switch

8.4.1 Fiber Separation in 2 by 2 MEMS Switches

The MEMS implementation of the *2 by 2* Matrix Switch has several unique advantages, starting with the fact that both sides of the micromirror can be simultaneously used to switch two input beams. The geometry of the switch allows for a very compact layout with few components and simple alignment. MEMS actuators also enable binary operation with no quiescent, only transient, power consumption.

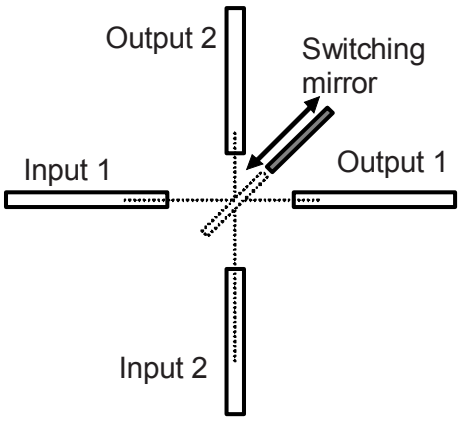


Figure 8.7. In their simplest implementation, Matrix Switches consist of no optical components other than the fibers and the switching mirror. Provided that the propagation distance is kept sufficiently short, lenses or other collimating optics are not needed.

This simplicity is most compelling feature of the 2 by 2 MEMS switch. In its most basic form, the switch is constructed from only the input and output fibers

plus the switching mirror. The question is if this simple geometry is viable, or if lenses must be used to make the diffraction losses between the fibers acceptable.

To answer this question and find the acceptable fiber-to-fiber separation in the 2 by 2 MEMS fiber switch, we use the formula we derived for fiber-to-fiber transmission in Chapter 6. Simplified for the case of transmission between perfectly-aligned, identical fibers separated by a distance D, the expression for the power coupling is

$$T = \frac{4\dfrac{\omega_{fiber}^2}{\omega_D^2}}{\left(1+\dfrac{\omega_{fiber}^2}{\omega_D^2}\right)^2 + \dfrac{k^2\omega_{fiber}^4}{4R_D^2}} \tag{8.1}$$

where the beam radius and the radius of curvature are given by

$$R_D = D\left[1+\left(\frac{\pi\omega_{fiber}^2}{\lambda D}\right)^2\right] \tag{8.2}$$

$$\omega_D = \omega_{fiber}\left[1+\left(\frac{\lambda D}{\pi\omega_{fiber}^2}\right)^2\right]^{\frac{1}{2}} \tag{8.3}$$

The solid line in Fig. 4.8 shows the optical power transmission as a function of separation for fibers with mode radii of 4.8 um at 1.55 um wavelength (Standard Single Mode Fiber). The dashed line shows the power transmission that would result if we set $\frac{1}{R_D} = 0$ in the above formula so that the transmission can be expressed

$$T = \frac{4\omega_{fiber}^2\omega_D^2}{\left(\omega_D^2+\omega_{fiber}^2\right)^2} \tag{8.4}$$

We recognize this as the formula for coupling between two Gaussian beams that are perfectly aligned and have flat wave fronts, but different mode radii. In other words, the dashed line shows the transmission we would get if the coupling into the fiber depended only on the size mismatch of the modes, but not the phase curvature.

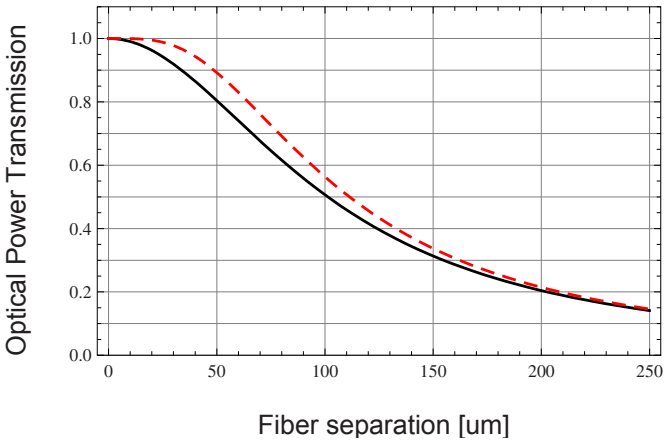


Figure 8.8 *Fiber-to-fiber transmission without collimation optics. The fibers are perfectly aligned Single Mode Fibers (Mode radius: 4.77 um, wavelength: 1.55 um, Rayleigh length: 46.7 um). The solid line shows the total loss, while the dashed line shows the loss contribution due to size mismatch. As the fiber separation increases, the size-mismatch contribution dominates over the phase-curvature contribution, and the two graphs asymptotically approach each other.*

As a practical matter the fibers have to be separated by more than their diameters in the 2 by 2 Switch geometry. The diameter of SMF is 125 um, so we see from Fig. 8.8 that the diffraction losses between the fibers in the 2 by 2 Switch will limit the transmission to less than 0.35. Comparing the solid and dashed lines of Fig. 8.8, we also see that size mismatch is dominant at these fiber separations and that the phase front curvature contributes only a minor part of the transmission losses[b]. We will use this fact to simplify our treatment of the transmission losses caused by the finite thickness of the switching mirror.

8.4.2 Mirror Thickness in 2 by 2 Matrix Switches

The diffraction loss due to the separation of the fibers is the only fundamental losses of the 2 by 2 MEMS fiber switch when the mirror is in its quiescent state (i.e. the mirror is not in the paths of the optical beams, and input 1 is coupled to output 1 and input 2 is coupled to output 2). When the mirror is actuated (i.e. the mirror is moved into the paths of the optical beams, and input 1 is coupled to output 2 and input 2 is coupled to output 1), then the finite thickness of the mirror will create an offset of the transmitted beams on either output 1 or output 2 or

[b] Note that the biggest difference between the solid and dashed lines of Fig. 8.8 is at the Rayleigh length of 47 um. That is expected, because that is where the radius of curvature has its smallest value.

both. If we want to minimize loss in both connections, then we place mirror such that its center plane is at the point of intersection between the two optical axes. Each output beam is then offset laterally by $d/\sqrt{2}$, where d is the mirror thickness.

This lateral offset between the axis of fiber and the axis Gaussian output beam reduces the transmission through the switch in the actuated state. In Chapter 6 we found that the coupling between laterally offset Gaussian beams is given by:

$$T = \frac{4\omega_{fiber}^2 \omega_D^2}{\left(\omega_D^2 + \omega_{fiber}^2\right)^2} e^{-\frac{2d^2}{\omega_D^2 + \omega_{fiber}^2}} \tag{8.5}$$

where ω_{fiber} is the fiber mode radius, ω_D is the Gaussian beam radius at the output fiber, and d is the lateral offset. This formula is, strictly speaking, only valid for the situation where the output beam is focused on the output fiber, i.e. the output beam has no curvature at the output fiber $\left(1/R_D = 0\right)$. However, we know from our treatment of the propagation losses in the 2 by 2 Switch that the wave front curvature plays a minor role in the transmission losses in practical 2 by 2 MEMS switches, so we can ignore the finite radius of curvature in rough calculations.

The above formula is used to plot the transmission as a function of mirror thickness for three different cases. The results are shown in Fig. 8.9. Again the input and output fibers are standard single mode fibers with a mode radius 4.8 um at 1.55 um wavelength. The solid line shows the dependence on offset for an output beam with a mode radius that matches that of the fiber. That match would require collimation and focusing lenses, which would change the dependence on the lateral offset, so the solid line is shown as a reference, not as a model for a practical 2 by 2 Switch design.

The short-dashed curve shows the transmission for an output mode radius of 10 um. That corresponds to a fiber separation of just under 100 um. To make this fiber separation practical would require thinning down the fiber diameter to be able to position the fibers closer.

The long-dashed curve shows the transmission for a fiber separation of 140 um. This is a practical separation that leaves enough room for the actuation mechanism of the switching mirror without having to modify the fibers. The problem with this design is of course that the mode-mismatch loss is large. Even with no lateral offset, the transmission is less than 0.35. The good news is that the sensitivity to mirror thickness is less than for the smaller beam radii. We see that mirror thicknesses of less than 5 um have little effect on the transmission, and that thicknesses in the 5 to 10 um range probably would be acceptable for many applications.

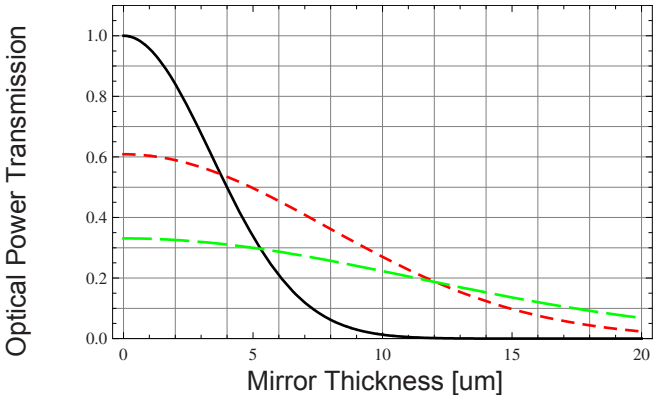


Figure 8.9 *Fiber-to-fiber optical power transmission as a function of mirror thickness (that causes a lateral offset that is $1/\sqrt{2}$ of the mirror thickness) in the actuated state of 2 by 2 MEMS fiber switches. The solid line is for an output mode radius of 4.8 um, corresponding to an impractical fiber separation of zero length. The short-dashed line is for an output mode radius of 10 um, corresponding to a fiber separation of just under 100 um, and the long-dashed line is for an output mode radius of 15.2 um, corresponding to a practical fiber separation of 140 um.*

Figure 8.9 makes it clear that the mirror in a 2 by 2 fiber switch must be very thin. High-reflectivity mirrors that are less than 5 um are difficult to fabricate with traditional manufacturing techniques. In MEMS technology such mirrors can be made by surface micromachining, by Deep Reactive Ion Etching, or by anisotropic etching of silicon. One of the first commercial uses of MEMS technology in fiber switches was indeed to create very thin mirrors defined by <111> crystalline planes in Silicon by anisotropic etching.

8.4.3 Low-loss 2 by 2 Matrix Switches

A power transmission of 0.35, or equivalently, a loss of 4.6 dB, is acceptable in many applications, so the simple 2 by 2 MEMS fiber switch without collimating optics is indeed practical and used in many systems. The losses are large enough, however, that there is strong motivation for improved designs. The obvious thing to do is to use a lens for each fiber so that the beams are collimated and diffraction losses can be reduced to insignificant levels. The beam size at the mirror will then also be larger so that thicker mirror can be used without undue increase of the transmission loss[c]. The problem with this straightforward solution is of course the

[c] Note that according to the principles derived in Chapter 2, the calculation of transmission loss due to lateral offset should be done using the larger mode ra-

extra complexity, chip area, and expense of adding the collimating lenses to the design.

More cost effective solutions include thinning down the fiber cladding, as suggested above, to be able to position the fiber facets closer, and expanding the mode size of the fibers. It is possible to use thermally-expanded-core fiber or in-line GRIN lenses to increase the mode size at the output of optical fibers. A relatively modest mode increase of a factor of three, well within the capability of TEC fibers and GRIN lenses, has a dramatic effect on transmission. This is shown in Fig. 8.10, where we plot the transmission as a function of fiber separation for a mode radius of 15 um. We see that for transmission distances below 200 um, the transmission is better than 0.95. In practice this level of loss is negligible due to the inescapable and much larger alignment losses.

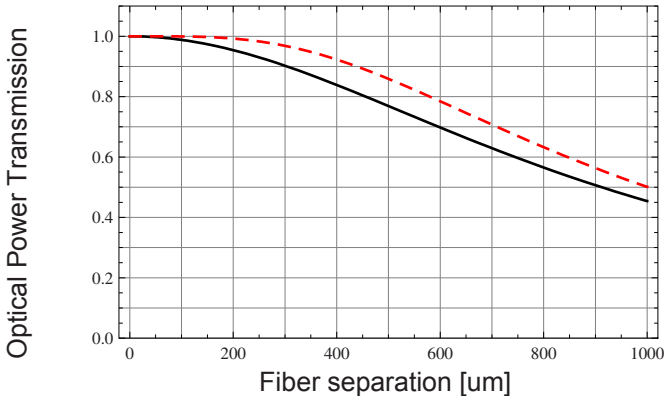


Figure 8.10 *Fiber-to-fiber transmission with an expanded mode radius of 15 um at 1.55 um wavelength (Rayleigh length: 456 um). As in Fig. 8.8, the fibers are perfectly aligned Single Mode Fiber at 1.55 um wavelength without collimation optics. The solid line shows the total loss, while the dashed line shows the loss contribution due to size mismatch.*

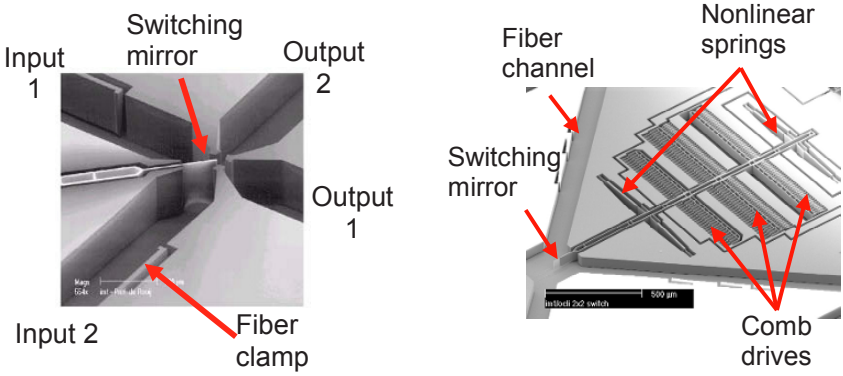
8.4.4 MEMS Implementation of 2 by 2 Fiber Switch

There have been many published reports on MEMS 2 by 2 fiber switches. The implementations are based on a variety of bulk and surface micromachining fabrication techniques. One technology that is both simple and take full advantage of all the attributes of MEMS is Deep Reactive Ion Etching (DRIE) of Silicon-on-

dius at the mirror, not the smaller one at the fiber. At the fiber lateral offset is reduced by the collimating lens.

insulator wafers. Two examples of DRIE-of-SOI switches are shown in Fig. 8.11. In both cases these switches are fabricated by a single DRIE process that defines the fiber channels, the switching mirror, and the electrostatic combdrives. The flexibility of DRIE makes it possible to create all these functions in a single mask.

Figure 8.11a shows a close up of the switching mirror at the cross-point of the fiber channels [1]. Note that the fiber channels have clamps to hold the fibers in place once they are positioned. The fiber channels are tapered towards the switching plane so that fibers with tapered claddings can be aligned both transversally and axially. The purpose of tapering the fiber ends is to reduce diffraction loss by reducing the separation between the input and output fibers.



..... a b

Figure 8.11 *2 by 2 MEMS switches shown without the fibers to highlight the MEMS design. The close-up in (a) shows the switching mirror in the fiber channels at the fiber-axes cross point [1], while (b) shows the design of the electrostatic comb drive that moves the mirror between the two switch states [2]. The mechanically non-linear springs gives the switch in (b) bi-stable operation. Both switches are realized by Deep Reactive Ion Etch of Silicon-on-Insulator wafers. Reprinted with permission.*

The switch shown in Fig. 8.11b has many of the same attributes as the one in (a). The switching mirror, the fiber channels and the fiber clamps are very similar. In addition, the switch in (b) has a unique bi-stable, electrostatic actuator [2]. The principle behind the bi-stability is that the springs supporting the moving parts of the actuators are designed so than when the mirror is moved, the springs are initially compressed in addition to being bent. As the mirror is moved further, the compression of the springs is reduced, and the spring becomes less stiff, leading to a second potential energy minimum (the first potential minimum is the relaxed state of the springs). There is therefore no need to apply a voltage to the actuators

to keep the switch in either of its two operational states. Voltage is only required when switching from one state to the other.

The biggest fabrication difficulty of the DRIE-of-SOI, 2-by-2 Switch is to create a good mirror. The mirror is made of silicon and must be metalized to provide good reflectivity. The bigger problem is to control the DRIE such that the etched surfaces are sufficiently smooth. In Chapter 7.2.4 we argue that scanning micromirrors should have an RMS surface roughness of better than $\lambda/20$ or 75 nm for fiber-optic communication wavelengths. It is a challenge to produce mirrors of this quality on surfaces defined by DRIE.

8.5 N by N Matrix Switches

The 2 by 2 Switch is readily extended to larger port counts. The basic layout of a 6 by 6 switch is shown in Fig. 8.12. In principle, this design can be extended to any number of input and output fibers. Once the switch is extended beyond 2 by2, we must in practice use lenses or other collimating optics to avoid excessive diffraction losses. This complicates the switch fabrication, but has the advantage that losses can be kept small even for large numbers of input and output fibers.

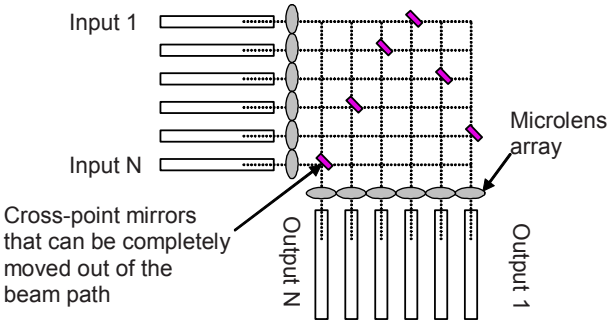


Figure 8.12. The basic N by N Matrix Switch works by moving mirrors in and out of the optical beams. There are N^2 mirrors, but only N of them are activated at any given time. With collimating optics to avoid diffraction losses, this design can in principle be extended to large numbers of fibers.

One of the problems with the switch of Fig. 8.12 is that the path length through the switch from a given input fiber to an output fiber will depend on which output fiber that is chosen. This can be avoided by offsetting the input and output fibers as shown in Fig. 8.13. In this switch, which we will call the constant-coupling architecture, the path lengths are always the same for all connections, irrespective of which switching mirrors that are activated.

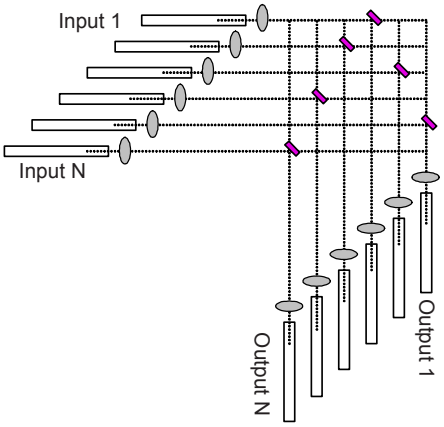


Figure 8.13. The constant-coupling, Matrix Switch have offset input and output fiber arrays so that any connection though the switch has the same path length. That means that all paths can be set up to have minimum diffraction losses.

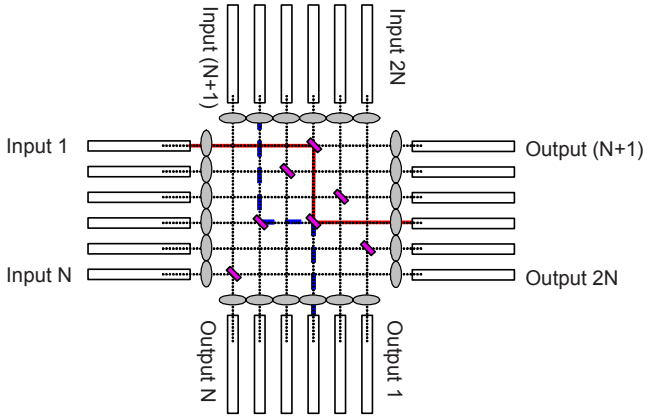


Figure 8.14. Matrix Switch with N^2 mirror that support switching of 2N input fibers into 2N output fibers. Two connections are shown as solid and dashed lines. Between N and 2N mirrors are activated for each state of the switch. The added functionality is achieved by using both sides of each switching mirror, just as in the case of a 2 by 2 Switch, but unlike the 2 by 2, the N by N does not have full connectivity.

The lower losses due to the constant path lengths for all connections is a practical advantage of the constant coupling architecture, but there is a price to pay in that the area of the switch increases by a factor of 2 (if we count only the area that is actually taken up by the components of the switch) or 4 (if we count the whole

square that the switch occupies). Area is often the limitation on matrix-switch design, so the constant coupling architecture is not commonly because the uniform insertion loss in not enough of an advantage to justify the extra area.

As for the 2 by 2 Switch, it is possible also in larger Matrix Switches to use both sides of the switching mirrors. The switch can then be configured to have a total of *2N* inputs by *2N* outputs as shown in Fig. 8.14. Any input fiber can be connected to any output fibers, but there are some combinations of connections that are impossible.

8.5.1 Scaling of N by N Matrix Switch

To understand the scaling of the Matrix Switch consider Fig. 8.15. We want to find the most compact design for a *N* by *N* switch, i.e. we want to minimize the length, *s*, of the side of the switch matrix. The optical beam from each input fiber is captured by a lens and focused (collimated) to a soft focus at a distance *s* from the lens. The beam then diverges again over a distance *s* till it is captured by another lens and coupled into an output fiber.

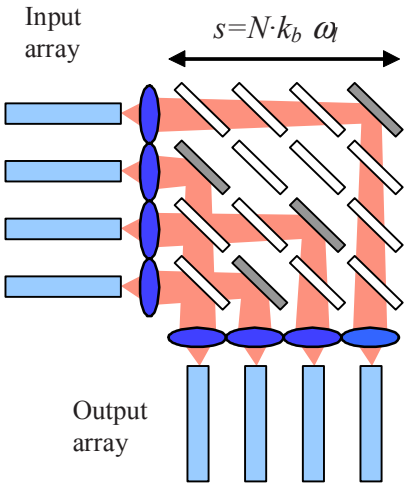


Figure 8.15. Scaling of the N by N Matrix Switch. The objective is to find the optimum beam configuration, i.e. the optimum beam size at the lens and the optimum focusing of the beam, to minimize the size of the mirrors and the size of the overall switch matrix that are required to support a given number of fiber ports.

If we assume that the beam radius is ω_l at the lens and ω_0 at the soft focus, then the length of the side of the matrix can be expressed

$$s = N \cdot k_m \omega_l = N \cdot k_m \omega_0 \left[1 + \left(\lambda s / \pi \omega_0^2 \right)^2 \right]^{1/2} \tag{8.6}$$

where N is the number of input and output fibers, and k_m is a constant that depend on how far we must separate the Gaussian beams from neighboring fibers.

The numerical value of k_m strongly affects scaling, so it must be carefully considered. There are two issues of importance; the lenses must be large enough to capture the beams and the cross talk between beams must be kept at acceptable levels. In Chapter 4 we showed that an aperture must be on the order of three times the Gaussian Beam radius to not create significant side lobes. In practice there also has to be some extra distance between the lenses, so it is reasonable to assume that the center-to-center lens spacing must be about $5\omega_0$. This lens spacing is also sufficient to ensure that the incoherent cross talk between fibers do not exceed acceptable levels, which typically is -40 dB or less for applications in telecommunications [3].

In contrast, the Gaussian approximation to the fiber mode leads to the prediction that $k_m=3$ is sufficient to ensure -40 dB cross-talk, demonstrating that the Gaussian approximation is not sufficiently accurate for cross-talk calculations, as we pointed out in Chapter 6. The reason that we cannot use the Gaussian beam approximation for cross-talk calculations is that the Gaussian beam approximation underestimates the optical power in the wings of the distribution, so that the cross talk is also underestimated. Measurements on single-mode fiber show that a separation of five times the mode radius [4,5] is required to get a cross talk of less than -40dB between channels. The conclusion is that the minimum fiber separation, as determined by the minimum acceptable lens aperture, matches the fiber separation determined by the cross-talk requirement, and that $k_m \approx 5$ is a reasonable value to use in our calculations.

To find the switch design that supports the largest number of fiber channels, we solve Eq. 8.5 for N

$$N = \frac{s}{k_m \omega_0 \left[1 + \left(\lambda s / \pi \omega_0^2 \right)^2 \right]^{1/2}} \tag{8.7}$$

and optimize over variations in ω_0.

$$\frac{dN}{d\omega_0} = 0 \Rightarrow s = \frac{\pi \omega_0^2}{\lambda} \tag{8.8}$$

Not surprisingly, we find that we get the largest number of mirrors when the total propagation length through the switch ($2s$) equals twice the Rayleigh length,

which is called the confocal parameter, of the Gaussian beam propagating through the switch. It follows that the beam radius at the lens is $\omega_l = \sqrt{2} \cdot \omega_0$.

Substituting the relationship $s = \frac{\pi \omega_0^2}{\lambda}$ into the expression for *N*, we find the maximum number of mirrors:

$$N_{\max} = \frac{s}{k_m \cdot \omega_l} = \frac{1}{k_m} \frac{\pi \omega_0}{\sqrt{2}\lambda} = \frac{1}{k_m} \sqrt{\frac{\pi s}{2\lambda}} \tag{8.9}$$

This expression shows that the Matrix Switch does not scale well to large channels numbers. As the size of the switch increases, the channel number grows only as the square root of the linear size, *s*, or in other words, the size grows as the square of the channel number.

At the fiber-optic communication wavelengths the relationships between size and channel number and beam radius and channel number can be expressed as

$$s = \frac{2N^2 \cdot k_m^2 \lambda}{\pi} = \frac{2N^2 \cdot 5^2 \cdot 1.55\mu m}{\pi} \approx N^2 \cdot 25\mu m \tag{8.10}$$

$$\omega_0 = \frac{\sqrt{2} \cdot N k_m \lambda}{\pi} = \frac{\sqrt{2} \cdot N \cdot 5 \cdot 1.55\mu m}{\pi} \approx N \cdot 3.5\mu m \tag{8.11}$$

We see that a 10 by 10 switch would have a mirror matrix that measure 2.5 by 2.5 mm, a very small chip. The mirrors have to be $\sqrt{2} \cdot 3 \approx 4.2$ times larger than the beam radius[d], so a 10 by 10 switch needs mirrors measuring about 150 um on a side, which is a manageable mirror size.

A 100 by 100 switch, on the other hand, would have to be 25 cm on a side and the beam radius is about 350 um, necessitating mirrors with diameters on the order of 1.5 mm. These values are borderline impractical for most MEMS fabrication technologies, so, although there are no hard limits, we conclude that in practice the MEMS Matrix Switch is limited to less than 100 input and output fibers.

[d] Strictly speaking it is only the mirrors close to the collimating lenses that need to be this large. Close to the beam waist, the mirrors can be smaller by a factor of $\sqrt{2}$, but this level of improvement on some mirrors will in practice typically not be worth the extra complexity of having different mirror designs in the same switch matrix.

8.5.2 MEMS Implementations of N by N Matrix Switch

Implementation of a N by N Matrix Switch requires integration of several different functions; mechanical structures for positioning of input fibers, output fibers, and lenses must be created on the same substrate where the switching mirrors are realized, and all these mechanical devices must be aligned, typically with sub-micron precision, to ensure proper operation of the switch. MEMS technology provides mechanical alignment with excellent precision and stability. The switching mirror itself, on the other hand, is a challenge because of the relatively large distance it has to be moved to bring it completely in and out of the beam path.

Anistropically etched v-grooves have emerged as a robust and cost effective technology for positioning and alignment of fibers and lenses with sub-micron precision. U-grooves formed by DRIE have similar performance and offer more flexibility in the geometrical layout and additional features, e.g. the the fiber clamps shown in Fig. 8.11. Traditional bulk micromachining and DRIE of silicon wafers therefore provide all the alignment features required for the N by N Matrix Switch.

The switching mirrors used in the N by N Matrix Switch are challenging for several reasons. First the mirrors have to be large, and they have to move over long distances to be able to capture the whole beam in the active position and leave the beam completely unobstructed in the passive position. We found that a 10 by 10 switch requires mirrors measuring 150 um on a side and that the size grows linearly with the fiber count. Even modest sized switches therefore require motion of hundreds of micron, which is difficult to achieve with electrostatic actuators (see Appendix B). The actuator design is further complicated by the fact that the N by N switch with N>2 require dense packing of the mirrors AND actuators. In the 2 by 2 Switch, the actuator is outside the area where the optical beams propagate, but that is not possible in larger arrays, so the actuators have to be tucked away in the third dimension, i.e. placed under or over the mirrors, and the actuators cannot take more space than the mirror separation allows.

The most difficult part of the design of large N by N Matrix Switches is therefore the actuator that has to provide long-distance travel and a very precise active position, while occupying a restricted area. Several different types of electrostatic MEMS inch-worm motors have been demonstrated, e.g. the scratch drive [6,7] and the vibromotor [8], but these all have problems meeting the requirements on precision, repeatability, and switching speed.

Magnetic fields are set up by currents or permanent magnets, so they do suffer from the electrode-spacing dependence that limits the forces available from electrostatic actuators. Magnetic actuators therefore represent an attractive alternative to electrostatics for driving the large mirrors of N by N Matrix Switches. Both purely magnetic drives and combinations of magnetic and electrostatic actuation have been demonstrated. In practice, this means that the fabrication process has to

incorporate magnetic materials and that the switch system has provide magnetic fields that are external to the MEMS chip. These represent cost-driving complications in fabrication and packaging, so magnetic drives are difficult to commercialize.

The difficulty in creating large, long-throw micromirrors have led to a search for other types of switching mechanisms, e.g. the Champagne switch (described in Chapter 3), that can be implemented in waveguides. Guided-wave architectures do not suffer the size restrictions caused by diffraction in free-space MEMS switches, but they still need a total of N^2 switches to create a N by N matrix. Unlike MEMS switches that do not contribute loss when they are in the inactive mode, most binary waveguide switches have insertion loss on the order of a dB in both states. This means that at most a few tens of switches can be arranged in series, thus limiting the port count in waveguide Matrix Switches to roughly the same port count as in MEMS implementations.

We conclude that the large mirror size and the associated large required motion are serious challenges to practical and reliable implementations of N by N Matrix Switch. So even though the scaling laws allow switches with several tens of fiber ports, and MEMS technology provides integration and alignment accuracy, there are serious technological obstacles that must be overcome before large N by N Matrix Switches can be successfully commercialized.

8.6 N by N Beam Steering Switches

Like all beloved children, the by Beam Steering Switch has many names. It is known as the Spanke Architecture, the 3-Dimensional MEMS Switch, the Confocal Switch, the Beam Steering Switch, and many others. We will use the name Beam Steering Switch, and we will consider two variations, the planar Beam Steering Switch and the 3-D Beam Steering Switch. We choose this name because beam-steering describes the basic operation of the active elements.

The Beam Steering Switch is essentially an application of the scanners we described in Chapter 7. As can be seen from Fig. 8.5, the optical field from each input fiber is collimated onto a dedicated input mirror that steers (or scans) the beam onto another beam-steering mirror that is dedicated to the desired output fiber. The chosen output mirror must be positioned to steer the optical beam into the output fiber. Each beam path is therefore set up by two beam-steering mirrors, or scanners.

It is clear from Fig. 8.5 and the explanation of the switch operation that any input fiber can be coupled to any output fiber, and that all combinations of one-to-one connections are possible. Just like the Matrix Switch, the Beam Steering Switch does not support one-to-many connections (broad casting) and many-to-one con-

nections (multiplexing). It is conceivable to extend the functionality of the Beam-Steering switch to include broadcast and multiplexing, but only at the cost of significant loss[e]., so we will not consider such extensions.

8.6.1 Scaling of the Beam Steering Switch

Planar Beam Steering Switch

The individual beam-steering mirrors are the workhorses of the Beam Steering Switch, and their resolving powers (number of spots they can resolve) determine the maximum number of input and output fibers that can be supported. To see how the scanning mirrors limit the channel number, consider the Beam-Steering switch of Fig. 8.16 [9]. In this layout, the input and output beams are parallel. It can be shown that relaxing this requirement, i.e. allowing the input and output beam to be non-parallel, does not increase the fiber-port count.

In the switch of Fig. 8.16, as in all Beam Steering Switches, the input mirror array directs each input optical beam to the desired output port, and the output mirror array aligns the optical beams for coupling into the output fibers. We assume that the mirror rotational axes lie in the center of each mirror, and the mirror centers are collinear in both arrays. Mirror size is proportional to the optical beam radius at the mirror, and spacing between adjacent mirrors is proportional to the size of the optical beam waist. The optical beam waist is centrally located between the mirror arrays to create a symmetric switch that minimizes mirror size and switching time.

To maximize the port count, we wish to have a large separation between the mirror arrays, while at the same time minimize optical beam divergence. From Fig. 8.16

$$\tan\alpha = \frac{d(N-1)\cos\beta}{p + d(N-1)\sin\beta} \tag{8.12}$$

where α is the optical scan angle, p is the separation between mirror arrays, d is the distance between mirror centers, and β is the angle between a line perpendicular to the mirror at zero deflection ($\alpha = 0$) and the input optical beam. In Fig. 8.16 we also introduce the parameters $s=d\cdot N$, which is the array size, and m, which is the mirror size.

[e] That broadcasting and multiplexing require loss follows directly from what we found in Chapter 2, namely that optical modes cannot be combined in loss-less, linear optical systems.

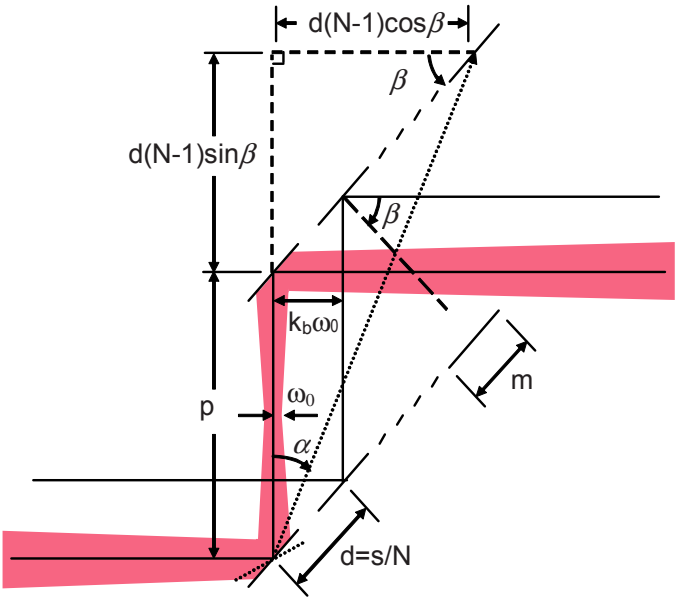


Figure 8.16. Gaussian beam propagation through a planar beam-steering switch. In the small-angle approximation, the optimum separation between the mirror arrays is the Rayleigh length Z_R. The product $k_b\omega_0$ is the separation between optical beams inside the switch matrix.

As for all optical scanners and beam-steering mirrors, the value of the optical angle is twice that of the mechanical angle. In other words, to achieve an optical deflection angle of α, the beam-steering mirror must rotate trough an angle $\alpha/2$. Note that to span the whole output array, the input mirror at the extreme ends of the input array must have an optical angular range of α in one direction, while the mirror in the middle of the array need to scan $\alpha/2$ in either direction. If the mirrors are identical and their reflective surfaces are in the same plane when not actuated, then each must be able to rotate through an optical angle of α in either direction. Clearly there are some improvements that can be made by having different mirror designs and/or having the mirrors oriented in slightly different directions in their quiescent state.

Solving the above equation for N, we find

$$d(N-1)(\cos\beta - \sin\beta\tan\alpha) = p\tan\alpha \Rightarrow N \approx 1 + \frac{p\alpha}{d\cos\beta} \tag{8.13}$$

where we have used a small-angle approximation to set $\alpha \approx \tan\alpha$ and $\cos\beta - \sin\beta\tan\alpha \approx \cos\beta$. This approximation significantly simplifies the analysis and is valid for most, if not all, practical implementations.

Now we assume that the Gaussian beam waist is positioned exactly at the midpoint between the two mirror arrays so we can make the following substitution

$$d\cos\beta = k_b \cdot \omega = k_b \cdot \omega_0 \left[1+\left(\frac{\lambda p}{2\pi\omega_o^2}\right)^2\right]^{0.5} \tag{8.14}$$

where ω_0 is the Gaussian beam radius at the mirrors, and k_b is the separation between adjacent optical ports measured in beam-waist radii. It is a constant determined by the required cross-talk between channels or by the allowable truncation of the Gaussians at the mirrors. It follows that N can be expressed

$$N = 1+\frac{p\cdot\alpha}{k_b\cdot\omega_0\left[1+\left(\frac{\lambda p}{2\pi\omega_o^2}\right)^2\right]^{0.5}} \tag{8.15}$$

We now optimize N with respect to variations in ω_0

$$\frac{dN}{d\omega_0} = 0 = \frac{-p\cdot\alpha}{(k_b\cdot\omega_0)^2\left[1+\left(\frac{\lambda p}{2\pi\omega_o^2}\right)^2\right]}\cdot$$

$$\left\{k_b\left[1+\left(\frac{\lambda p}{2\pi\omega_o^2}\right)^2\right]^{0.5} + k_b\cdot\omega_0\frac{2\frac{\lambda p}{2\pi\omega_o^2}\cdot\left(-2\frac{\lambda p}{2\pi\omega_o^3}\right)}{2\left[1+\left(\frac{\lambda p}{2\pi\omega_o^2}\right)^2\right]^{0.5}}\right\} \Rightarrow \frac{p}{2} = \frac{\pi\omega_o^2}{\lambda} \tag{8.16}$$

The corresponding optimum mirror separation and channel number are

$$d\cos\beta = k_b\cdot\omega = k_b\cdot\omega_0\sqrt{2} \tag{8.17}$$

$$N = 1+\frac{\pi\cdot\alpha\cdot\sqrt{2}\omega_0}{\lambda\cdot k_b} = 1+\frac{\pi\cdot\alpha\cdot d\cos\beta}{\lambda\cdot k_b^2} \tag{8.18}$$

We note that using the small-angle assumption and the Gaussian optical beam model, we find the optimum value of p to be the Rayleigh length of the optical beam. This is what we would expect from our earlier treatment of the Matrix

Switch and our fundamental considerations of Gaussian Beam propagation in Chapter 4.

The port count expressed in terms of the mirror size, $m=f\cdot d$, becomes

$$N = 1 + \frac{\pi \cdot \alpha \cdot m \cos\beta}{\lambda \cdot f \cdot k_b^2} \tag{8.19}$$

where f is the one-dimensional mirror fill-factor. It is also useful sometimes to express the maximum port count as a function of t, the product of the mirror radius and its maximum mechanical scan angle:

$$t = \frac{m}{2} \cdot \frac{\alpha}{2} \tag{8.20}$$

where $m/2$ is the mirror radius (or half of the length of a side length if the mirror is square).

The parameter t, called the space-bandwidth product of the mirror, tells us how much the edge of the mirror have to move to achieve a mechanical scan angle a. This important parameter is typically determined by the actuation technology used to drive the steering mirrors. With this expression for t, we find:

$$N = 1 + \frac{4\pi \cos\beta \cdot t}{\lambda \cdot f \cdot k_b^2} \tag{8.21}$$

We see that for a given wavelength, we want the space-bandwidth product (t) as large as possible, and the incident angle (β), the fill factor (f), and the beam separation (k_b) at the mirrors to be as small as possible.

Just as for the Matrix Switch, the factor k_b in the above expressions is determined by the amount of power loss and distortion we can tolerate in the beam after reflection from the beam steering mirrors. However, we also have to make sure that the beams are separated by a sufficiently large distance so that the cross talk between adjacent channels does not become excessive. We calculate the cross talk between adjacent channels at the position of the beam waist. (Remember that the cross talk doesn't change by passing through loss-less, linear optical devices). To ensure sufficiently low cross talk we require a minimum beam separation at the beam waists. We can express this as

$$d \cdot \cos\beta = k_0 \cdot \omega_0 \tag{8.22}$$

As stated in the discussion of cross talk in the Matrix Switch, measurements on SMF show that a port separation corresponding to $k_0 \cong 5$ is required to achieve -40

dB cross-talk. With $k_0 \cong 5$, we have $k_b = \frac{5}{\sqrt{2}} \approx 3.5$, which is a slightly high, but reasonable, value for the mirror-to-mirror spacing in stringent applications like fiber switches. In Chapter 4 we found that the open aperture perpendicular to the optical axis should be 3 times the beam radius, so we can set the fill factor *f=3/3.5=0.86*. The conclusion is that the mirror separation is limited at very closely the same level when we consider cross talk and aperture effects. This is of course the same as what we found for the Matrix Switch. In our calculations we will use the parameter value $k_b = 3.5$ to ensure acceptable loss, distortion, and cross talk.

With the value for k_b established, we can evaluate the scaling performance of the planar Matrix Switch. In terms of the length ($s=Nd$) of the linear array, the maximum fiber-port count is

$$N = 1 + \frac{\pi \cdot \alpha \cdot d \cos\beta}{\lambda \cdot k_b^2} = 1 + \frac{\pi \cdot \alpha \cdot \frac{s}{N} \cos\beta}{\lambda \cdot k_b^2} \Rightarrow$$
$$N \approx \frac{1}{k_b}\sqrt{\frac{\pi \cdot \alpha \cdot s \cos\beta}{\lambda}} = \sqrt{\frac{\pi \cdot 0.4 \cdot s \cdot 0.8}{1.55um \cdot 3.5^2}} \approx 0.23 \cdot s \cdot um^{-1} \tag{8.23}$$

Solving for the length of the linear switching array we find

$$s \approx \frac{N^2 k_b^2 \lambda}{\pi \cdot \alpha \cdot \cos\beta} \approx \frac{N^2 \cdot 3.5^2 \cdot 1.55um}{\pi \cdot 0.4 \cdot 0.8} = N^2 \cdot 19um \tag{8.24}$$

We conclude that the planar Beam Steering Switch scales almost exactly like the Matrix Switch. (For the Matrix Switch we found: $s = \frac{2N^2 \cdot k_m^2 \lambda}{\pi} \approx N^2 \cdot 25\mu m$). This is not a complete coincidence, because both switches are fundamentally limited by the same effect, namely diffraction. Just as the Matrix Switch, the planar Beam Steering Switch will in practice be limited to substantially less than 100 by 100 ports.

Figure 8.16 shows that there are significant differences in the path length through the Beam Steering Switch for different connections. These optical path-length differences cause insertion loss due to the optical-mode mismatch at the output fibers. Using the formulas we have found for the number of fiber ports $\left(N - 1 = \frac{\pi \cdot \alpha \cdot \sqrt{2}\omega_0}{\lambda \cdot k_b}\right)$, mirror separation $\left(d = \frac{k_b \cdot \omega_0 \sqrt{2}}{\cos\beta}\right)$, and array separation

$\left(\frac{p}{2} = \frac{\pi\omega_o^2}{\lambda}\right)$, we find that the difference between the longest and shortest path lengths is

$$\Delta p = \sqrt{p^2 + d^2(N-1)^2} - p = \sqrt{p^2 + p^2\frac{\alpha^2}{\cos^2\beta}} - p \Rightarrow$$
$$\frac{\Delta p}{p} \approx \frac{\alpha}{2\cos\beta} \approx \frac{0.4}{2\cdot 0.8} = 0.25 \tag{8.25}$$

The variation in path length is only a small fraction of the optimum array separation, which is equal to the Rayleigh length of the beam. Differences of less than a quarter of the Rayleigh length lead to transmission larger than 0.98 (see Fig. 8.8 and 8.10). In practice, the insertion loss will be much larger and dominated by alignment errors, so the path-length differences through the planar Beam Steering Switch are of no practical consequence.

3-D Beam Steering Switch

The planar Beam Steering Switch we have analyzed (Fig. 8.16) has a linear array of input mirrors and a matching array of output mirrors. The arrays are configured so that every beam-steering mirror directs the beams within one plane and, consequently, the mirror-rotation axes are perpendicular to that plane. Non-planar configurations leads to higher port counts if the value of the incident angle β can be reduced or even set to zero as shown in the out-of-plane switching geometry of Fig 8.17.

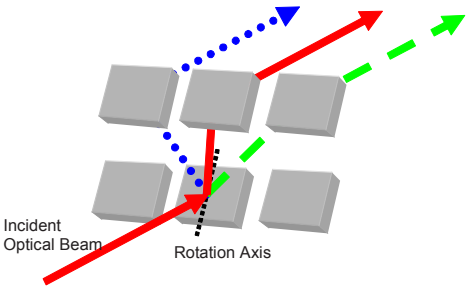


Figure 8.17. Schematic of Beam Steering Switch with out-of-plane switching. This architecture is more efficient than the planar one of Fig. 8.16, because the tilt of the mirror is not in the direction of switching, so the mirrors can be smaller and closer together in the switching dimension. Consequently, more fiber channels can be supported.

The optimum solution from the point of view of maximizing the number of fiber-ports is clearly to combine the in-plane and out-of-plane designs of Figs. 8.16 and 8.17. That requires the mirrors to tilt around two axes. Two-axes rotation complicates the MEMS implementation, but the extra complexity comes with a very important advantage; the maximum possible fiber port count increases dramatically! With 2-axes beam steering the maximum number of ports can be expressed

$$N = 1 + \cos\beta \left(\frac{\pi \cdot \alpha \cdot d}{\lambda \cdot k_b^2} \right)^2 \tag{8.26}$$

or in terms of the space-bandwidth product

$$N = 1 + \cos\beta \left(\frac{4\pi \cdot t}{\lambda \cdot f \cdot k_b^2} \right)^2 \tag{8.27}$$

where it is assumed that the incident angle is zero in one dimension and β in the other. Solving the first of these equations for the mirror size d, we find the following expressions for mirror size

$$d \approx \sqrt{\frac{N}{\cos\beta}} \frac{\cdot k_b^2 \lambda}{\pi \cdot \alpha} \tag{8.28}$$

and array size[f]

$$s = d\sqrt{N} \approx N \frac{k_b^2 \lambda}{\pi \cdot \alpha \sqrt{\cos\beta}} \approx N \frac{3.5^2 \cdot 1.55 \mu m}{\pi \cdot 0.4 \cdot \sqrt{0.8}} \approx N \cdot 17 \mu m \tag{8.29}$$

We see that reasonably small mirror arrays (~ 20 mm) can support very large switches with a thousand fiber ports or more. By pushing the chip size further it is possible to extend the port count beyond 5,000. The path-length differences between the shortest and longest paths through the switch are somewhat longer than then for the planar Beam Steering Switch, but the losses caused by path-length differences are still insignificant compared to losses caused by alignment errors.

The enormous port count supported by the 3-D Beam Steering Switch is due to the exceptional resolution of the beam-steering mirrors. The ability of scanning micromirrors to (de)multiplex thousands of channels is unmatched by other optical (de)multiplexing technologies, irrespective of the (de)multiplexing dimension (wavelength, frequency, time, polarization).

[f] We are interested in the big picture here, so we won't worry about the fact that, due to the tilt, the array is rectangular, not square.

8.6.2 MEMS Implementations of the N by N Beam Steering Switch

The implementation of the planar Beam Steering Switch is similar to that of the Matrix Switch, and therefore has many of the same advantages and challenges. On the plus side we have that the planar geometry makes it possible to use lithographically defined structures to position and align the fibers, the lenses, and the beam-steering mirrors in a single plane. That allows single-chip solutions that are compact, well aligned, and mechanically robust.

Just as is the case for the Matrix Switch, however, the switching mirrors of the Beam Steering Switch represent an implementation challenge. The mirrors have to be oriented and rotated around an axis that is perpendicular to the chip surface. The difficulty of designing and fabricating such mirrors, particularly for the mm sizes required for large port-count switches, is so formidable that at present very few demonstrations and no commercialization of the planar Beam Steering Switch have been reported. The exception is the WDM optical cross connect (WDM OXC) shown in Fig. 8.6. This switch is constructed from a set of planar Beam Steering Switches, one for each wavelength. However, the favored implementation of the WDM OXC is to arrange all the input mirrors on one substrate and all the output mirrors on another. In this configuration, the implementation of the WDM OXC is very close to that of the 3-D Beam Steering Switch that is discussed in the next paragraph.

The very nature of the 3-D Beam Steering Switch makes it impossible to implement it in a planar geometry. In fact, this switch is very much a three-dimensional structure that consists of a set of two-dimensional arrays as shown in Fig. 8.18. Each of the arrays presents a challenge to the switch manufacturer. One-dimensional fiber arrays are common, but two-dimensional arrays are not yet offered commercially, so each switch designer has to develop their own solution. The situation is a little better for the micro-lens arrays. Commercial products are available, but the degrees of freedom in the design of microlenses and microlens arrays are vast, so typically each switch requires a custom design. The good news is that there is an increasing number of vendors that offer custom microlens design and fabrication, so the manufacturers of 3-D Beam Steering Switches do not need to develop that expertise in house.

The mirror arrays shown in Fig. 8.18 are the heart of the 3-D Beam Steering Switch. These mirrors are significantly easier to design, fabricate, and operate than the Mirrors of the Matrix Switch, because the Beam-Steering mirrors do not have to move complete out to the beam path. On the contrary, each input beam is always directed to the same input mirror, and each output beam is coming off the same output mirror. This means that mirrors do not need to be linearly translated, just rotated. The mirrors are in essence scanners and subject to the same implementation consideration that we discussed in Chapter 7. As pointed out in there, it is relatively speaking easy to design and fabricated single-axes mirrors that can be

used to build the WDM OXC, but significantly harder to create the two-axes scanning mirrors required to implement the full-blown 3-D Beam Steering Switch.

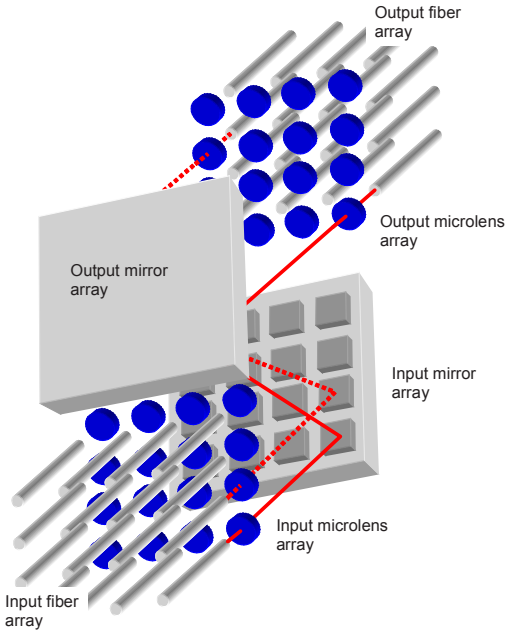


Figure 8.18. Schematic of the 3-D Beam Steering Switch. Two different paths through the switch are shown as solid and dotted lines. The packaging of this system is difficult because the two mirror arrays, the two lens arrays, and the two fiber arrays must all be well aligned and robustly held to ensure correct and stable operation.

In addition to the challenges of creating the fiber, lens, and mirror arrays required by the 3-D Beam Steering Switch, it is also very difficult to find cost effective ways to position and stabilize the arrays. The individual components of the arrays must be aligned to sub-micron precision and be stable over long lifetimes under adverse conditions, including large temperature variations. For a complex mechanical structure as the one shown in Fig. 8.18, this means that the package has to be large, sturdy, and precise; not a combination that comes cheap. The large number of channels supported by the Beam-Steering Switch also complicates and drives the cost of the electronic control. Package design and electronic integration are therefore very important, maybe the most important, parts of the Beam Steering Switch implementation.

8.7 Summary of MEMS Fiber Switches

This chapter is focused on the fundamental limits of scaling with the goal of understanding how different MEMS switch architectures compare. We consider four switches, starting with the 2 by 2 Matrix Switch, continuing with the N by N Matrix Switch and the Planar Beam Steering Switch, and finally ending with the 3-D Beam Steering Switch. Going through this hierarchy we find that the switches become more complex, but also more powerful in terms of the number of fiber ports that they can accommodate.

The 2 by 2 Matrix Switch, as the name suggests, has only two input fibers and two output fibers. In its simplest form the 2 by 2 consists of only the fibers and one switching mirror in addition to the mechanical alignment structures. In this configuration, the switch suffers a diffraction loss that limits the transmission to about 35%, but that loss can be reduced by adding collimating lenses to the design.

There are two compelling reasons to use MEMS technology to implement the 2 by 2 Matrix Switch. First the switching mirror can be made with thicknesses in the 5 to 10 um range which is what is required to not add significant state-dependent loss. Just as important is the fact that MEMS fiber grooves and other positioning structures enable simple and stable alignment of the switch.

The scaling of the N by N Matrix Switch can be expressed in the formula

$$N_{\max} = \frac{1}{k_m}\sqrt{\frac{\pi \cdot s}{2\lambda}} \tag{8.30}$$

where s is the length of the side of the switch matrix, λ is the wavelength, and k_b is a constant that for typical switch conditions and requirements takes a numerical value approximately equal to 5. At a wavelength of $\lambda=1.55um$, the relationship between N and s evaluates to $s=N^2 \cdot 25um$, showing that $N=100$ requires a switch matrix measuring about 25 cm on a side. This is an impractical MEMS device, so scaling properties limits the port count of the N by N matrix Switch to well less than 100 by 100.

As for the 2 by 2, the N by N benefits from the precise and stable alignment structures that can be created by MEMS fabrication technology. The MEMS implementation is, however, complicated by the fact that the mirrors must be moved all the way in and out of the beam paths in the Matrix Switch. This means that large switches require large mirrors that must be able to move over long distances, creating a challenging MEMS design problem.

As is the case for the Matrix Switch, the scaling of the Planar Beam Steering Switch is determined by diffraction of the optical beams propagating through the switch. The total number of fiber ports supported by the switch can be expressed

$$N-1=\frac{1}{k_b}\sqrt{\frac{\pi\cdot\alpha\cdot s\cos\beta}{\lambda}} \tag{8.31}$$

where α is the optical scan angle, s is the linear array size, and β is the angle between a line perpendicular to the mirror at zero deflection ($\alpha = 0$) and the input optical beam. The constant k_b is again given the numerical value of 5. Assuming a wavelength of λ=*1.55um*, an incident angle *cos*β=*0.8*, and an angle range α=0.4, the relationship between N and s evaluates to s=N^2*·19 um*, showing that the Planar Beam Steering Switch scales very similarly to the N by N Matrix Switch and is therefore limited to similar port counts and mirror sizes. Like the Matrix Switch, the Planar Beam Steering Switch makes good use of the precise and stable alignment structures that can be fabricated in MEMS, but the mirror design represents a challenge.

By extending the Beam Steering Switch to 3 dimensions, its scaling is substantially improved, and can be expressed

$$N-1=\frac{s\cdot\pi\cdot\alpha\sqrt{\cos\beta}}{k_b^2\lambda} \tag{8.32}$$

where the symbols have the same meanings as above. The radical difference from the Matrix Switch and the Planar Beam Steering Switch is that the port count N here depends linearly, rather than as the square-root, on the array size s. With the same vales as above, the relationship between N and s evaluates to $s \approx N\cdot 17\mu m$. This demonstrates that the 3-D Beam Steering Switch can be scaled to port counts in the thousands!

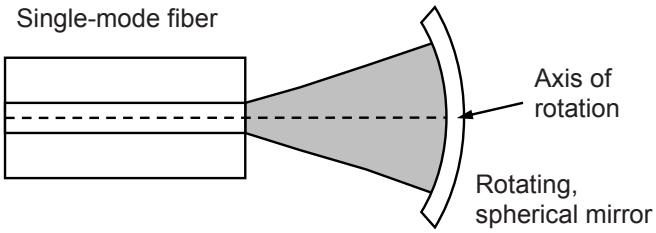
The MEMS implementation of the 3-D Beam Steering Switch is quite different from that of the Matrix and Planar Beam Steering Switches. The essential difference is that the three dimensional structure of the switch cannot be aligned using the same MEMS alignment structures that works so well for planar geometries. Instead, bulky mechanics must be used to position the fiber arrays, lens arrays, and mirror arrays in a stable configuration.

The beam-steering mirrors themselves are relatively straight-forward to implement. Each beam steering mirror is a 2-axis scanner like the ones described in Chapter 7. They can be made and operated in the plane of the substrate or chip, and they only need rotational motion, not translation, because they do not need to translate out of the path of the optical beams.

Exercises

Problem 8.1 - Spherical-Mirror Modulator

Consider the 1 by 1 switch (modulator) in the figure below. The input is the forward-propagating light on the fiber, and the output is the backward propagating light. The beam from the fiber is focused on the spherical mirror, which has a radius of curvature that matches that of the beam. So with the mirror in the quiescent position, all the light is coupled back on the fiber, while rotation of the mirror leads to reduction of the back coupling.



Spherical-mirror modulator.

a. Express the back-coupled optical power relative to the input power. Give your answer in terms of the wavelength, the fiber-mirror distance, and the angle of rotation of the mirror.
b. Is there a mirror size that minimizes the required maximum motion (i.e. the motion of the mirror ends that must move the longest distance under rotation)?
c. How would you choose the mirror size in a practical implementation? What is the smallest volume that the complete switch (not counting the fiber) must occupy?

Now consider a variation of the spherical-mirror modulator, in which the mirror isn't rotated, but rather its radius of curvature is changed by a MEMS actuator.

d. Express the relative back-coupled optical power as a function of the radius of curvature of the mirror. Give your answer in terms of the wavelength, and the fiber-mirror distance.
e. Compare the two modulator principles. Which one is better from an operational point of view and which one is simpler to implement?

Problem 8.2 - Optimized Corner-Cube Modulator

Optimize the corner-cube modulator of problem 7.3. What is the smallest volume that the complete corner-cube switch (not counting the fiber) must occupy?

Problem 8.3 - Fourier Lens

One alternative implementation of the beam-steering switch is to place a Fourier lens between the mirror arrays in Fig. 8.18, such that the lens is one focal length away from either array.

a. How does such a Fourier lens change the scaling of the Beam-Steering Switch?
b. Is there a specific type of mirror array (array size, mirror size, and mirror rotation angle) that favors the use of Fourier lenses?

Problem 8.4 - Beam-Steering Switch with Normal Incidence

a. How can you use polarizing beam splitters and $\lambda/4$ plates to avoid the difficulty of non-normal incidence on the mirror arrays in the Beam-Steering Switch?
b. How does this implementation change the scaling of the Beam-Steering Switch?
c. How can this implementation be combined with a Fourier lens?

Problem 8.5 - "Immersion" Switches

a. How is the scaling of the Matrix Switch changed by immersing the whole switch in an oil of index 1.5?
b. How is the scaling of the Beam-Steering Switch changed by immersing the whole switch in an oil of index 1.5?
c. How is the scaling of the Matrix Switch changed by filling the volumes between the mirrors with silicon of index 3.5? Here we have to leave room for the mirrors to move, so the beams must cross two air-silicon interfaces at each mirrors. Assume that these are Anti-Reflection coated so we can ignore reflections from these interfaces.
d. How is the scaling of the Beam-Steering Switch changed by filling the volumes between the mirror arrays with silicon of index 3.5? Again we must leave room for the mirrors to move, and we assume that we can ignore reflections from the air-Si interfaces.

Problem 8.6 - Non-planar Beam-Steering Switches

The Beam-Steering-Switch implementation of Fig. 8.18 has planar mirrors, planar array substrates, and the mirrors are parallel to the substrates in their quiescent positions.

a. Can we improve the scaling of the switch by relaxing any one or any combination of these three conditions?
b. Can we improve any practical aspect of the switch implementation by relaxing any one or any combination of these three conditions?

Problem 8.7 - Hybrid Switch

a. Design a hybrid switch, i.e. one that uses the operational principles of both the Matrix Switch and the Beam-Steering Switch.
b. What advantages does such a hybrid design have?

References

1 C. Marxer, N.F. de Rooij, Journal of Lightwave Tech., Vol. 17, No. 1, Jan 1999.

2 Bryant Hichwa etal, OCLI/JDS Uniphase, "A Unique Latching 2x2 MEMS Fiber Optics Switch", Optical MEMS 2000, Kauai, August 21-24th, 2000.

3 E. L. Goldstein, L. Eskildsen, A.F. Elrefaie, "Performance Implications of Component Crosstalk in Transparent Lightwave Networks", Photonics Technology Letters, Vol. 6, No. 5, May 1994, pp. 657-660.

4 E.M. Kim, D.L. Franzen, "Measurement of farfield and near-field radiation patterns from optical fibers", National Bureau of Standards, Report No.: NBS-TN-1032, Washington DC, USA, February 1981.

5 U. Krishnamoorthy, "Design and Fabrication of Micromirrors for Optical Applications", Ph. D. thesis, University of California, Davis, 2002.

6 T. Akiyama, D. Collard, H. Fujita, "Scratch drive actuator with mechanical links for self-assembly of three-dimensional MEMS," Journal of MicroElectroMechanical Systems, vol. 6, no. 1, March 1997, pp. 10–17.

7 S.-S. Lee, L.-S. Huang, C.-J. Kim, M.C. Wu, " Free-Space Fiber-Optic Switches Based on MEMS Vertical Torsion Mirrors", Journal of Lightwave Technology, vol. 17, no. 1, January 1999, pp. 7-13.

8 M.J. Daneman, N. C. Tien, O. Solgaard, A.P. Pisano, K. Y. Lau, R. S. Muller, "Linear Microvibromotor for Positioning Optical Components", IEEE Journal of MicroElectroMechanical Systems (JMEMS), vol. 5, no. 3, September 1996, pp. 159-165.

9 P.M. Hagelin, U. Krishnamoorthy, J.P. Heritage, O. Solgaard, "Scalable Optical Cross-Connect Switch Using Micromachined Mirrors", IEEE Photonics Technology Letters, vol. 12, no. 7, July 2000, pp. 882-885.

9: Micromirror Arrays – Amplitude and Phase Modulation

9.1 Introduction to Micromirror Arrays

The success of ICs and MEMS alike is based on the photolithographic fabrication process and is its ability to create large arrays of devices. Projection displays based on micromirrors [1] like the Digital Light Processing (DLP)® technology [2,3,4], is a good example from the MEMS world. Large DLP chips contain arrays of more than one million micromirrors that each can be mechanically tilted to direct incoming light in a prescribed direction. Each micromirror is a relatively complex, three-dimensional, electro-mechanical structure of very small dimensions. Given the size, complexity, and required precision, it would be a formidable challenge to make even small numbers of these micromirrors using traditional mechanical fabrication technologies. Fabricating millions of micromirrors on a single substrate is therefore a unique capability of MEMS technology and a practical impossibility without the parallel-processing power of optical lithography.

In this chapter we will describe the optical properties of large arrays. We start by taking a closer look at the DLP® technology, particularly the implementation and integration challenges that the DLP® presents. In a typical application of the DLP®, each mirror acts as an amplitude modulator. It is, however, also possible to make arrays of micromirrors that act as phase modulators. In fact, one of the compelling possibilities of large mirror arrays is that they can be used to perturb the phase of an incident wave front over an extended area to enable flexible and precise manipulation of the optical field.

One of the main purposes of this chapter is to describe phase-modulating mirror arrays and compare their properties to those of amplitude-modulating arrays. We will find that in some well-defined ways, phase modulators perform better than amplitude modulators. This is an observation that is repeated in many areas of optics, and it follows from the fundamental fact that light is radiating electromagnetic ***fields***. A phase modulator controls the optical field by setting up interfering fields that can add or subtract according to the state of the phase modulator. This is more efficient than simply attenuating the optical intensity, which is what an amplitude modulator does. Phase modulation is therefore a very powerful tool

that is used in a wide range of Optical Microsystems. The basic concepts are described in this Chapter together with several examples of their use with micromirror arrays, and the principles of phase modulation will be used again and again throughout the book to describe diffractive Optical MEMS, MEMS filters, and interactions in Photonic Crystals.

9.2 Amplitude Modulating Mirror Arrays

Amplitude-modulating micromirror arrays are used in many image-forming applications, including projection displays, mask less lithography, spectroscopy, and confocal microscopy. The Texas Instrument's DLP® technology is the most successful of these technologies. The schematic drawing shown in Fig. 9.1 illustrates how complex the DLP® mirrors are, and solidifies the point made in the introduction that a chip with millions of such structures cannot be made in a serial process.

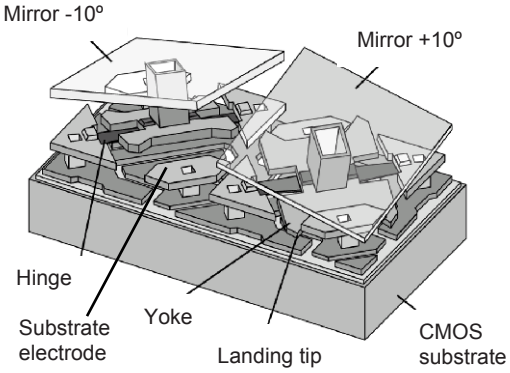


Figure 9.1 *Schematic drawing showing two pixels in TI's DLP® technology. DLP® chips contain as many as one million of these complex micromirrors. Each mirror can be in one of three states; With no applied voltage the mirror is parallel to the substrate surface. With an actuation voltage applied between the mirror and the left-most substrate electrode, the mirror tilts to the left, shown as -10º, and when the an actuation voltage is applied between the mirror and the right-most substrate electrode, the mirror tilts to the right, shown as +10º. Reprinted with permission.*

The DLP® mirror has three states. With no applied voltage, there is no force to make the mirror rotate around the rotation hinges, so the mirror is in its quiescent state with the mirror surface parallel to the chip surface. With a voltage applied to one of the substrate electrodes, the mirror rotates around the hinge in the direction of the voltage. Once the landing tips contacts the substrate, the rotation stops and the mirror is in either the +10º rotation state or the -10º rotation state, depending on whether the applied voltage is on the right or left.

Figure 9.1 shows that the points where the lading tips contacts the substrate are electrically connected to mirrors, so there is no potential difference between the contacting bodies. Note further that the mirror itself never contacts the substrate electrodes. The landing tips stops the rotation before that can happen, so there is no electrical shorting of the mirror to the substrate electrode. Both these points are important for reliable operation and long-term stability of the mirror.

9.2.1 Projection Display

The micromirror arrays can be used for many different types of applications, including mask-less lithography [5], spectroscopy [6] and confocal microscopy, in which the mirror array is used as a programmable spatial mask [7]. The most common application, and the one the DLP® was designed for, is projection displays. A schematic version of a projection display is shown in Fig. 9.2.

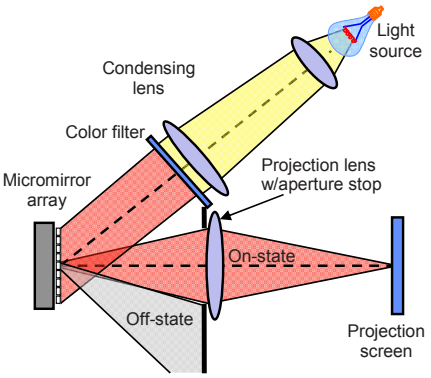


Figure 9.2 *Projection system used to form a monochromatic image. Three aligned such systems are needed to create a full color display. Mirrors that are in the on state (tilted towards the incident light) are imaged as bright on the screen, while mirrors in the off state (tilted away) appear dark on the screen.*

In this system, the light from a traditional source[a] is collimated onto a micromirror chip, typically through a color filter that transmits blue, red, or green. A full color display would then need three aligned systems, together forming aligned blue, green, and red images that combine to form a color image. Each mirror in the micromirror array is either tilted towards the light source (+10º) or away from it (-10º). If the mirror is tilted toward the incoming light, then the light hitting it is directed through the projection lens and focused on the projection screen, so that the image of that mirror on the screen appears bright. If the mirror is tilted away from

[a] Lasers can be used, but one of the strengths of the micromirror array is that it works with traditional light sources, so that it the typical choice.

the incoming light, then the incoming light is directed outside the projection lens, and the image of the particular mirror appears dark on the screen.

In the introduction we point out that MEMS is the only practical option for making 2-D arrays of large numbers of switching mirrors as required by the projection display of Fig. 9.2. In addition, the MEMS implementation has several structural and functional advantages. First and foremost MEMS enable direct integration of electronics and optics. Under each mirror in the DLP® array is a switching and memory circuit that holds the state of the mirror until the next desired state is sent. The direct integration is made possible by the fact that the MEMS is fabricated on a silicon substrate. The procedure is to first create the switching circuitry in the silicon substrate, and then fabricate the MEMS on top. This means that the post processing must be done at low temperatures[b], which puts restrictions on the MEMS technology that can be used. The DLP® is therefore built using Aluminum[c] structural layers and polymer sacrificial layers that can be removed by standard dry-etch techniques.

An important functional advantage of the MEMS implementation is that the small size and mass of the DLP® mirrors gives them high switching speeds. The switching time for the individual mirrors of the array is in the tens of microseconds, which is more than fast enough for gray scale to be created by pulse-length modulation. This means that the mirror can be operated in the binary fashion described above, and that the complications of analog control of the tilting angles to create grayscale can be avoided.

Finally, the small size of the MEMS array leads to a compact system design. This makes the overall system inexpensive to fabricate, distribute, and operate, and makes is practical to combine multiple arrays in one system, e.g. using three DLP arrays to create a full color display.

The reflective, as opposed to transmissive, operation of mirror arrays is both an advantage and a challenge. Aluminum-coated mirrors have reflectivity around 92% at visible wavelengths (see Table 7.1), and the mirror fill factor (the ratio of the mirror area to the total area of the array) is around 90% for a total light throughput of 83%. Compared to Liquid Crystals and other transmissive Spatial Light Modulators, this is a very high light efficiency. Reflective optics is also

[b] Exactly what low temperatures means will depend on the IC technology that is used. If the circuits have been metalized, then temperatures have to be restricted to well below 500º C, while slightly higher temperatures can be used if the IC process is designed for MEMS post fabrication.

[c] Not just any Aluminum, however. The exact alloy is optimized for long-term mechanical stability.

much less dispersive (wavelength dependent) than even the best chromatically-compensated transmissive optics[d].

The reflective operation of mirror array creates problems in the system design, however. Optical systems based on reflective components tend to be bulkier and more difficult to align than those based on transmissive optics. This is because reflective devices require folded geometries and off-axis operation, as can be appreciated by considering the system shown schematically in Fig. 9.2. Creating good folded-optics solution is one of the main challenges for the Optical MEMS designer.

Projection Display Resolution

Compactness is important for the functionality and the cost of micromirror arrays, so we need to determine how small each mirror in the array can be and still perform its function. Rotating mirrors like the DLP mirror are examples of the 1-axis scanners that are described in Chapter 7, and we can use the tools developed there to determine how big each mirror has to be.

We model an individual mirror in an array as shown in Fig. 9.3. The micromirror rotates, without translation, around a fixed axis of rotation. The illumination of the individual micromirrors in an array will typically be uniform across the mirror, but we will use the Gaussian Beam theory we developed in Chapters 4 and 7 to calculate the diffraction angle from the mirrors. The results will not be exact due to the discrepancy in the illumination profile, but the errors are relatively small and do not change the overall conclusions we draw from the modeling.

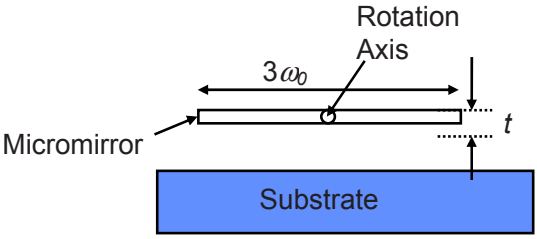


Figure 9.3 *Simple micromirror model showing the definitions of the parameters used to calculate the required size and angular deflection.*

In Chapter 7 we found that an ideal mirror illuminated by a Gaussian profile on gives a number of resolvable spots that can be written

[d] This is the reason that reflective microscope objectives are popular for some applications, in spite of their extra complexity and cost.

$$N = \frac{\theta_{tilt}}{\theta_{diff}} + 1 \tag{9.1}$$

The exact definition of the diffraction angle depends on the specified contrast, C_{spec}, which is the required ratio of the highest to lowest intensity in the projected image of the micromirror. The beam radius where the intensity is reduced to meet the contrast specification is

$$e^{-2\frac{r_c^2}{\omega^2}} = \frac{1}{C_{spec}} \Rightarrow r_c = \omega\sqrt{0.5\ln\left(C_{spec}\right)} \tag{9.2}$$

In projection displays the contrast requirements are between -20 and -30 dB, so using the models we developed in Section 7.2.1, we conservatively set

$$e^{-2\frac{r_c^2}{\omega^2}} = \frac{1}{1000} \Rightarrow r_c = \omega\sqrt{0.5\ln 1000} \approx 1.86 \cdot \omega \tag{9.3}$$

The corresponding diffraction angle is

$$\theta_{diff} = \lim_{z\to\infty} \frac{2r_c}{z} = \lim_{z\to\infty} \frac{2\cdot\omega(z)\cdot\sqrt{0.5\ln\left(C_{spec}\right)}}{z} = 2\sqrt{0.5\ln\left(C_{spec}\right)}\frac{\lambda}{\pi\cdot\omega_0} \approx 3.7\frac{\lambda}{\pi\cdot\omega_0} \tag{9.4}$$

Assuming a mirror size $d=3\omega_0$, we find

$$N = \Delta\theta\cdot\frac{\pi\cdot d}{6\cdot\sqrt{0.5\ln\left(C_{spec}\right)}\cdot\lambda} + 1 \approx \Delta\theta\cdot\frac{\pi\cdot d}{11.2\cdot\lambda} + 1 \tag{9.5}$$

It follows from the operation of amplitude-modulating micromirrors in projection displays that the number of resolvable spots of the mirror must be 2, corresponding to the two states (dark and bright) that the mirror must have. If we assume a wavelength in the middle of the visible region, i.e. $\lambda=500nm$, and a scan angle $\Delta\theta=0.7$, the minimum mirror size is given by

$$d = \frac{6\sqrt{0.5\ln\left(C_{spec}\right)}\cdot\lambda}{\pi\cdot\Delta\theta} \approx \frac{11.2\cdot 500nm}{\pi\cdot 0.7} \approx 2.5um \tag{9.6}$$

This is the result of a rough calculation based on an incorrect assumption of Gaussian Beam Illumination. It is certainly possible to push this limit lower, but not much. The choice $\Delta\theta=0.7$ was made to match the DLP mirror that has a mechanical scan angle of $\pm 10^o$ and therefore a total optical angle range of

$4 \cdot 10^{o} \cdot (\pi/180^{o}) = 0.7$. Increasing the scan angle beyond this value quickly becomes impractical.

The point is that even conservative assumptions lead to the realization that micromirrors for displays can be made as small as five times the wavelength. An array with a million mirrors only has to be about 2.5mm on a side. This limit set by diffraction is so small that in most practical situations, the mirror size is determined by considerations such as the size of the switching circuitry and the MEMS actuators, rather than by the fundamental limit set by diffraction.

It is instructive to express the mirror requirement in terms of t, the minimum deflection of the end point of the mirror as defined in Fig. 9.3. The scan angle can be expressed

$$\Delta\theta = \frac{t}{d/2} = 2\frac{t}{3\omega_0} \tag{9.7}$$

Combining this with the equation for the number of resolvable spots gives

$$N = \frac{t}{6\sqrt{0.5\ln\left(C_{spec}\right)}} \cdot \frac{2\pi}{\lambda} + 1 = \frac{t}{11.2} \cdot \frac{2\pi}{\lambda} + 1 \tag{9.8}$$

From this expression it seems that the number of resolvable spots is to first order not dependent on the mirror size! This is of course a fallacy, because the number of resolvable spots does depend on t, which again sets a lower limit on the mirror size. Assuming $N=2$ and $\lambda=500nm$ and solving for t we find

$$t \geq \frac{6\sqrt{0.5\ln\left(C_{spec}\right)} \cdot \lambda}{2\pi} \approx \frac{11.2 \cdot 500nm}{2\pi} \approx 0.9um \tag{9.9}$$

The necessary maximum deflection is less than two wavelengths. Again this is not a hard limit, but rather a practical consideration that is based on quite conservative assumptions. By relaxing the specified contrast, which we set to 30 dB, the required deflection will be reduced, and for many applications, a maximum deflection on the order of a wavelength is sufficient.

9.3 Projection of Micromirror Arrays

In Section 9.2.1 we analyzed the resolution of micromirror arrays in isolation without explicit consideration of the optical system. Implicitly we made the assumption that the optical system was able to capture the full diffraction angle from each micromirror as shown in Fig. 9.2. Now we must take a closer look at this assumption and learn what kind of optical system that is required for projection of micromirror arrays.

We will start by considering amplitude modulation as in the preceding section, but our investigation will show that phase modulation allows smaller projected image features, so phase modulation is preferable if image resolution is the highest priority. We will also find, however, that there is a price to pay for the improved resolution; phase-modulated arrays are more complicated to operate than amplitude-modulated arrays.

9.3.1 The Point Spread Function

Consider the imaging system of Fig. 9.4. The system projects an image of the micromirror array onto the screen. It is shown here as having only a single lens. In practice a combination of several lenses will be used to optimize different aspects of the performance of the system.

The appearance of an individual mirror on the screen depends on the illumination, the state of the mirror, and the magnification of the projection system. In addition it also depends on the Point Spread Function (PSF) of the lens system. The PSF is defined as the impulse response of an optical system, i.e. the image of a perfect point source. No optical system is able to create an impulse in the image plane, so the PSF is always a pattern of finite size. The finite PSF of real optical systems therefore blurs the projected images. If the PSF is large compared to the projected size of a micromirror, then the image of that micromirror is dominated by the characteristics of the optical system rather than by the micromirror itself. Under those conditions, the image of the micromirror bears little resemblance of the object. It is blurred by the finite PSF of the lens system, and we say that the micromirror cannot be resolved by the optical system.

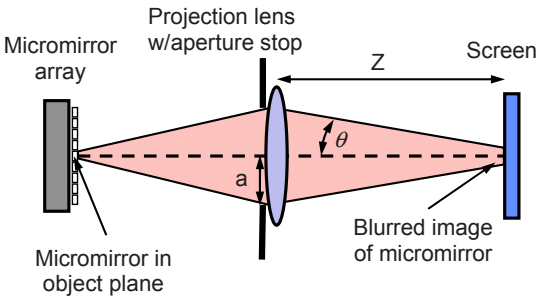


Figure 9.4 *Detail of a micromirror-array projection system. The mirror array, which defines the object plane, is imaged onto the screen, which defines the image plane. The Point Spread Function (PSF) of the projection system blurs the images of the micromirrors on the screen, so that objects with images that are smaller than the PSF cannot be resolved on the screen.*

The PSF of a simple projection system is the famous Airy disc [8] with the following mathematical definition

$$I(\theta) = I_0\left(\frac{2J_1(x)}{x}\right)^2 = I_0\left(\frac{2J_1(ka\cdot r/Z)}{ka\cdot r/Z}\right)^2 \tag{9.10}$$

where I_0 is the field at the center of the disc, J_1 is the Bessel function of the first kind and order one, $k = \frac{2\pi}{\lambda}$ is the propagation constant or wave number, λ is the wavelength, a is the radius of the lens aperture, r is the radial distance from the disc center and Z the lens-screen distance. Most often the Airy-disc formula is given for the specific case of imaging from infinity, i.e. the case where an image formed in the focal plane at a distance of one focal length from the lens. The formula works just as well for imaging at any distance, however, as long as Z is not taken to be the focal length, but to be the correct lens-screen distance for the imaging set up.

The Airy disc is usually defined in terms of its optical intensity as in Eq. 9.10, due to the fact that this concept originates with incoherent imaging. For our purposes we need to consider addition of optical fields, so the field distribution is the appropriate object for our calculations. It can be expressed

$$\vec{E}(\theta) = \vec{E}_0\frac{2J_1(x)}{x} = \vec{E}_0\frac{2J_1(ka\cdot r/Z)}{ka\cdot r/Z} \tag{9.11}$$

where I_0 is the field at the center of the disc, and the other parameters are defined above.

In many cases it is convenient to rewrite the Airy-disc expression in terms of the system *f*-number, which is defined as $N=Z/2a$[e]. The expression then becomes

$$\vec{E}(\theta) = \vec{E}_0\frac{2J_1(\pi\cdot r/\lambda N)}{\pi\cdot r/\lambda N} \tag{9.12}$$

This form highlights the dependence on the *f*-number of the imaging system; the smaller the *f*-number the smaller the PSF.

The Airy-disc formula is also the far-field pattern from a circular aperture without an imaging lens. In that case it is useful to express the pattern in terms of the observation angle, θ, which is related to r and Z as $\sin\theta = r/Z$. Using this formulation, the Airy disc pattern becomes

[e] Here we are talking about the generalized system *f*-number that simplifies to the standard *f/2a* for imaging at infinity.

$$\vec{\mathrm{E}}(\theta) = \vec{\mathrm{E}}_0 \frac{2J_1(ka\sin\theta)}{ka\sin\theta} \tag{9.13}$$

The Airy Disc field distribution is shown in Fig. 9.6. The central lobe of the curve is bell-shaped as we would expect. The side lobes are created by the sharp cut-off of the aperture, just as hard edges creates side lobes in truncated Gaussian as we observed in section 4.7.

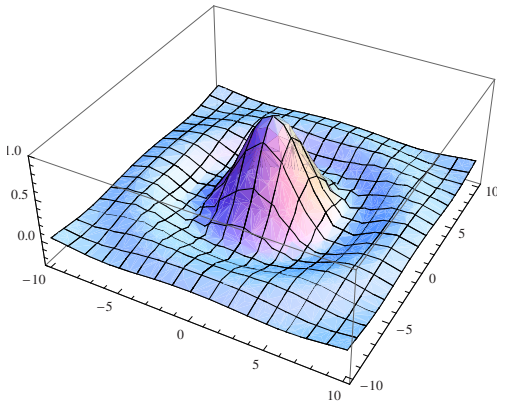


Figure 9.5 *The two-dimensional Airy Disc field pattern is circularly symmetric with a bell-shaped central lobe and significant side lobes. The argument is* $\sqrt{x^2+y^2} = \pi \cdot r/(\lambda N)$*, where r is the radial distance from the disc center,* λ *is the wavelength, and N is the system f-number.*

The field and intensity patterns of the Airy Disc are compared in Fig. 9.6. Notice the dramatic reduction in the appearance of the side lobes in the intensity compared to the field distribution. It is tempting to conclude that these side lobes are mere curiosities that play no significant practical role. For coherent systems, that is a fallacy, because we must consider the field PSF. The effects of the side lobes become pronounced when they interfere with neighboring distributions.

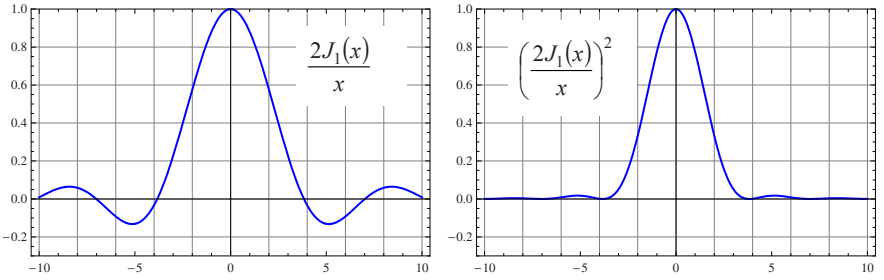


Figure 9.6 *Comparison of the Airy Disc field (a) and intensity (b) distributions.*

The first nulls in the distributions appear at $x \approx 3.83$. Taking the radius of the first null to be the radius of the image, we find the famous expression

$$x_{1st} = \frac{\pi \cdot r_{1st}}{\lambda \cdot N} \approx 3.83 \Rightarrow r_{1st} \approx 1.22 \cdot N \cdot \lambda \tag{9.14}$$

In the case of imaging of an object plane at long distances, the lens-screen distance equals the focal length ($Z=f$) and the denominator becomes the standard *f*-number, $2a/f$, which is how this expression is typically presented.

The half maximum value of the intensity distribution appears at $x \approx 1.62$, so the full width at half maximum (FWHM) of the intensity is

$$x_{HM} = \frac{\pi \cdot r_{HM}}{\lambda \cdot N} \approx 1.62 \Rightarrow FWHM = 2r_{HM} \approx 2\frac{1.62 \cdot N \cdot \lambda}{\pi} \approx N \cdot \lambda$$

This simple-to-remember formula says that the FWHM of the central lobe of the Airy Disc equals the system *f*-number multiplied by the wavelength!

The Airy Disc is an exact expression for the far-field diffraction from a circular aperture. Most practical lenses will have imperfections that make the far-field pattern deviate from this ideal. In high-quality systems optimized for monochromatic operation, e.g. optical lithography machines used in the IC and MEMS industries, these deviations are minor, and the system PSF approaches the Airy Disc. In other cases, the aperture function is engineered to achieve a desired point-source (impulse) response. For example, it is possible to use a "Gaussian Aperture"[9] to get a Gaussian PSF that has the advantage of uniformly decreasing distribution without side lobes.

Even the best designed lens systems will at least to some degree deviate from the ideal Airy-Disc PSF. In addition, the Airy Disc is difficult to use in analytical models and even create problem in some numerical calculations. These issues lead us to use simpler PSDs. In particular we will make extensive use of Gaussian approximations in our analytical calculations, because they greatly simplify the math and allow us to derive closed-form solutions to problems that are analytically intractable if we use the full-blown Airy disc formulation.

We define the Gaussian PSF for the optical field and for the intensity as

$$\vec{E}(x) = \vec{E}_0 e^{-\frac{x^2}{c_G^2}} = \vec{E}_0 e^{-\frac{\ln 2}{2} \cdot \frac{x^2}{1.62^2}} \tag{9.15}$$

$$I(x) = I_0 \left(e^{-\frac{x^2}{c_G^2}} \right)^2 = I_0 e^{-\ln 2 \frac{x^2}{1.62^2}} \tag{9.16}$$

where the parameter x is defined as for the Airy Disc PSF, i.e. $x = \frac{ka \cdot r}{Z} = \frac{\pi \cdot r}{\lambda N}$. Here again r is the radial distance from the disc center, λ is the wavelength, and N is the system f-number. The value of the constant $c_G = \frac{1.62}{\sqrt{\ln 2/2}}$ is chosen so that the FWHM of the intensity Gaussian PSF matches that of the Airy disc intensity distribution.

This choice is made to closely match the central lobes of the two intensity PSDs. Rewriting the Gaussian PSDs in terms of the projection-system parameters we find

$$\vec{E}(x) = \vec{E}_0 e^{-\frac{\ln 2}{2} \cdot \frac{\left(\frac{\pi \cdot r}{\lambda N}\right)^2}{1.62^2}} = \vec{E}_0 e^{-\frac{r^2}{\omega_{PSF}^2}} \tag{9.17}$$

$$I(x)^2 = I_0 e^{-\ln 2 \frac{\left(\frac{\pi \cdot r}{\lambda N}\right)^2}{1.62^2}} = I_0 e^{-2\frac{r^2}{\omega_{PSF}^2}} \tag{9.18}$$

where we have introduced the PSF beam radius $\omega_{PSF} = \frac{1.62}{\ln 2 \cdot \pi} \lambda N \approx 0.74 \lambda N$.

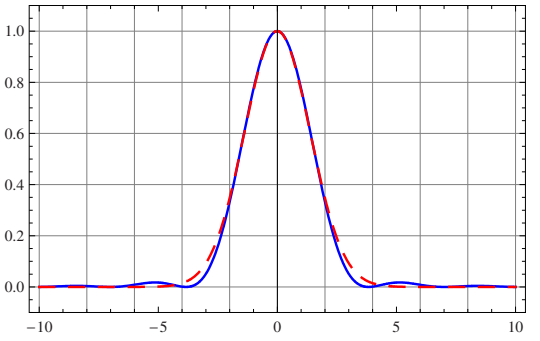


Figure 9.7 *Comparison of the Airy Disc ($(2J_1(x)/x)^2$ - solid) and Gaussian ($I(x) = I_0 e^{-\ln 2 \frac{x^2}{1.62^2}}$ - dashed) intensity PDFs. The FWHM of the two distributions are the same, but the Gaussian PDF does not have the side lobes that are present in the PDFs of all optical systems with hard-edged apertures.*

The Airy Disc and the Gaussian approximation with the same FWHM are compared in Fig. 9.7. We see that the two are well matched except around the nulls of

the Airy Disc. The main error of using this Gaussian approximation is that our calculations don't capture the effects of the side lobes that are typically present in all optical systems that are not been specifically engineered to avoid them. This is not an insignificant error. In fact, dealing with the side lobes created by hard apertures is a major concern in many optical designs.

9.3.2 Image formation with finite Point Spread Functions

To understand the effect of a finite-sized PSF on image formation, we start by considering an idealized imaging system, in which the PSF is independent of the position of the point source in the object plane. Such systems are called shift invariant. In addition, we will assume that there is no distortion in the imaging system (other than that caused by the finite size of the PSF itself). That means that if the magnification of the projection systems is M, then the coordinates in the object and image planes are related as

$$x_i, y_i = Mx_o, My_o \tag{9.19}$$

where x_i,y_i are the image-plane coordinates, and x_o,y_o are the object-plane coordinates.

The object plane can be expressed as a sum (integral) of impulses that are weighted in amplitude and phase

$$E_o(x_o, y_o) = \iint_{x,y} E_o(x, y) \cdot \delta(x_o - x, y_o - y) dxdy \tag{9.20}$$

where $E_o(x,y)$ is the electric field distribution (amplitude and phase) in the object plane, and $\delta(x,y)$ is the two-dimensional impulse function. Imaging systems are linear, so we can find the field distribution, $E_i(x,y)$, in the imaging plane by summing the images of the individual impulses in the object plane, i.e. by summing the PSFs

$$\begin{aligned} E_i(x_i, y_i) &= \frac{1}{M} \iint_{x,y} E_o(x, y) \cdot PSF(x_i - Mx, y_i - My) d(Mx) \cdot d(My) \Rightarrow \\ E_i(x_i, y_i) &= M \iint_{x,y} E_o(x, y) \cdot PSF(x_i - Mx, y_i - My) dxdy \end{aligned} \tag{9.21}$$

The image is simply the object appropriately magnified and convolved with the PSF! The factor $1/M$ in the first line reflects the fact that the field strength is reduced as the square root of the area (the intensity is reduced linearly in the area). Equation 9.21 gives us a convenient mathematical tool for investigation of image formation with micromirror arrays.

9.3.3 Projection of a Gaussian Source

Now we consider the simple case of projection of a Gaussian source by an imaging system with a Gaussian PSF as depicted in Fig. 9.8. The questions we ask ourselves are: How does the PSF change the image of the Gaussian source, and how small of a source can the system resolve?

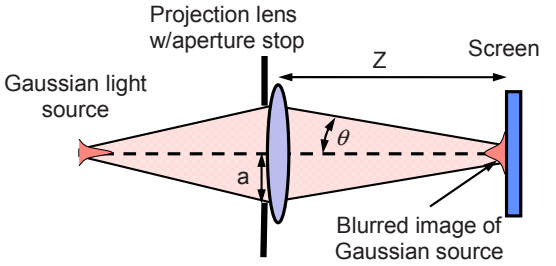


Figure 9.8 *Projection of a Gaussian source. The finite PSF of the projection system blurs the image of the Gaussian on the screen.*

Using the formalism developed in the preceding section, we write (for simplicity of notation we set the field at the center of the source $E_o(0,0)=1$ in these calculations)

$$E_i(x_i,y_i)=\frac{M}{\pi\cdot\omega_{PSF}^2}\iint_{x,y} e^{-\frac{x^2+y^2}{\omega_o^2}}\cdot e^{-\frac{(x_i-Mx)^2+(y_i-My)^2}{\omega_{PSF}^2}}\,dxdy \tag{9.22}$$

where ω_o is the Gaussian beam radius of the object and ω_{PSF} is the beam radius of the PSF. The scale factor in front of the integral normalizes the integral of the PSF to unity. This form shows how the Gaussian approximation to the PSF simplifies our calculations; the integral is separable in the two integration parameters and readily carried out

$$E_i(x_i,y_i)=\frac{M}{\pi\cdot\omega_{PSF}^2}\int_x e^{-\frac{x^2}{\omega_o^2}}\cdot e^{-\frac{(x_i-Mx)^2}{\omega_{PSF}^2}}\,dx\cdot\int_y e^{-\frac{y^2}{\omega_o^2}}\cdot e^{-\frac{(y_i-My)^2}{\omega_{PSF}^2}}\,dxdy \tag{9.23}$$

$$E_i(x_i,y_i)=\frac{M}{\pi\cdot\omega_{PSF}^2}\left\{\int_x e^{-\frac{\omega_{PSF}^2x^2+\omega_o^2x_i^2-\omega_o^2\,2x_iMx+\omega_o^2M^2x^2}{\omega_o^2\omega_{PSF}^2}}dx\cdot....\right\}=$$

$$\frac{M}{\pi\cdot\omega_{PSF}^2}\left\{\int_x e^{\begin{matrix}-\frac{\omega_{PSF}^2+\omega_o^2M^2}{\omega_{PSF}^2\omega_o^2}x^2+2\frac{\omega_o^2M}{\omega_o^2\omega_{PSF}^2}x_ix-\frac{\omega_o^4M^2}{\left(\omega_{PSF}^2+\omega_o^2M^2\right)\omega_o^2\omega_{PSF}^2}x_i^2\\ +\frac{\omega_o^4M^2}{\left(\omega_{PSF}^2+\omega_o^2M^2\right)\omega_o^2\omega_{PSF}^2}x_i^2-\frac{\omega_o^2}{\omega_o^2\omega_{PSF}^2}x_i^2\end{matrix}}dx\cdot....\right\} \tag{9.24}$$

$$E_i(x_i,y_i)=\frac{M}{\pi\cdot\omega_{PSF}^2}\left\{e^{-\frac{x_i^2}{\omega_{PSF}^2+\omega_o^2M^2}}\int_x e^{-\left(\frac{\sqrt{\omega_{PSF}^2+\omega_o^2M^2}}{\omega_{PSF}\cdot\omega_o}x-\frac{\omega_oM}{\sqrt{\omega_{PSF}^2+\omega_o^2M^2}\cdot\omega_{PSF}}x_i\right)^2}dx\cdot...\right\} \tag{9.25}$$

$$E_i(x_i,y_i)=\frac{M}{\pi\cdot\omega_{PSF}^2}\cdot$$

$$\left\{\begin{matrix}e^{-\frac{x_i^2}{\omega_{PSF}^2+\omega_o^2M^2}}\cdot\frac{\omega_{PSF}\cdot\omega_o}{\sqrt{\omega_{PSF}^2+\omega_o^2M^2}}\cdot\\ \int_x e^{-\left(\frac{\sqrt{\omega_{PSF}^2+\omega_o^2M^2}}{\omega_{PSF}\cdot\omega_o}x-\frac{\omega_oM}{\sqrt{\omega_{PSF}^2+\omega_o^2M^2}\cdot\omega_{PSF}}x_i\right)^2}d\left(\frac{\sqrt{\omega_{PSF}^2+\omega_o^2M^2}}{\omega_{PSF}\cdot\omega_o}x\right)\cdot...\end{matrix}\right\} \tag{9.26}$$

$$E_i(x_i,y_i)=\frac{M}{\pi\cdot\omega_{PSF}^2}\left\{e^{-\frac{x_i^2}{\omega_{PSF}^2+\omega_o^2M^2}}\cdot\frac{\omega_{PSF}\cdot\omega_o}{\sqrt{\omega_{PSF}^2+\omega_o^2M^2}}\cdot\sqrt{\pi}\cdot...\right\} \tag{9.27}$$

$$E_i(x_i,y_i)=e^{-\frac{x_i^2+y_i^2}{\omega_{PSF}^2+\omega_o^2M^2}}\cdot\frac{M}{\pi\cdot\omega_{PSF}^2}\frac{\omega_{PSF}^2\cdot\omega_o^2}{\omega_{PSF}^2+\omega_o^2M^2}\cdot\pi \tag{9.28}$$

$$E_i(x_i,y_i)=\frac{1}{M}e^{-\frac{x_i^2+y_i^2}{\omega_{PSF}^2+\omega_o^2M^2}}\cdot\frac{\omega_o^2M^2}{\omega_{PSF}^2+\omega_o^2M^2} \tag{9.29}$$

$$E_i(x_i,y_i)=\frac{1}{M}e^{-\frac{x_i^2+y_i^2}{(0.74\lambda N)^2+\omega_o^2M^2}}\cdot\frac{\omega_o^2M^2}{\omega_{PSF}^2+\omega_o^2M^2} \tag{9.30}$$

The projection is a Gaussian with a beam radius equal to the sum of the radii of the magnified source and the PSF! The factor $\frac{1}{M}$ shows that the field is reduced as the square root of the increase in area caused by the magnification, and the scale factor $\frac{\omega_o^2 M^2}{\omega_{PSF}^2 + \omega_o^2 M^2}$ simply reflects the fact that as the radius of the magnified source becomes less than the radius of the PSF, the image grows dimmer, because it is blurred, i.e. spread over a bigger area. This verifies what we would have guessed, namely that the image is determined by the smaller of the object and PSF. There is little to be gained by making the magnified source smaller than the PSF of the imaging system. Later we will see that for phase-modulating arrays we must match the size of the individual micromirrors to the projection-system PSF.

9.3.4 Projection of a Gaussian Micromirror

To apply the results of section 9.3.3 to micromirror arrays, we have to limit the maximum intensity to that of the illumination. That means that the scale factor in the above equation becomes important. To see how, let's consider a dark Gaussian pixel on a bright background. Following the derivation of section 9.3.4, we find

$$E_i(x_i, y_i) = \frac{M}{\pi \cdot \omega_{PSF}^2} \iint_{x,y} \left(1 - e^{-\frac{x^2+y^2}{\omega_o^2}} \right) \cdot e^{-\frac{(x_i - Mx)^2 + (y_i - My)^2}{\omega_{PSF}^2}} dxdy \tag{9.31}$$

$$E_i(x_i, y_i) = \frac{M}{\pi \cdot \omega_{PSF}^2} \left\{ \iint_{x,y} e^{-\frac{(x_i - Mx)^2 + (y_i - My)^2}{\omega_{PSF}^2}} dxdy - \iint_{x,y} e^{-\frac{x^2+y^2}{\omega_o^2}} \cdot e^{-\frac{(x_i - Mx)^2 + (y_i - My)^2}{\omega_{PSF}^2}} dxdy \right\} \tag{9.32}$$

$$E_i(x_i, y_i) = \frac{1}{M} \left(1 - e^{-\frac{x_i^2 + y_i^2}{\omega_{PSF}^2 + \omega_o^2 M^2}} \cdot \frac{\omega_o^2 M^2}{\omega_{PSF}^2 + \omega_o^2 M^2} \right) \tag{9.33}$$

This equation reveals the importance of the scale factor! It reduces the contrast of the dark pixel on the bright background by increasing the minimum field of the pixel.

The image intensity as a function of the distance from the image center is shown in Fig. 9.9. The parameters chosen for these plot are $\omega_{PSF}=1$, $\omega_o M=4$ (dashed), $\omega_o M=2$ (dotted), $\omega_o M=1$ (dash-dotted), and $\omega_o M=0.5$ (dashed-dotted). For comparison, we also show the intensity PSF, e^{-2x^2/ω_{PSF}^2}. The pre-factor $1/M$ is omitted in these plots to simplify comparison.

The graphs show clearly that to create a truly dark image, the magnified pixel size must be very much larger than the PSF of the projection system. Dark pixels with radii smaller than the radius of the PSF will indeed project as Gaussian distributions with close to the beam radius of the PSF, but at the cost of significantly reduced contrast.

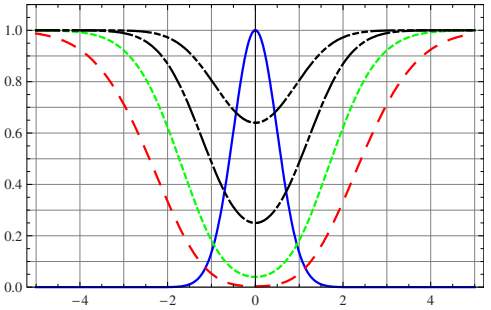


Figure 9.9 *Projection of a dark Gaussian pixel on a bright background for object beam radii of $\omega_0 M=4$ (dashed), $\omega_0 M=2$ (dotted), $\omega_0 M=1$ (dash-dotted), and $\omega_0 M=0.5$ (dash-double-dotted). The solid line shows the intensity PSF, e^{-2x^2/ω_{PSF}^2} where $\omega_{PSF}=1$.*

Our discussion of the PSF demonstrates that it sets a limit on how small a micromirror should be for optimum performance. In section 9.2 we saw that the mirror size was limited by the micromirror modulation contrast. Whether it is the PSF or the modulation contrast that ultimately limits how small we can make the mirror in a given micromirror array depends on the application.

In a projection display, which is the most common application of the DLP technology, the magnification M is large. The magnified source beam radius is typically much larger than ω_{PSF}, so the PSF plays a minor role in determining the size of the projected pixels on the screen. For this type of application, the mirrors size is limited by the modulation contrast considerations of section 9.2.

The situation is the opposite for systems that create microscopic images of micromirror arrays. Examples of such systems are maskless lithography machines, in which a pattern created by a micromirror array is projected onto a photoresist-covered wafer for printing of ICs and MEMS. Here the micromirror image is demagnified, i.e. $M<1$. Consequently, the PSF tend to dominate and it becomes the factor that limits the mirror size.

9.3.5 Projection of a 1-D Gaussian Source

It is also instructive, as well as useful for the comparison of amplitude and phase modulation that we will carry out later, to briefly consider a 1-D Gaussian source of the form

$$\vec{E}(x) = \vec{E}_0 e^{-\frac{x^2}{\omega_{PSF}^2}} \tag{9.34}$$

where again we use the definition we $\omega_{PSF} = \frac{1.62}{\ln 2 \cdot \pi} \lambda N \approx 0.74 \lambda N$. Using the same approach as in the preceding section, we find

$$E_i(x_i, y_i) = \frac{M}{\pi \cdot \omega_{PSF}^2} \iint_{x,y} e^{-\frac{x^2}{\omega_o^2}} \cdot e^{-\frac{(x_i - Mx)^2 + (y_i - My)^2}{\omega_{PSF}^2}} dxdy \tag{9.35}$$

$$E_i(x_i, y_i) = \frac{1}{\sqrt{\pi} \cdot \omega_{PSF}} \int_x e^{-\frac{x^2}{\omega_o^2}} \cdot e^{-\frac{(x_i - Mx)^2}{\omega_{PSF}^2}} dx \tag{9.36}$$

$$E_i(x_i, y_i) = \frac{1}{M} \cdot e^{-\frac{x_i^2}{\omega_{PSF}^2 + \omega_o^2 M^2}} \cdot \frac{\omega_o M}{\sqrt{\omega_{PSF}^2 + \omega_o^2 M^2}} \tag{9.37}$$

Due to the fact that the integral is separable, the 1-D and 2-D Gaussian give essentially the same result; the beam radius of the image is the sum of the object and PSF radii. There is a difference in the scale factor, but that is simply due to the source is independent of the *y* coordinate.

Extending this to dark lines on a bright background we find

$$E_i(x_i, y_i) = \frac{1}{M} \left(1 - e^{-\frac{x_i^2}{\omega_{PSF}^2 + \omega_o^2 M^2}} \cdot \frac{\omega_o M}{\sqrt{\omega_{PSF}^2 + \omega_o^2 M^2}} \right) \tag{9.38}$$

Just as the dark pixel, the dark line has reduced contrast, caused by the scale factor, when the magnified image approaches the radius of the PSF.

9.4 Micromirrors with Phase Modulation

So far we have considered rotating micromirrors that modulate the projected light intensity by directing light outside the aperture of the projection optics. This

should be considered a type of amplitude modulation, because it is primarily the amplitude of the reflected light that is being controlled by the position of the micromirror. It is also possible to create micromirrors that primarily control the phase of the reflected light. For example a micromirror that moves vertically in a piston-like motion will delay or advance the phase of the light that is reflected off them compared to light that is reflected off the neighboring mirrors.

One typical example is shown in Fig. 9.10. Here one mirror is pulled towards the substrate by a distance that corresponds to a quarter of a wavelength of the incident light. The light that is reflected from this micromirror is therefore exactly out of phase (i.e. having a phase that is advanced by a total of π) with the light that is reflected from other parts of the array. Depending on the size of the mirror and the projection system that is used, either the edges or the whole mirror will appear dark in projection. The gaps between the mirrors are assumed to be non-reflecting, so their only effect is to set up a weak amplitude grating that is of negligible consequence for the projection of the array.

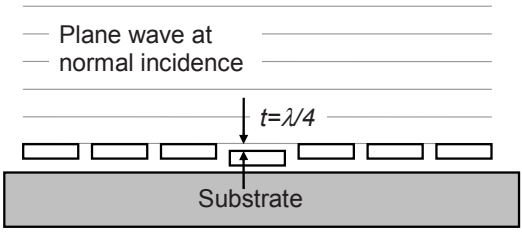


Figure 9.10 *Phase shifting micromirror array. Each mirror can translate vertically to set up a local variation in the phase of the reflected light. In the configuration shown there is one micromirror that is translated downwards by a quarter of a wavelength, so that the path length of, the light reflected from that mirror has an extra phase shift of π radians compared to the light that is reflected from the rest of the array.*

Phase shifting mirrors behave very differently than amplitude-modulating mirrors. In this section we will develop the tools and the insight required to compare and contrast phase shifting micromirrors to the rotating micromirrors of section 9.3. We will find that phase modulation is more complicated than amplitude modulation, but also that it offers significant advantages under certain conditions.

9.4.1 Projection of a Phase Step

From sections 9.3.3-9.3.5, it is tempting to draw the conclusion that a projected feature with full contrast cannot have a beam radius that is comparable to that of the PSF. That is, however, not true. Consider a simple phase step along the y-axis, i.e. a function that has a constant unity absolute value and a phase of $e^{j\theta_1}$ for

$x<0$ and $e^{j\theta_2}$ for $x>0$. This object field results in the following optical field in the image plane

$$E_i(x_i,y_i)=\frac{M}{\pi\cdot\omega_{PSF}^2}\left\{\begin{array}{l}\int_{-\infty}^{0}\int_{-\infty}^{\infty}e^{j\theta_1}e^{-\frac{(x_i-Mx)^2+(y_i-My)^2}{\omega_{PSF}^2}}dxdy\\ +\int_{0}^{\infty}\int_{-\infty}^{\infty}e^{j\theta_2}e^{-\frac{(x_i-Mx)^2+(y_i-My)^2}{\omega_{PSF}^2}}dxdy\end{array}\right\} \tag{9.39}$$

Again we benefit from the fact that the Gaussian distribution is separable in Cartesian coordinates, and write

$$E_i(x_i,y_i)=\frac{1}{\sqrt{\pi}\cdot\omega_{PSF}}\left\{e^{j\theta_1}\int_{-\infty}^{0}e^{-\frac{(Mx-x_i)^2}{\omega_{PSF}^2}}dx+e^{j\theta_2}\int_{0}^{\infty}e^{-\frac{(Mx-x_i)^2}{\omega_{PSF}^2}}dx\right\} \tag{9.40}$$

$$E_i(x_i,y_i)=\frac{1}{M\sqrt{\pi}}\cdot\left\{e^{j\theta_1}\int_{-\infty}^{-\frac{x_i}{\omega_{PSF}}}e^{-u^2}du+e^{j\theta_2}\int_{-\frac{x_i}{\omega_{PSF}}}^{\infty}e^{-u^2}du\right\} \tag{9.41}$$

$$E_i(x_i,y_i)=\frac{1}{M\sqrt{\pi}}\cdot\left\{e^{j\theta_1}\left[\frac{\sqrt{\pi}}{2}-\int_{0}^{\frac{x_i}{\omega_{PSF}}}e^{-u^2}du\right]+e^{j\theta_2}\left[\frac{\sqrt{\pi}}{2}+\int_{0}^{\frac{x_i}{\omega_{PSF}}}e^{-u^2}du\right]\right\} \tag{9.42}$$

$$E_i(x_i,y_i)=\frac{1}{2M}\cdot\left\{e^{j\theta_1}\left[1-Erf\left(\frac{x_i}{\omega_{PSF}}\right)\right]+e^{j\theta_2}\left[1+Erf\left(\frac{x_i}{\omega_{PSF}}\right)\right]\right\} \tag{9.43}$$

$$E_i(x_i,y_i)=\frac{1}{2M}\cdot\left\{\left[e^{j\theta_1}+e^{j\theta_2}\right]+\left[e^{j\theta_2}-e^{j\theta_1}\right]Erf\left(\frac{x_i}{\omega_{PSF}}\right)\right\} \tag{9.44}$$

For $e^{j\theta_1}=-1$ and $e^{j\theta_2}=1$, i.e. a phase step where the field is uniformly negative for negative x values and positive for positive x values, this evaluates to

$$E_i(x_i, y_i) = \frac{1}{M} \cdot \frac{2}{\sqrt{\pi}} \int_0^{\frac{x_i}{\omega_{PSF}}} e^{-u^2} du = \frac{1}{M} \cdot Erf\left(\frac{x_i}{\omega_{PSF}}\right) \tag{9.45}$$

and the intensity becomes

$$I_i(x_i, y_i) = \left(\frac{1}{M} \cdot Erf\left(\frac{x_i}{\omega_{PSF}}\right)\right)^2 \tag{9.46}$$

where $Erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-u^2} du$ is the error function.

This closed-form solution to the intensity projection of a phase step is plotted as the solid line in Fig. 9.11 together with projections of dark Gaussian lines with $\omega_o M=4$ (dashed), $\omega_o M=2$ (dotted), $\omega_o M=1$ (dash-dotted), and $\omega_o M=0.5$ (dashed-dotted). The beam radius of the PSF is again set to unity ($\omega_{PSF}=1$). Note that these projected Gaussian lines have better contrast than the Gaussian pixels of Fig. 9.9. The reason is that the scale factor is larger for lines than for pixels.

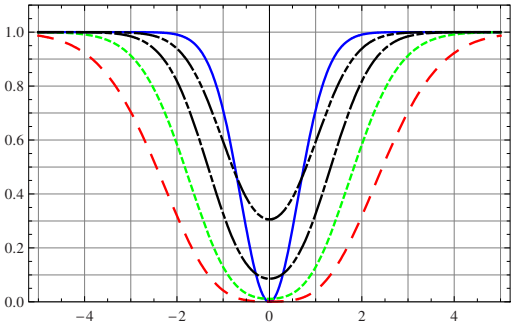


Figure 9.11 *Projection of a phase step (solid) compared to dark Gaussian lines on a bright background for object beam radii of $\omega_0 M=4$ (dashed), $\omega_0 M=2$ (dotted), $\omega_0 M=1$ (dash-dotted), and $\omega_0 M=0.5$ (dash-double-dotted). The PSF has a beam radius of unity ($\omega_{PSF}=1$).*

Just like the projections of the dark lines, the phase step projection has an intensity minimum exactly at the position of the step. As opposed to the dark-line projections, the phase step projection reaches zero intensity at its center. The existence of this null, that we have found from our mathematical modeling, could have been

predicted from the observation that the phase of the optical field is undefined at the phase step, so the field has to go to zero at that point to avoid ambiguity[f].

Figure 9.11 shows that the projection of the phase step is significantly narrower than the projection of any dark line. If we compare the phase step to lines that give close to unity contrast, we see that the dark lines are about twice as wide. In fact, the phase-step projection is not very much wider than the intensity PSF itself as shown in Fig. 9.12.

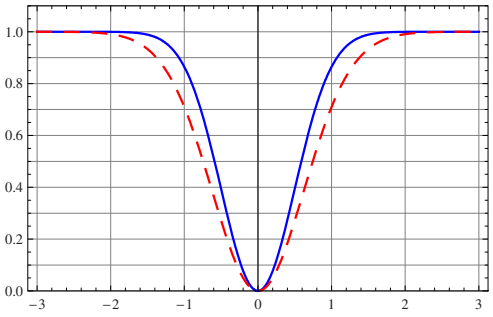


Figure 9.12 *Comparison of projection of a phase step (dashed) to the PSF (ω_{PSF}=1). To simplify the comparison, the function 1-PFS is shown (solid).*

9.4.2 Projection of a Phase Modulated Line

It is straightforward to extend our treatment of a single phase step to multiple phase steps. Consider again an object of uniform intensity, but this time with a phase of θ_1 for $x<-w/2$, θ_2 for $-w/2<x<w/2$, and θ_3 for $x>-w/2$. This is the phase distribution we would get in reflection from a micromirror array where one line of mirrors is displaced with respect to the mirrors on either side. To find the projection of this phase-shifted line of width w, we adopt the same approach as we used in section 9.4.1 to analyze the projection of a single line phase shift.

[f] The same type of phase uncertainty and vanishing of the optical field is associated with an optical vortex, which is the point, or zero-dimensional, equivalent to the one-dimensional phase step that we have considered.

$$E_i(x_i,y_i)=\frac{M}{\pi\cdot\omega_{PSF}^2}\left\{\int_{-\infty}^{-w/2}\int_{-\infty}^{\infty}e^{j\theta_1}e^{-\frac{(x_i-Mx)^2+(y_i-My)^2}{\omega_{PSF}^2}}dxdy+\right.$$
$$\left.\int_{-w/2}^{w/2}\int_{-\infty}^{\infty}e^{j\theta_2}e^{-\frac{(x_i-Mx)^2+(y_i-My)^2}{\omega_{PSF}^2}}dxdy+\int_{w/2}^{\infty}\int_{-\infty}^{\infty}e^{j\theta_3}e^{-\frac{(x_i-Mx)^2+(y_i-My)^2}{\omega_{PSF}^2}}dxdy\right\} \tag{9.47}$$

$$E_i(x_i,y_i)=\frac{1}{\sqrt{\pi}\cdot\omega_{PSF}}$$
$$\left\{\int_{-\infty}^{-w/2}e^{j\theta_1}e^{-\frac{(x_i-Mx)^2}{\omega_{PSF}^2}}dx+\int_{-w/2}^{w/2}e^{j\theta_2}e^{-\frac{(x_i-Mx)^2}{\omega_{PSF}^2}}dx+\int_{w/2}^{\infty}e^{j\theta_3}e^{-\frac{(x_i-Mx)^2}{\omega_{PSF}^2}}dx\right\} \tag{9.48}$$

$$E_i(x_i,y_i)=\frac{1}{M\sqrt{\pi}}\left\{e^{j\theta_1}\int_{-\infty}^{\frac{-Mw/2-x_i}{\omega_{PSF}}}e^{-u^2}du+e^{j\theta_2}\int_{\frac{-Mw/2-x_i}{\omega_{PSF}}}^{\frac{Mw/2-x_i}{\omega_{PSF}}}e^{-u^2}du+e^{j\theta_3}\int_{\frac{Mw/2-x_i}{\omega_{PSF}}}^{\infty}e^{-u^2}du\right\} \tag{9.49}$$

$$E_i(x_i,y_i)=\frac{1}{M\sqrt{\pi}}\left\{\begin{array}{l}\frac{\sqrt{\pi}}{2}e^{j\theta_1}-e^{j\theta_1}\int_0^{\frac{x_i+Mw/2}{\omega_{PSF}}}e^{-u^2}du-e^{j\theta_2}\int_0^{\frac{x_i-Mw/2}{\omega_{PSF}}}e^{-u^2}du\\ +e^{j\theta_2}\int_0^{\frac{x_i+Mw/2}{\omega_{PSF}}}e^{-u^2}du+\frac{\sqrt{\pi}}{2}e^{j\theta_3}+e^{j\theta_3}\int_0^{\frac{x_i-Mw/2}{\omega_{PSF}}}e^{-u^2}du\end{array}\right\} \tag{9.50}$$

$$E_i(x_i,y_i)=\frac{1}{2M}\left[\begin{array}{l}\left(e^{j\theta_1}+e^{j\theta_3}\right)+\left(e^{j\theta_2}-e^{j\theta_1}\right)Erf\left(\frac{x_i+Mw/2}{\omega_{PSF}}\right)\\ +\left(e^{j\theta_3}-e^{j\theta_2}\right)Erf\left(\frac{x_i-Mw/2}{\omega_{PSF}}\right)\end{array}\right] \tag{9.51}$$

This formula shows that even a single phase-modulated line results in a quite complex intensity distribution, given by $I_i(x_i,y_i)=E_i(x_i,y_i)E_i^*(x_i,y_i)$. To get an overview we evaluate the expression for the special case $\theta_1=\theta_3=0$ and $\theta_2=\pi$. This is the phase distribution we would get in reflection from a micromirror array where one line of mirrors is displaced by a quarter wavelength, which means a

path length difference of a half wavelength in reflection. The projected field and intensity from such a phase-modulated line is given by

$$E_i(x_i, y_i) = \frac{1}{M}\left[1 - Erf\left(\frac{x_i + Mw/2}{\omega_{PSF}}\right) + Erf\left(\frac{x_i - Mw/2}{\omega_{PSF}}\right)\right] \tag{9.52}$$

$$I_i(x_i, y_i) = \frac{1}{M^2}\left[1 - Erf\left(\frac{x_i + Mw/2}{\omega_{PSF}}\right) + Erf\left(\frac{x_i - Mw/2}{\omega_{PSF}}\right)\right]^2 \tag{9.53}$$

Figure 9.13 compares this intensity for four different values of the magnified mirror width with the projection of a π phase step. As expected, we see that only the edges of wide lines project as dark, while the centers of the wider lines project as bright independent of their phase shift. As the line width is reduced, the dark edges merge, and the whole line appear dark. Further reduction in line width results in an associated reduction in the projection width, but also in reduced contrast as the edges get too close to each other.

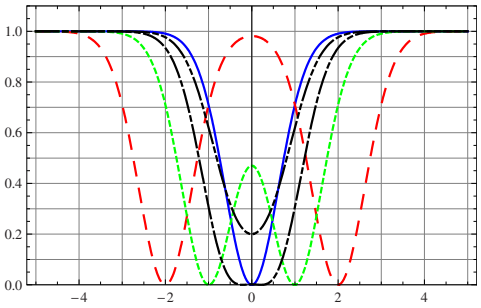


Figure 9.13 *Comparison of the projection of a π phase step to projections of π phase modulated lines of different magnified widths; $Mw=4\omega_{PSF}$ (dashed), $Mw=2\omega_{PSF}$ (dotted), $Mw=\omega_{PSF}$ (dash-dotted), and $Mw=0.5\omega_{PSF}$ (dash-double-dotted) For simplicity, the beam radius of the PSF is set to unity ($\omega_{PSF}=1$) in these plots.*

The optimum width of the mirrors is approximately equal to the Gaussian beam diameter of the PSF ($2\omega_{PSF}$). This is very different from what we learned about amplitude-modulating micromirrors; If a rotating, amplitude-modulating mirror has sufficient rotation, the whole mirror projects dark, so there is no need to make sure that the mirror is not too large for the projection system.

9.4.3 Sub-Pixel Shifts in Phase-Modulated Micromirror arrays

The complexity of the projected intensity of a phase-modulated micromirror array arises from the interference between the fields from the different mirrors, but this effect also gives phase-modulated mirror arrays some unique and powerful characteristics. Consider as an example a micromirror array that is set up with N phase modulated rows such that the phase is θ_0 for $x<0$, θ_1 for $0<x<w$, θ_2 for $w<x<2w$ and so on up to θ_N for $(N-1)w<x<Nw$ and θ_{N+1} for $x>Nw$. Following the same procedure as in sections 9.4.1 and 9.4.2, it is straightforward to show that the projected field is

$$E_i(x_i,y_i)=\frac{1}{2M}\left\{\begin{array}{l}\left(e^{j\theta_0}+e^{j\theta_{N+1}}\right)+\left(e^{j\theta_1}-e^{j\theta_0}\right)Erf\left(\dfrac{x_i}{\omega_{PSF}}\right)\\ +\left(e^{j\theta_2}-e^{j\theta_1}\right)Erf\left(\dfrac{x_i-Mw}{\omega_{PSF}}\right)\\ +.....+\left(e^{j\theta_{N+1}}-e^{j\theta_N}\right)Erf\left(\dfrac{x_i-N\cdot Mw}{\omega_{PSF}}\right)\end{array}\right\} \tag{9.54}$$

If we restrict the phase differences to 0 and 1, then the intensity that results from this distribution has minima at the points (lines) between the mirrors as in Fig. 9.13. If, however, we chose the phases correctly, then we can place the intensity minima at intermediate locations, i.e. it is possible to control the position of spatial features with an accuracy that is better than the size of the micromirrors [10,11]. This is of course of great utility in applications where the *positioning* of image features must be controlled with finer resolution than the images themselves.

9.5 Projection of Micromirrors through Hard Apertures

Our treatment of imaging through a Gaussian PSF gave us valuable insight into the limits of micromirror design that can be used to analyze a variety of applications. The Gaussian approximation to the PSF simplifies the analysis and allows us to derive closed-form solutions. We must emphasize, however, that certain aspects of our modeling is idealized to the point where some important effects are lost. To get an appreciation for this effect, we go back to the general equation giving the object field as a convolution of the object field and the PSF

$$E_i(x_i,y_i)=$$
$$\frac{M}{4\pi}\iint_{x,y}E_o(x,y)\cdot\frac{2J_1\left(\dfrac{\pi}{\lambda N}\cdot\sqrt{(x_i-Mx)^2+(y_i-My)^2}\right)}{\dfrac{\pi}{\lambda N}\cdot\sqrt{(x_i-Mx)^2+(y_i-My)^2}}dxdy \tag{9.55}$$

The scale factor in front of the integral normalizes the integral of the PSF to unity. As before, we are interested in amplitude and phase modulated lines, so we integrate over the y variable

$$E_i(x_i) = \frac{1}{\pi}\int_x E_o(x)\cdot Sinc\left[\frac{M\pi}{\lambda N}\cdot(x_i - Mx)\right]dx \tag{9.56}$$

$$I_i(x_i) = \left(\frac{1}{\pi}\int_x E_o(x)\cdot Sinc\left[\frac{M\pi}{\lambda N}\cdot(x_i - Mx)\right]dx\right)^2 \tag{9.57}$$

where $Sinc[z] = Sin[z]/z$. This image-plane intensity is plotted in Fig. 9.14 for amplitude-modulated dark lines (solid) and for phase-modulated lines with π phase shifts (dash-dotted). The magnified widths are $Mw = 4\lambda N/\pi$ (a), $Mw = 2\lambda N/\pi$ (b), $Mw = \lambda N/\pi$ (c), and $Mw = 0.5\lambda N/\pi$ (d), and for simplicity we set $\lambda N/\pi = 1$.

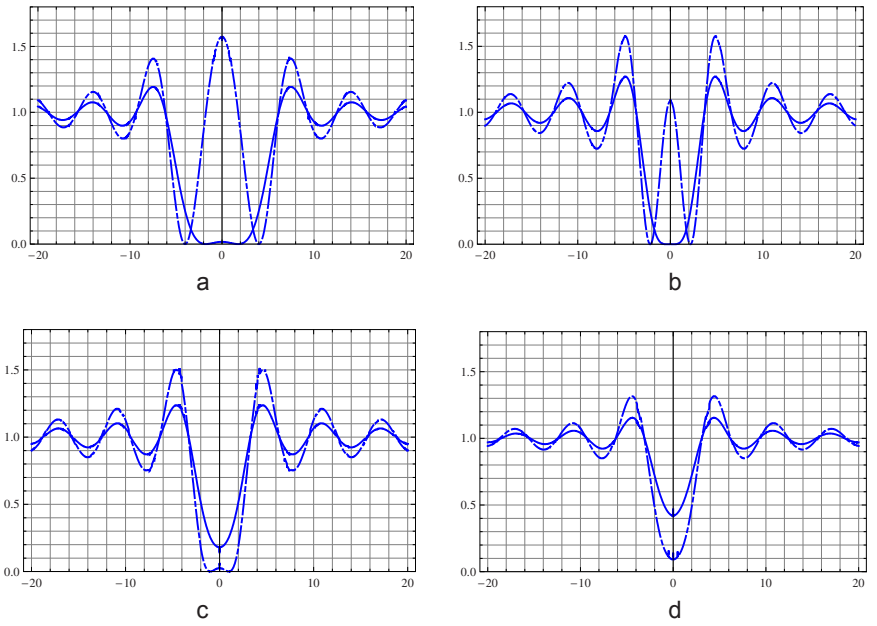


Figure 9.14 *Image intensity of a single dark line on a bright background projected through a system with an Airy Disc PSF. The plots show both amplitude-modulated dark lines (solid) and phase-modulated lines with π phase shifts (dash-dotted). The magnified widths are $Mw=4\lambda N/\pi$ (a), $Mw=2\lambda N/\pi$ (b), $Mw=\lambda N/\pi$ (c), and $Mw=0.5\lambda N/\pi$ (d), and $\lambda N/\pi=1$.*

The plots demonstrate that the Airy-Disc PSF, or more accurately the *Sinc PSF* because we are considering lines, that results from hard-edged, circular apertures, create projections that in most respects are very similar to the images created by the Gaussian PSF. The interdependence of the size of the magnified object and the PSF are the same, and the relationship between the phase modulated and the amplitude modulated lines are the same. The biggest difference is the side lobes of the projections. These are caused by the side lobes in the Airy Disc, and create problems in many applications.

Figure 9.14 also shows that the projections of the phase-modulated lines have larger side lobes than the amplitude-modulated images. That is due to the fact that in the phase modulated images, the side lobes are interference between the light originating outside and inside the phase-modulated line. In the amplitude-modulated images, only the by the light outside the line participates, because there is no light originating from inside the line.

9.6 Adaptive Optics

In the treatment of phase-modulating micromirror arrays in sections 9.4 and 9.5 we implicitly assumed that the purpose of the arrays were to create images suitable for projection. In other word, the ultimate objective was to form spatially varying amplitudes, and the phase modulation was a means to an end, rather than an end to itself. In Adaptive Optics the situation is diametrically opposite. Here spatially varying phase modulation is the desired outcome, and any associated amplitude modulation is at best a nuisance and at worst a serious source of errors. The purpose of the micromirror arrays in adaptive-optics systems is to compensate for arbitrary phase errors in an incoming optical wavefront. The phase errors might be caused by density variations in the propagation medium (e.g. turbulence in the atmosphere), or by flaws in the optical system caused by temperature variations or other time-varying environmental influences.

A typical Adaptive Optics system is shown in Fig. 9.15. An incoming wave front that is corrupted by transmission through a turbulent medium is collimated onto a phase-modulating mirror array and subsequently focused onto a camera. Before reaching the camera the light from a known reference is separated from the light from the target and directed towards a wave-front analyzer. By analyzing the aberrated light from the known reference, the phase corruption of the transmission medium can be determined and corrected by the phase-modulating mirror. If, for example, the reference is a point source far from the imaging system, then the collimated beam should have flat wave fronts, so the adaptive-optics mirror is deformed such that the wave front of the reference light coming off the mirror is indeed flat.

The assumption is that the light from the target has been transmitted through the same path in the same medium so that it has accumulated the same phase distortions. After coming off the phase-modulating mirror, the light from the target has the phase fronts it would have had if the transmission medium had been homogeneous and therefore without phase distortion. The system is dynamically updating the settings of the adaptive optics mirror as the phase distortion of the turbulent transmission medium changes over time.

Adaptive Optics is used in three very important application areas. The original motivation for the development of the technology was improvement of astronomical observations. By removing the aberrations caused by transmissions through atmosphere, resolution can be substantially improved. The reference signal is either a bright star in the vicinity of the constellation under observation, or it is an artificial laser guide star that is created by energizing the sodium atoms in the mesosphere with a laser beam at 589.2 nm wavelength. The sodium atoms re-emit the absorbed radiation, creating an artificial star that can be placed in any desired location, making possible aberration-free observations in all directions. Existing Adaptive Optics for astronomical observations use large deformable mirrors actuated by bulk piezoelectric transducers. MEMS arrays are compelling for this application, because of their relatively low cost and the large number of mirrors that can be integrated into a single array.

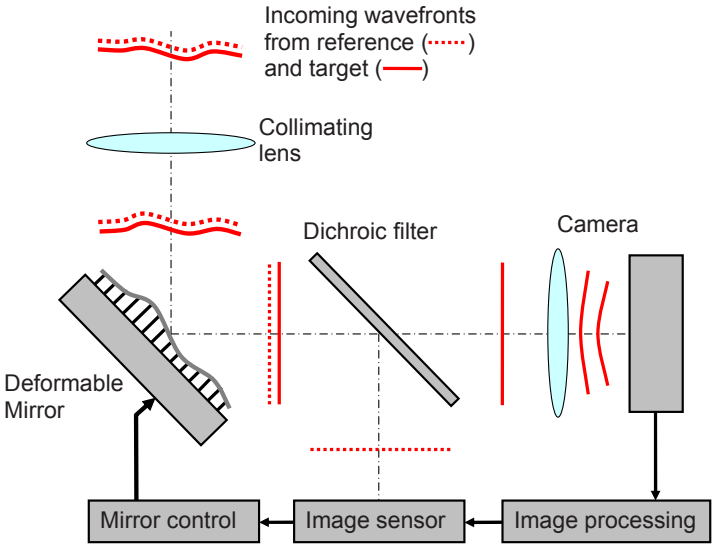


Figure 9.15 *Adaptive Optics system designed to receive an optical signal from a distant target. The signal is distorted by inhomogenities of the transmission medium, but the aberrations of the wave fronts are compensated by a deformable mirror. Light from a reference is used to determine the settings of the deformable mirror.*

Vision science is the other driver of Adaptive-Optics technology. The vitreous body (vitreous humor) of the human eye contains inhomogeneities that make it difficult to image the retina with a resolution better than the size of the rods and cones. By compensating the phase aberrations created by these inhomogeneities, one can obtain a more detailed picture, useful for diagnosis of retinal disease. The small size and large number of degrees of freedom make MEMS arrays ideal for this application.

Adaptive Optics is also very useful for cleaning up optical beams for free-space (as opposed to guided wave) communication through the atmosphere. In these systems, the target can typically be thought of as a point source, so the target is its own reference. Communication over long distances through the lower parts of the atmosphere is very demanding because of the large and rapidly varying phase aberrations. The small size and associated high speed of MEMS phase-modulating arrays are therefore compelling reasons to consider them for free-space communication.

9.6.1 Micromirror Arrays for Adaptive Optics

The deformable mirrors of Adaptive Optics are quite different from arrays used for image projection. Phase distortions caused by transmission typically has longer-range variations than the images that we are trying to collect, so spatial resolution is not the primary concern in the design of adaptive optics mirrors. In fact, the mirrors typically have magnified sizes that are larger than the PSF of the imaging system. This makes technical sense in systems that projects images with fine detail in the presence phase distortions that are slow functions of position in the image.

In such systems the phase singularities associated with the transitions between mirrors lead to unwanted amplitude modulation. We found in section 9.4 that associated with each phase discontinuity there is a step (error function) in the amplitude distribution with a height determined by the phase difference across the boundary. Projection systems can utilize these amplitude steps, and it is an advantage that their heights are independent of the optical projection systems (the widths of the steps are proportional to the width of the PSF). In Adaptive Optics, however, this amplitude modulation is a nuisance.

From an optics point of view it is therefore advantageous to use continuous adaptive-optics mirrors like the one shown in Fig. 9.16. This type of mirror avoids phase discontinuities that appear at the boundaries between individual elements of segmented mirrors like the one shown in Fig. 9.10. Continuous mirrors are therefore standard in bulk implementations of adaptive optics.

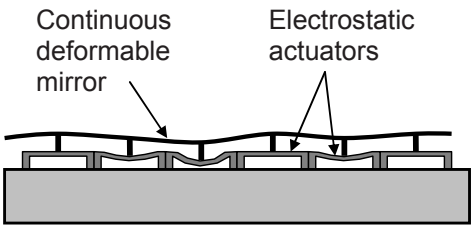


Figure 9.16 *Schematic of a continuous adaptive-optics mirror that can be deformed into the desired shape by electrostatic actuators. The continuous mirror avoids the phase singularities associated with the transitions between mirrors in segmented structures.*

The problem with MEMS implementations of continuous mirrors is that the position of each actuator is strongly affected by the setting of other actuators in the array. If we consider a single element in Fig. 9.16, we see that the continuous mirror connects the actuators such that each actuator is subject to forces from its neighbors. This effect is particularly pronounced in arrays where the actuators can only exert a downward force, i.e. it is not possible to use electrostatics to push a given part of the mirror up. To get good spatial control of the mirror with such arrays, the forces of the mirror surface on each actuator has to be at most comparable to, and preferably smaller than, the spring forces developed in the actuator itself. These are again limited by the available electrostatic forces. The result is that the relatively weak electrostatic forces set an upper limit on the bending strength of the reflecting diaphragm, resulting in weak mirrors with long-term reliability problems. This problem is not present in traditional implementations that are not based on MEMS, because high-force actuators are available, so the deformable mirrors can have high bending stiffness.

Figure 9.17 shows an individual micromirror that is better suited to MEMS implementations. The mirror has three degrees of freedom of motion; piston motion perpendicularly to the mirror plane and rotation on two orthogonal axes in the mirror plane. These three degrees of freedom of motion are referred to as Tip-Tilt-Piston motion. This complex mechanical functionality allows the array to form a linear approximation to a desired phase profile, so that phase discontinuities can be avoided, or at least significantly reduced. This type of array can be thought of as a general diffractive optical element that can be configured for a wide range of different applications in addition to adaptive optics.

The down side of the tip-tilt-piston design is that it is complex, and therefore more difficult and costly to fabricate and operate. In Optical Micro Systems, as in MEMS in general, it is often advantageous to sacrifice functionality for simplicity of fabrication. So even though the tip-tilt-piston functions better from an optics point of view, a simple piston-only mirror is often the better choice for practical systems.

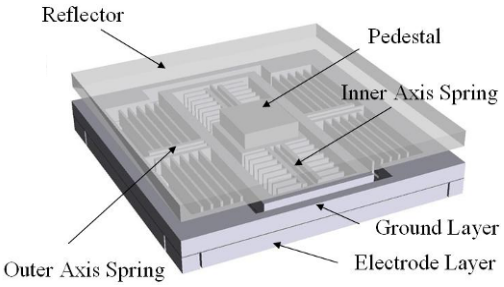


Figure 9.17 *Single pixel of a tip-tilt-piston micromirror adaptive-optics array used to control the phase front of optical beams [12]. The reflector is shown as transparent to allow an unobstructed view of the mirror architecture.*

Figure 9.18 shows an array of piston-motion, metal micromirrors that are directly integrated with CMOS circuitry. Large arrays of micromirrors require integrated electronics for signal conditioning and multiplexing, so direct integration on CMOS is an enabling advantage. The problem of phase discontinuities and the associated amplitude modulation is mitigated simply by decreasing the mirror size and increasing their numbers. Smaller mirrors means smaller phase step at each mirror boundary. The associated amplitude modulation also has higher spatial frequencies that are easier to remove by spatial filtering. Small size and CMOS compatibility that enable arrays with large numbers of mirrors therefore make MEMS the ideal technology for piston-only, segmented deformable mirrors.

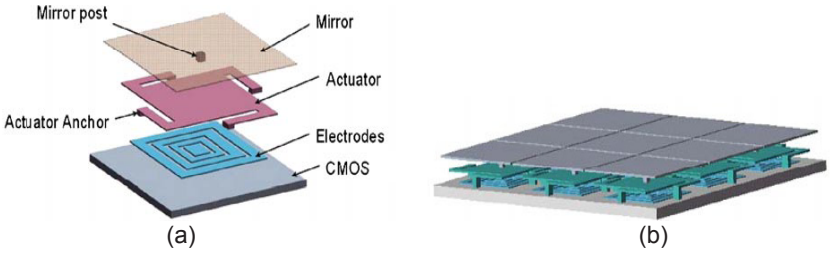


Figure 9.18 *Expanded individual mirror (a) and 3 by 3 segment of a simple piston-only, adaptive-optics metal micromirror array integrated on CMOS. Reprinted from [13] with permission.*

9.7 Phase vs. Amplitude Modulation

The main results of sections 9.4 and 9.5 are the formulas for the narrowest projected dark lines created by micromirror arrays. If we assume that the imaging systems has a Gaussian PSF, then a dark line created by amplitude modulation has a projected intensity given by (Eq. 9.38)

$$I_i(x_i, y_i) = \frac{1}{M^2}\left(1 - e^{-\frac{x_i^2}{\omega_{PSF}^2 + \omega_o^2 M^2}} \cdot \frac{\omega_o M}{\sqrt{\omega_{PSF}^2 + \omega_o^2 M^2}}\right)^2 \tag{9.58}$$

where I_0 is the source intensity on each side of the dark line, M is the linear magnification factor, ω_{PSF} is the beam radius of the Gaussian PSF, and ω_0 is the beam radius of the Gaussian dark line. The corresponding expression for a phase step, i.e. the narrowest phase-modulated feature is (Eq. 9.46)

$$I_i(x_i, y_i) = \frac{1}{M^2} Erf\left(\frac{x_i}{\omega_{PSF}}\right)^2 \tag{9.59}$$

Comparing these two formulas we found that the projected phase step gives a dark line of about half the width of an amplitude modulated line of good contrast. This factor-of-two improvement in spatial resolution of PM over AM is well known and exploited in a number of technologically important fields, including phase-contrast microscopy, phase-mask lithography, and optical data storage.

Equally important for implementations of microoptical systems is the fact that phase modulation requires less mechanical motion than what is needed for amplitude modulation. In section 9.2.1 we found that rotating mirrors needed to move their extreme ends on the order of a couple of wavelengths to achieve good contrast. That corresponds to an average motion of about one wavelength. Phase-modulation mirrors, on the other hand, create high-contrast images with only a quarter wavelength of motion. This difference is very significant, because most miniaturized systems are limited by the actuation technology, so reducing the required range of motion simplifies implementation.

As engineers we know that nothing comes for free, so it is no surprise that phase modulation also has some very serious drawbacks. Chief among these is dispersion. The method we use to create phase delay in micromirror arrays is to create physical path length differences as when we move a mirror vertically with respect to its neighbors in an array. A given path length difference, ΔL, results a phase difference of $\frac{2\pi \cdot \Delta L}{\lambda}$, where λ is the wavelength of light. This wavelength dependence is of little consequence in monochromatic applications, but complicates the use of phase modulating micromirrors for white-light or broad-band systems.

The numerical calculations in section 9.5 show that phase modulation creates larger side lobes than amplitude modulation does. This is a complicating issue in many systems. As we have seen, it can be dealt with by grading the transmission through the apertures to create Gaussian-like PSFs, but that is a non-standard, and therefore expensive, solution.

From section 9.4 we make the observation that it is phase differences that create features in the image, i.e. in phase modulated mirror arrays it is really the mirror edges that are important. That means that the mirrors have to be designed to work with a specific PSF, and that neighboring pixels will influence each other more strongly than in amplitude modulated arrays. This is both an opportunity and a problem. It is disadvantageous that PM mirror arrays only work well in systems that are specifically designed to have the correct PSF, and it complicates the control of the array that the setting of one mirror will influence the setting of its neighbors. On the other hand, it is clearly a great advantage to be able to use multiple mirrors to achieve sub-mirror positioning accuracy as discussed in section 9.4.3.

When we ask the question of which is better, amplitude modulation or phase modulation, the obvious answer is that it depends on the application. For example, direct phase modulation as shown in Fig. 9.10 is of little use in micromirror projection displays. These systems have large magnification, and the mirror size is determined not by fundamental considerations of diffraction, but by technological constraints. It is therefore practical to design the magnified views of the rotating mirrors slightly larger than the PSF of the projection system. This makes the alignment of the optical system non-critical, and each mirror in the array can be controlled without regard to the state of its neighbors. The non-dispersive character of amplitude-modulating mirrors is also a strong practical advantage for white light imaging. These are the reasons that rotating mirrors are preferred for large-screen projections.

The fundamentally superior spatial resolution and smaller displacement requirements of phase-modulating mirrors can be used to advantage in a number of practical system, however. Systems that require the ultimate in spatial resolution, like active illumination microscopy and mask-less lithography, are obvious examples. Another opportunity presented by phase modulation is to use actuated arrays to create tunable diffractive optical elements. The dispersive nature of phase-modulating micromirrors can also be utilized to create optical filters and sensors. In the following chapters we will investigate the characteristics of optical microsystems that are based on phase modulation, diffraction, and/or interference, but before we go into detail, we will in the next section introduce some of the operational principles to clarify the opportunities offered by optical phase modulation.

9.7.1 Diffractive Optical MEMS

Traditional diffractive optics like gratings, zone plates, and holograms are static devices, i.e. they cannot be reconfigured to provide multiple degrees of freedom in optical manipulation or sensing applications. Using MEMS technology we can create large arrays of optical devices that can be precisely positioned using microactuators. Typically, the optical devices are simple mirrors that impart a phase

shift as shown in Fig. 9.10, but more complex optical modulators with amplitude, phase, and/or polarization control are also possible.

Taken to the extreme, such arrays can be thought of as a type of universal optics, i.e. an optical device that can be configured to act as a lens, an imager, a beam splitter, a wavelength filter or any other optical component that can be implemented by spatially modulating the properties of an incident optical field. Such universal optical synthesizers will be expensive, require complex control algorithms, and will not perform as well as devices that are optimized for a specific task, so specialized devices that have the minimum complexity required for a given function are better choices for most practical applications.

Phase-modulating or diffractive optical MEMS therefore come in many different variations, some of which are described in the following. The examples are chosen to illuminate operational principles. The goal of this section is not comprehensive coverage, but rather to provide a titillating glimpse of the opportunities.

The Spatial Light Modulators for Maskless lithography shown in Fig. 9.19a and b are designed to produce images with the smallest possible feature size and the best possible ability to position small features in off-grid positions. The contrast in the image, on the other hand, does not have to be very good because modern photoresists exhibit a threshold in their light sensitivity that allows sharp features to be defined by images that contain some background light.

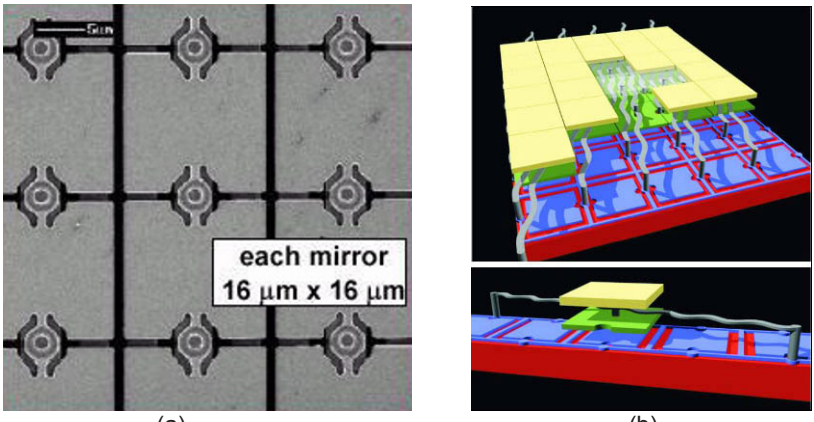
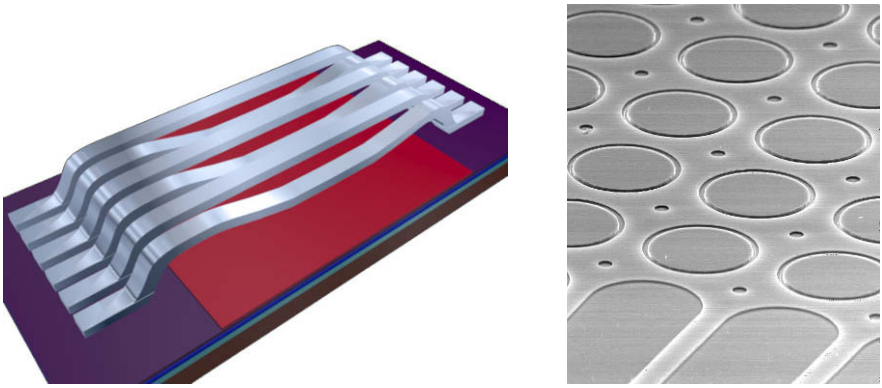


Figure 9.19 *Examples of micromirror arrays for mask-less lithography. (a) Micronic's Spatial Light Modulator [14]. (b) Conceptual drawing showing the design of Lucent's MEMS mirror array for maskless lithography [15]. Reprinted with permission.*

The grating light modulator [16] of Fig. 9.20 (a) is essentially an amplitude modulator, even though it locally affects only the phase of the incident light. The amount of phase modulation determines the diffraction from the grating modula-

tor, so that the phase modulation is converted into amplitude modulation by an aperture that selects either the diffractive or the reflective light from the grating.



(a) (b)

Figure 9.20 *Spatial Light Modulators (SLMs) based on phase modulation in micrograting. Each micrograting can be switched between a reflecting and a diffracting state. The drawing (a) shows details of a single pixel in a Grating Light Modulator for projection displays [17], and the SEM (b) shows Lightconnect's SLM used in fiber-optic Voltage Controlled Optical Attenuators [18]. Reprinted with permission.*

The simple mechanical structure of the grating modulator gives it several advantages. First and most importantly it is easy to manufacture and therefore mechanically precise which gives it high optical fidelity. Second, it can be made small and lightweight and therefore fast. Its small size together with its robustness also makes it relatively easy and inexpensive to package.

The Lightconnect SLM of Fig. 9.20 (b) operates on the same principle as the GLV; it converts phase modulation into amplitude modulation by controlling the diffraction efficiency of a micrograting. Compared to the GLV, the Lightconnect SLM has much reduced polarization sensitivity due to its fourfold rotational-symmetric design. This device is therefore well suited to fiber-optic applications that require polarization insensitivity.

The tunable blazed grating of Fig. 9.21 represents a class of diffractive optical MEMS that extends the GLV principle to optical filtering and spectroscopy. In these devices the MEMS SLM is designed to create a tunable diffraction filter. By correctly arranging the positions of the individual reflectors of the MEMS SLM, the spectral components of the incident light are preferentially diffracted into the output aperture of the filter. To enable a wide range of filter functions to be synthesized, the individual reflectors of the SLM must be individually controlled and have a range of motion that is substantially larger than that required for mono-

chromatic operation. The tunable blazed grating of Fig. 9.19 is designed to operate on a higher diffraction order that corresponds to the blaze angle of the grating. This gives the Tunable Blazed Grating high dispersion (wavelength dependence), and therefore superior spectral resolution.

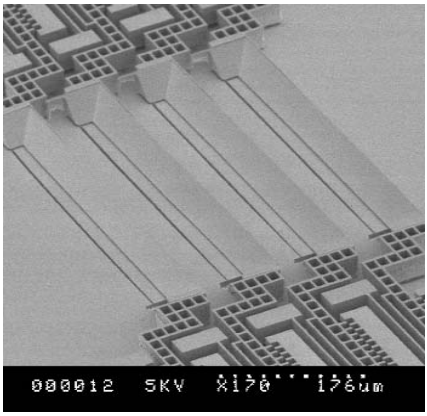


Figure 9.21 Tunable Blazed Grating for optical filtering and spectroscopy [19].

The examples shown in Figs. 9.19 through 9.21 are all actuators that influence or modulate the optical field. Their function is to create a light field with specific properties. The last example, shown in Figure 9.22, shows a sensor for which the optical output field is a function of some measurand.

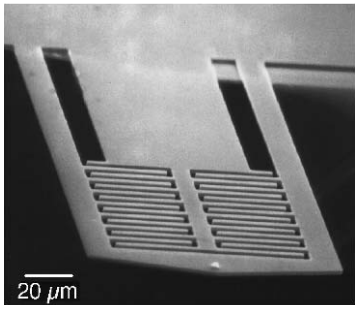


Figure 9.22 AFM array with interferometric displacement sensing. Reprinted from [20] with permission.

In the specific example shown in Fig. 9.22, the measurand is the force on the tip of an Atomic Force Microscope. This force changes the diffraction efficiency of the grating that is integrated into the AFM cantilever beam, and the change in diffraction efficiency is determined by measurement of the diffracted light. The grating creates an interferometer so that the relative position of the AFM tip can be measured with interferometric precision without requiring sub-wavelength alignment of

the light source and optical detector. The use of this type of displacement sensor is not restricted to AFMs. Accelerometers [21], microphones [22,23,24]], IR detectors [25,26], and sensors for bio-molecular associations [27] have been realized based on this principle.

The above examples demonstrate the versatility and wide field of use of diffractive optical MEMS. These diverse optical devices are all based on the same optical interactions principles, although different applications clearly require different MEMS designs and implementations. In the following chapters we will describe in detail several types of diffractive optical MEMS, starting with Grating Light Modulators for monochromatic light in Chapter 10. After describing the Grating Light Valve, which in many ways is the simplest example of diffractive optical MEMS, we extend the treatment to modulators optimized for fiber optics in Chapters 11. In Chapter 12 we study the properties of interferometers, including deformable gratings, as displacement sensors. We complete the description of diffractive MEMS in Chapter 13, where we turn our attention to the filtering properties of micromirror arrays.

9.8 Summary of Micromirror Arrays

This chapter starts with an in-depth discussion of the scaling of rotating, or tilting, micromirrors. In an analysis that parallels the approach we used to study optical scanners in Chapter 7, we show that diffraction impose fundamental limits on the scaling of rotating mirrors used in projection displays. The most important results of the analysis are the formulas that relate minimum mirror size (d) and minimum deflection (t) to specified contrast (C_{spec}), wavelength (λ), and rotation angle ($\Delta\theta$) of the micromirror

$$d = \frac{6\sqrt{0.5\ln\left(C_{spec}\right)}\cdot\lambda}{\pi\cdot\Delta\theta} \tag{9.60}$$

$$t \geq \frac{6\sqrt{0.5\ln\left(C_{spec}\right)}\cdot\lambda}{2\pi} \tag{9.61}$$

which evaluate to 2.5 um and 0.9 um respectively for C_{spec}=1000, λ=500nm, and $\Delta\theta$=0.7. These expressions show that micromirrors can be scaled to just a few wavelengths and still provide high contrast. In fact, most systems are constrained by practical consideration, rather than by these fundamental limits.

Section 9.3 looks at micromirror scaling from a different perspective, focusing on the characteristics of the system that is used to project the image created by the micromirror array. We find that the projected image of a micromirror is given by the scaled convolution of the magnified optical field and the Point Spread Func-

tion (PSF) of the projection system. That has several important consequences. We find that it is useless to make micromirrors so small that their magnified image is smaller than the PSF. In fact, maintaining good contrast requires that the magnified mirror size is about twice the PSF or larger. The conclusion is that the type of amplitude modulation provided by rotating micromirrors does not allow projected images with minimum features that are comparable to the PSF of the imaging system.

This shortcoming of amplitude-modulating mirrors leads us to consider phase modulation. In section 9.4 we show that phase-modulating mirrors give about twice the spatial resolution of rotating mirrors, and that this can be achieved with only a quarter wavelength displacement, compared to the one to two wavelengths that are necessary for rotating mirrors. We also find that fact that the images of neighboring phase-modulating mirrors interact. This gives us the opportunity to position dark features with an accuracy that is better than the size of the individual micromirrors, but it also present a complicating, and sometimes cost driving, control problem.

In the final section of the chapter, we compare and contrast the characteristics of phase and amplitude modulating micromirro arrays, and conclude that the optimum modulation format depends on the application. As a general rule, however, phase modulation provides better functionality at the cost of increased complexity. To illustrate some of the opportunities of phase modulating microsystems, we finish the chapter with a set of examples of phase-modulating, or diffractive, optical MEMS.

Exercises

Problem 9.1 - Gaussian Point Spread Function

In the text we use a Gaussian Point Spread Function mostly because it simplifies analytical modeling, but a Gaussian PSF also has other advantages.

a. How can you create a Gaussian PSF?
b. What are the practical advantages and disadvantages of a Gaussian PSF?

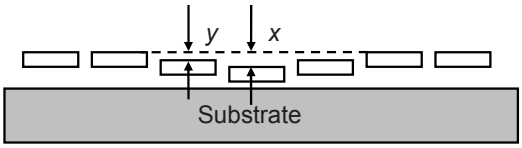
Problem 9.2 - Curved Micromirrors

The micromirrors in the arrays we have considered are all flat and they are created on flat substrates.

a. What would be the advantages of arrays of curved micromirrors?
b. Ditto for curved substrates?

Problem 9.3 - Non-local Response

The non-local response of phase-modulating mirrors creates both opportunities and challenges. One potential advantage is that the effect of motion of one micromirror can be maximized by adjusting the position of its neighbors. We will restrict our considerations to the situation where the nearest neighbors are moved symmetrically as shown in the figure below.



The response to motion (x) of the central mirror is affected by the positions of the surrounding mirrors.

a. Write an expression for the projection of the three lines as a function of x and y, the wave length of the incident light, the PSF, and the magnification of the projection system.
b. Can you find a situation, in which a non-zero y improves the response of the central mirror? Use numerics and/or plotting if necessary.
c. How can this effect be utilized for practical purposes?

Problem 9.4 - Vortices

The 2-D equivalent to a phase step is a vortex, i.e. a phase singularity around which the phase increases from zero to 2p .

a. How can we create an approximation to a vortex at the intersection of four micromirrors in a square lattice?
b. Write an expression for the projection of a vortex.
c. What are the advantages and difficulties in creating small features using vortices?

Problem 9.5 - Tilting Micromirrors

In the section on tilting micromirrors we made the assumption that these devices are used as amplitude modulators, i.e. the function of the mirrors are to direct some part (or none) of the incident light outside the aperture of the projection optics. It is, however, also possible to use them as phase modulators. In that case we set up the projection systems to capture all the light from each micromirror irrespective of its rotation angle. The only things that varies with rotation is then the linearly varying phase across the mirrors.

a. Use the techniques developed in this chapter to find the response of a tilting, phase modulating micromirror. Make simplifying assumptions to arrive at an analytical solution.
b. Under what conditions does tilting, phase-modulating micromirrors perform better than piston-motion, phase-modulating micromirrors, and under what conditions is the opposite true?
c. Can you use phasors as a simple means to illustrate the point made in b)?

Problem 9.6 - Phase vs. Amplitude

Compare amplitude and phase modulation by listing their relative strengths and weaknesses.

Problem 9.7 - Wild Duck Chase

Invent something to do with photography based on micromirrors. (Don't expect help from Gregers.)

References

1 R.N. Thomas, J. Guldberg, H.C. Nathanson, P.R. Malmberg, "The mirror matrix tube, A novel light valve matrix for projection displays", IEEE transactions on Electron Devices, vol. ED-22, p.765, 1975.
2 L.J. Hornbeck, "Deformable-mirror spatial-light modulators", Proceedings of the SPIE, vol. 1150, pp. 86-102, August 1989.
3 L.J. Hornbeck, "The DMD™ projection display chip: a MEMS-based technology", MRS Bulletin; April 2001; vol.26, no.4, pp.325-7.
4 http://www.dlp.com/
5 K.F. Chan, Z. Feng, R. Yang, A. Ishikawa, W. Mei, "High-resolution maskless lithography", Journal of Microlithography, Microfabrication, and Microsystems; Oct. 2003; vol.2, no.4, pp.331-9.
6 http://www.newport.com/lambdacommander/
7 V. Bansal, S. Patel, P. Saggau, "High-speed addressable confocal microscopy for functional imaging of cellular activity", Journal of Biomedical Optics; May 2006; vol.11, no.3, pp.34003-1-9.
8 E. Hecht, Optics, 2nd ed., Addison Wesley, 2001, ISBN 0-201-11609-X. See also http://en.wikipedia.org/wiki/Airy_disc
9 J.H. Debes, J. Ge, A. Chakraborty, "First High-Contrast Imaging Using a Gaussian Aperture Pupil Mask", The Astrophysical Journal, 572:L165–L168, 2002 June 20.

10 T. Sandstrom, H. Martinson , “RET for optical maskless lithography”, Proceedings of the SPIE - The International Society for Optical Engineering; 2004; vol.5377, no.1, pp.1750-63.
11 J.-S. Wang, S. Hafeman, A. R. Neureuther, and O. Solgaard, "Effects of Through-Focus Symmetry in Maskless Lithography Using Micromirror Arrays", Journal of Vacuum Science and Technology B, Vol. 23, No.6, pp.2738-2742 (November/December 2005).
12 I.W. Jung, U. Krishnamoorthy, O. Solgaard, “High Fill-Factor Two-Axis Gimbaled Tip-Tilt-Piston Micromirror Array Actuated by Self-Aligned Vertical Combdrives“, Journal of Microelectromechanical Systems, vol. 15, no. 3, June 2006, pp. 563-571.
13 H. Lee, M.H. Miller, T.G. Bifano, “CMOS chip planarization by chemical mechanical polishing for a vertically stacked metal MEMS integration”, Journal of Micromechanics and Microengineering, vol. 14, 2004, pp. 108–115.
14 E. Croffie, N. Eib, N. Callan, N. Baba Ali, A. Latypov, J. Hintersteiner, T. Sandstrom, A. Bleeker, K. Cummings, “Application of Rigorous Electromagnetic Simulation to SLM-based Maskless Lithography for 65nm Node”, Proceedings of SPIE Vol. 5256, 23rd Annual BACUS Symposium on Photomask Technology, edited by Kurt R. Kimmel, Wolfgang Staud.
15 V.A. Aksyuk, D. López, G.P. Watson, M.E. Simon, R.A. Cirelli, F. Pardo, F. Klemens, A.R. Papazian, C. Bolle, J.E. Bower, E. Ferry, W.M. Mansfield, J. Miner, T.W. Sorsch, D. Tennant, “Mems Spatial Light Modulator for Optical Maskless Lithography”, Proceedings of the Solid-State Sensor and Actuator Workshop, pp. 352-355, Hilton Head, South Carolina, June 6-10, 2006.
16 O. Solgaard, F. S. A. Sandejas, D. M. Bloom, "A deformable grating optical modulator", Optics Letters, vol. 17, no. 9, pp. 688-690, 1 May 1992.
17 http://www.siliconlight.com
18 A. Godil, “Diffractive MEMS technology Offers a New Platform for Optical Networks”, Laser Focus World, May 2002.
19 X. Li, C. Antoine, D. Lee, J.-S. Wang, O. Solgaard, “Tunable Blazed Gratings“, Journal of Microelectromechanical Systems, vol. 15, no. 3, June 2006, pp. 597-604
20 G.G. Yaralioglu, A. Atalar, S.R. Manalis, C.F. Quate, “Analysis and Design of an interdigital cantilever as a displacement sensor”, Journal of Applied Physics, vol. 83, no. 12, 15 June 1998, pp. 7405-7415.
21 N.C. Loh, M.A. Schmidt, S.R. Manalis, “Sub-10 cm^3 Interferometric Accelerometer with Nano-g Resolution”, Journal of Microelectromechanical Systems, vol. 11, no. 3, June 2002, pp. 182-187.
22 H. Sagberg, A. Sudbo, O. Solgaard, K.A. Hestnes Bakke, I.-R. Johansen, “Optical Microphone Based on a Modulated Diffractive Lens”, IEEE Photonics Technology Letters, vol.15, no.10, October 2003, pp. 1431-1433.

23 N.A. Hall, F.L. Degertekin, “Self-calibrating micromachined microphones with integrated optical displacement detection”, 11th International Conference on Solid State Sensors and Actuators Transducers '01/Eurosensors XV, Munich, Germany, 10-14 June 2001, pp.118-121, vol.1.

24 L. Wook, N.A. Hall, Z. Zhiping, F.L. Degertekin, “Fabrication and characterization of a micromachined acoustic sensor with integrated optical readout”, IEEE Journal of Selected Topics in Quantum Electronics, vol.10, no.3, May-June 2004, pp.643-651.

25 S.R. Manalis, S.C. Minne, C.F. Quate, G.G. Yaralioglu, A. Atalar, “Two-dimensional micromechanical bimorph arrays for detection of thermal radiation”, Applied Physics Letters, vol. 70, no. 24, 16 June 1997, pp. 3311-3313.

26 Y. Zhao, M. Mao, R. Horowitz, A. Majumdar, J. Varesi, P. Norton, J. Kitching, “Optomechanical Uncooled Infrared Imaging System: Design, Microfabrication, and Performance”, Journal of Microelectromechanical Systems, vol. 11, no. 2, April 2002, pp. 136-146.

27 C.A. Savran, T.P. Burg, J. Fritz, S.R. Manalis, “Microfabricated mechanical biosensor with inherently differential readout”, Applied Physics Letters, vol. 83, no. 8, 25 August 2003, pp. 1659-1661.

10: Grating Light Modulators

10.1 Introduction to Grating Light Modulators

In Chapter 9 we described the optical properties of mirror arrays and demonstrated that phase modulation is preferable to amplitude modulation for many applications. The advantages of phase modulation are explored further in this chapter on Grating Light Modulators (GLMs) [1]. The GLM is a simple, yet versatile, optical modulator that use diffraction to convert phase modulation to amplitude modulation. In other words, it performs the function of an amplitude modulator, but is based fundamentally on phase modulation and retains many of the advantageous characteristics of phase modulators.

The Chapter starts with a phenomenological description of the operational principles, the mechanics, and the optics of GLMs. We then use phasor notation to explore some of the first-order design issues. This treatment leads us to the high-contrast GLM. The mathematical modeling is refined in Section 10.5 and used to determine the scaling properties of GLMs.

The last part of the chapter is devoted to a detailed description of projection displays based on grating modulators. In this section we look closer at the implementation of the grating modulator itself, concentrating on actuation and mechanics. The most important and unique elements of the optical system design is also described.

10.2 Phenomenological Description of MEMS Grating Modulators

10.2.1 Mechanical Design and Actuation of Grating Light Modulators

MEMS actuators enable positioning of micromirrors with high spatial precision and high speed. That allows us to turn diffraction gratings into high quality optical switches or amplitude modulators. A simple MEMS implementation of such a

deformable grating is shown in Fig. 10.1. The fabrication starts by deposition of a sacrificial film and a structural film on a silicon substrate. A grating is defined in the structural layer by photolithography. Following this step, the sacrificial layer is etched away under the grating so that the grating elements become free-standing ribbons as shown.

The amount of reflected light from the grating modulator is controlled by positioning the movable ribbons. There is a multitude of actuation principles that can be applied to the task of moving and positioning the ribbons, including thermal, magnetic, piezoelectric, and electrostatic actuation. However, the particular device structure, with ribbons that are suspended relatively close to a flat substrate, makes electrostatic actuation particularly simple to implement. All that is needed is that the ribbons and the substrate are conducting and that they can be held at different electrostatic potentials. Electrostatic actuation is therefore by far the most common actuation mechanism used in practical grating light modulators.

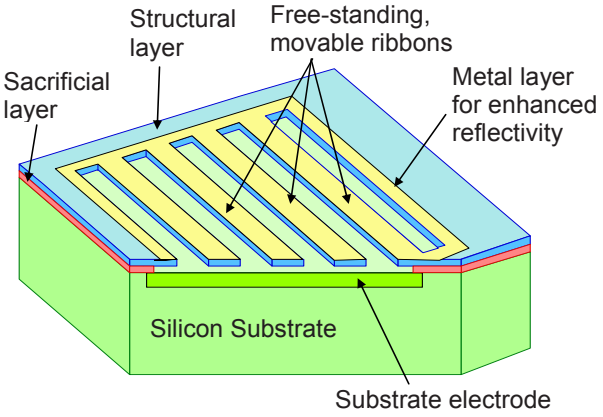


Figure 10.1 Schematic of a single deformable rating modulator with a cut-out to show the ribbon structure.

The grating modulator of Fig. 10.1 has several important advantages. First, the fabrication is very simple with only a single mask needed to define the grating itself. In practice the fabrication is more complicated, requiring several masks to define wiring, reflectors, and sacrificial layer to be removed, but the number of photolithography steps that have to be carried out is still quite small. This simplicity is important in and of itself, but here it achieves extra significance because it simplifies the integration of grating modulators into arrays and with other semiconductor devices, e.g. transistors, so that integrated, multifunctional microsystems can be built.

The second important advantage of the grating modulator is that the required film thicknesses are on the order of an optical wavelength or less. This is fortuitous

because the wavelengths of visible (400 nm to 650 nm) and of near infra-red light used in optical fibers (1,300 to 1,600 nm) are comparable to the film thicknesses used in modern Integrated Circuit technology. Optical components are more sensitive to surface roughness and curvature than electronic devices, but with careful attention, most thin-film technologies used for electronics can be adapted to diffractive optical MEMS.

The lateral dimensions of grating modulators range anywhere from less than 10 microns to several hundreds of microns. This is orders of magnitude larger than state-of-the-art transistors, but we nevertheless are able to place large arrays with thousands or even millions of modulators on chips of standard size. All together, this means that the advanced methods and tools, as well as the enormous capacity, of the IC industry can be readily applied to the fabrication of deformable grating modulators.

Being able to build grating modulators using IC technology is of little value if the end product is not also robust and reliable. It is clear from Fig. 10.1 that the grating light modulator, although small, is a machine with moving parts. The difference from the typical machinery we use every day is that the grating modulator cannot practically be maintained or repaired. It has to function unassisted for its entire lifetime, which for some applications might reach 10^{14} cycles! It has been shown that grating modulators can indeed function without measurable wear and tear over such large umber of operations, provided that they are made from high quality ceramics or single-crystalline materials, and that they are packaged in an inert atmosphere. The reason for this is that the mechanical motion of a grating modulator required to produce nearly complete off-on switching (or 100% contrast) is one quarter of the wavelength of the light, i.e. only about 150 nm at visible wavelengths. This small required motion leads to stresses that are much lower than the yield stress of the ribbon material.

Another advantage of the small mass of the ribbons and the small distance that they must be moved is that the switching speed can be very high. Rise and fall times as short as 20 ns have been demonstrated [2,3]. Clearly switching speeds on this order are low compared to optoelectronic devices like semiconductor lasers and modulators, but the fact that the grating modulator can be readily integrated into large arrays means that the combined information throughput of a grating modulator array can be very high.

The demonstrably high speed of the basic grating light modulator of Fig. 10.1 is partially a function of the small gap between the ribbons and the underlying substrate. As we will see, the gap is of the same size as the ribbon thickness, and equal to one quarter of the wavelength of the light that is to be modulated. This small distance enables relatively large electrostatic forces to be applied to the ribbons, but is also makes accurate position control of the ribbons difficult because of the electrostatic instability. This makes the basic grating modulator as shown in

Fig. 10.1 unsuitable for analog operation. It is effectively a binary device with the ribbons either relaxed or pulled all the way to the substrate.

It is, however, straight forward to modify the basic grating modulator for analog operation. All that is needed is that the ribbons are moved further from the substrate such that it is possible with electrostatic actuation to continuously control the position of the movable ribbons over a range corresponding to a quarter of the wavelength of the light. Typically this is done in conjunction with another structural change, which is to fill the gaps between the movable ribbons with fixed ribbons at the level of the movable ribbons in their relaxed state. This reduces the dispersion of the device, as we will see in the treatment of optical functionality in the following.

From this discussion we can conclude that the grating modulator is fast and reliable, that it can be integrated into multifunctional microsystems, and that it can be produced at low cost in large numbers due to its compatibility with IC fabrication technology. Unfortunately, this IC compatibility does not extend to packaging. The relatively small chip size will in principle make grating arrays simple and inexpensive to package, but, unfortunately, optical modulators require an optical interface and must be packaged in an inert atmosphere, so the packaging techniques of the IC industry cannot be applied without significant modifications. The wide range of different shapes and sizes of the systems that are based on grating modulators has also made it difficult to standardize packaging. The result is that custom solutions have to be developed separately for each grating modulator application. Much of the cost advantage of the grating-modulator chips therefore disappears by the time the chips are packaged. At the present time this creates a significant hurdle for commercialization of new products, but presumably this hurdle will be lowered as standards for optical MEMS packaging are developed.

10.2.2 Optical Design and Operation of Grating Light Modulators

Now we will turn our attention from the mechanical design and properties of the grating light modulator to its optical functionality. As with any grating, the deformable grating modulator diffracts light. The unique feature of the GLV is that the amount of diffraction, the diffraction efficiency, can be controlled through the application of an electrostatic voltage.

The principle of operation is explained in Fig. 10.2. Figure 10.2a shows what happens to a monochromatic plane wave at normal incidence on the grating modulator when the grating ribbons are in the upper position, i.e. the relaxed state without an applied electrostatic force. In this case, the height difference between the top of the ribbons and the substrate corresponds to exactly one half of the wavelength of the monochromatic plane wave. The path length difference for the light that is reflected from the ribbons and the light that is reflected from the substrate is

therefore exactly one wavelength, which means that the two reflected parts are in phase and interfere constructively in reflection.

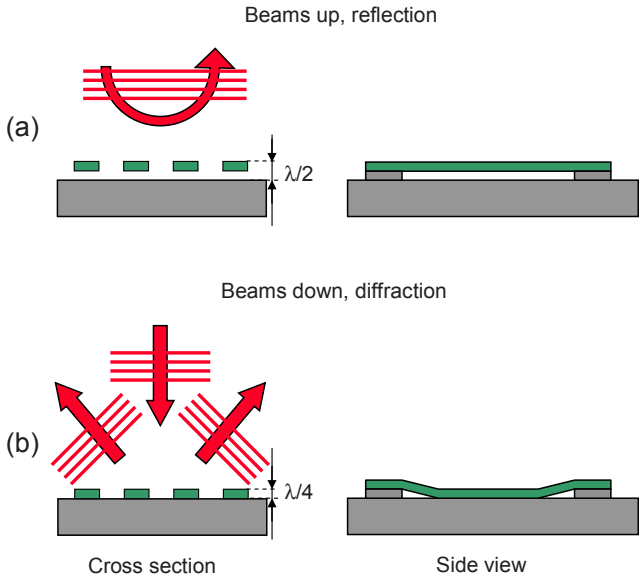


Figure 10.2. The grating modulator switches light by controlling diffraction. When the ribbons of the grating are in the relaxed state, as in a, the parts of the incident light that are reflected from the ribbons and from the substrate are in phase, and all the incident optical power is reflected. When the ribbons are pulled to the substrate, as in b, the reflections from the ribbons and the substrate are in opposite phase, and all the incident optical power is diffracted.

If, on the other hand, the ribbons are pulled down to the substrate by an applied electrostatic force, as shown in Fig. 10.2b, then the height difference between the top of the ribbons and the substrate is exactly one quarter of the wavelength of the light. The total path length difference is now exactly one half wavelength, and the two reflected parts are out of phase and interfere destructively in reflection. The plane wave is therefore not reflected from the grating, but instead its optical power is diffracted into a set of diffraction modes at discreet angles.

In the implementation of a deformable grating shown in Fig. 10.2, optical path length differences are created by moving a reflector so that the reflected light must propagate over a longer distance. This variation of path length difference allows us to control the optical phase. There are many other physical phenomena that can be used to control optical phase, but displacement of mirrors are among the most robust and conceptually most straight forward. It is also particularly simple to implement in MEMS technology.

The relationship between path length and phase delay is, however, only valid at one single wavelength. If we use grating light modulators to manipulate broad band light (i.e. light that is not to a good approximation monochromatic), we must consider the variation of phase delay across the optical input spectrum. We will see later that that might lead to slightly different implementations, depending on what type of response we want across the input spectrum.

The two states of the grating shown in Fig. 10.2a and b are extremes in the sense that in (a) the path length difference is one wavelength and ALL the incident light is reflected, while in (b) the path length difference is half a wavelength and ALL the incident light is diffracted. If the path length difference is somewhere between half and one wavelength, the returning light from the grating will be a mix of reflected and diffracted light. By controlling the exact position of the ribbons and thereby the path length difference, we have analog control over how much light is reflected and how much is diffracted. The net effect is that we can control the state of the light coming off the grating modulator by controlling the height difference, or grating amplitude, of the deformable grating. Of course this requires that we have a structure and an actuation mechanism that allows continuous positioning of the ribbons.

The principle of operation of the grating modulator as described in Fig. 10.2 demonstrates that it is an extension of the Eidophore [4,5,6], a display technology that was first demonstrated in 1943. The display engine, or light modulator, of the Eidophore is a thin reflective and conductive diaphragm that is suspended over an oil film. The reflective layer can be deformed electrostatically to form a diffraction grating, so just like the grating modulator, the Eidophore display can be switched between a reflective and a diffractive state.

The MEMS implementation gives the grating modulator several advantages over the Eidophore. In particular, MEMS technology allows better dimensional control, more flexible grating design, better compatibility with electronics, and better reliability. All together this makes grating modulators smaller, faster, and easier and less expensive to fabricate, integrate, package, and operate.

10.2.3 Schlieren Projection System

The Schlieren system shown in Fig. 10.3 is an example of how arrays of grating light modulator can be used in a projection display. Here the GLV is uniformly illuminated at near-normal incidence. Any pixel that is in the reflective state will send the light incident on it back towards the light source via the turning mirror on the optical axis. Light from the reflecting pixels will therefore not reach the screen, and they will appear dark. Pixels that are in the diffractive state will send

the incident light around the turning mirror so that it is imaged on the screen by the imaging lens. The diffracting pixels will therefore appear bright on the screen.

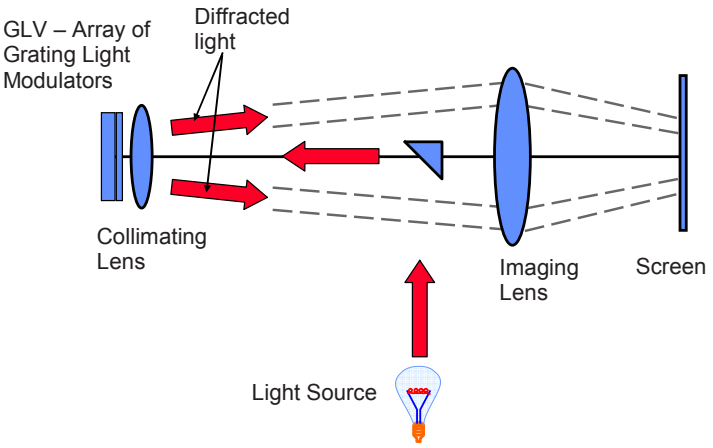


Figure 10.3. Basic Schliern projection system using an array of grating light modulators as the image forming element. Each pixel on the screen corresponds to a modulator in the array. If the modulator is in the reflecting state, the light is reflected back out of the system, and the pixel on the screen is dark. If, on the other hand, the modulator is in the diffracting state, the light is imaged onto the screen, and the pixel appears bright.

10.3 Phasor Representation of Grating Modulator Operation

To get a quantitative understanding of the operation of the grating light modulator, particularly its analog operation and its response to broadband light, it is instructive to use the phasor representation introduced in Chapter 2. The phasor diagrams of Fig. 10.4. represent three different states of a Grating Light Modulator. Each diagram shows three phasors representing optical fields. The phasors representing the fields reflected from the ribbons and from the substrate are shown as solid lines, while their sum, representing the total reflected field, is shown as a dashed line.

What is not shown in Fig. 10.4 is the diffracted field. The gratings we are considering in this chapter are metal coated (i.e. the reflectivity is high and uniform) and have periods that are larger than the wavelength, so we can assume that the sum of the reflected and diffracted powers is constant and equal to the incident power multiplied by the reflection coefficient of the metal. In other words, whatever power is not in the reflected field, will be in the diffracted field.

Figure 10.4a) is the phasor representation of the physical situation shown in Fig. 10.2a). The two phasors are in phase so their sum attains its maximum value, which is the sum of the absolute values of the two reflected parts. In this state, there is no particular significance associated with the relative size of the two phasors.

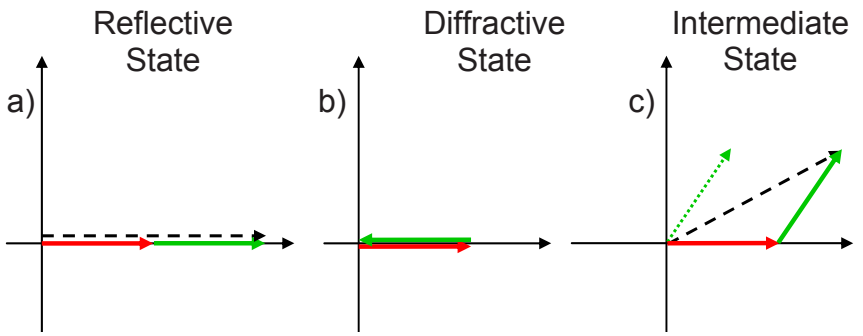


Figure 10.4. Phasor representation of the optical reflected fields from a Grating Light Modulator. The two phasors representing the light reflected from the ribbons and from the substrate are drawn as solid lines, while total reflected field is drawn as a dashed line. In (a) the two reflected parts are in phase, resulting in a maximum value for the total reflected field (shown offset for clarity). In (b) the two reflected parts are exactly out of phase, so the total reflected field is zero. In (c) the phase difference between the two reflected parts is between zero and π radians, so the resulting total reflected field is between zero and its maximum value.

In Fig. 10.4b) the path-length difference for the light reflected from the two parts of the modulator is π radians, i.e. the two parts of the reflected light are in exactly opposite phase. The result is that there is no reflected light from the modulator in this state. The incident light is therefore completely diffracted. It is clear from the figure that we only get complete suppression of reflection when the two phasors, representing the two parts of the reflected light, are equal so that they exactly cancel when they are in opposite phase. For the physical implementation shown in Figs. 10.1 and 10.2, this means that the areas of the ribbons must match the areas between the ribbons to achieve complete diffraction.

The usefulness of the phasor representation becomes clear when we consider Fig. 10.4c) that shows the reflected light when the two parts of the reflections have a relative phase between zero and π radians. The resulting reflected field now has a value that is somewhere in between zero for the out-of-phase configuration and the maximum value for the in-phase configuration. We can find the resulting reflected field for an arbitrary relative phase, θ, by vector summation. Here we are not interested in the phase of the reflected light, so we write:

$$\begin{aligned} &\vec{E}_{tot} = \vec{E}_{ribbon} + \vec{E}_{substrate} \Rightarrow \\ &\left|E_{tot}\right| = \sqrt{\left(\left|E_{ribbon}\right| + \left|E_{substrate}\right| \cdot \cos\theta\right)^2 + \left(\left|E_{substrate}\right| \cdot \sin\theta\right)^2} \end{aligned} \tag{10.1}$$

where $\vec{E}_{tot}$ is the total reflected field, $\vec{E}_{ribbon}$ is the reflected field from the ribbons, and $\vec{E}_{substrate}$ is the reflected field from the substrate. To simplify the calculations, we assume that the reflected fields from the ribbons and from the substrate are of equal magnitude, i.e. $\left|E_{ribbon}\right| = \left|E_{substrate}\right|$.

$$\begin{aligned} \left|E_{tot}\right| &= \left|E_{ribbon}\right| \cdot \sqrt{1 + 2 \cdot \cos\theta + \cos^2\theta + \sin^2\theta} \\ \left|E_{tot}\right| &= \left|E_{ribbon}\right| \cdot \sqrt{2(1 + \cos\theta)} \end{aligned} \tag{10.2}$$

$$\left|E_{tot}\right| = 2 \cdot \left|E_{ribbon}\right| \cdot \cos^2\frac{\theta}{2} \tag{10.3}$$

The reflected optical power then becomes:

$$P_{reflected} = R \cdot P_{incident} \cdot \cos^2\frac{\theta}{2} \tag{10.4}$$

where $P_{reflected}$ and $P_{incident}$ are the reflected and incident optical powers, respectively. As pointed out above, the light that is not reflected must be diffracted, so the diffracted power, $P_{diffracted}$, can be expressed as:

$$P_{diffracted} = R \cdot P_{incident} \cdot \left(1 - \cos^2\frac{\theta}{2}\right) = R \cdot P_{incident} \cdot \sin^2\frac{\theta}{2} \tag{10.5}$$

These two simple harmonic expressions for the reflected and diffracted optical powers are shown graphically in Fig. 10.5. Note that the diffracted light here means the total diffracted light summed over all diffraction orders. Later we will develop a more complete model that allows us to distinguish between the powers in different diffraction modes.

The curves confirm our earlier assertion that to switch light we must change the relative phase by an amount of π radians. To get complete switching we can for example go from a state with 2π radians to one with π radians relative phase difference. This is the switching strategy employed in the implementation shown in Figs. 10.1 and 10.2. The periodic curves of Fig. 10.5 demonstrate, however, that that is not the only switching strategy. In fact we will see that we can achieve less dispersive (less wavelength-dependent) switching by changing the relative phase from zero relative phase to π radians.

One of the important figures of merit for optical modulators is their contrast, which we define simply as the ratio of maximum optical power output to minimum optical power output. The required contrast is application dependent, but most systems require well in excess of 20dB contrast, and many high-quality display systems require contrast as high as 30 dB.

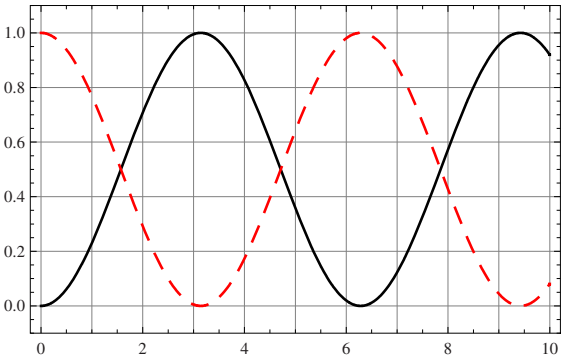


Figure 10.5. Total reflected (green line) and total diffracted (blue line) optical powers normalized to the power reflectivity of the grating modulator as a function of relative phase difference of the two parts of the reflected light. Both the reflected and diffracted optical powers are harmonic functions of the phase difference.

It is clear from its definition that contrast is more strongly dependent on errors in the dark state than error in the bright state, so in any system that requires high contrast, we must pay special attention to the dark state. Figure 10.5 shows that in principle a grating modulator can achieve zero output power, so it is theoretically possible to get infinite contrast. In practice, it is of course impossible to verify that the output is exactly zero, so even in the best case we have a finite contrast that is given by our measurement resolution. More typically, the dark state is determined by scattered light or by dispersion (wavelength dependence) of the modulator as we will discuss in detail later.

As mentioned above, in optical MEMS technology we control optical phase indirectly by creating variation in optical propagation length or path length. That means that we only control the phase differences at one specific wavelength, and light at all other wavelengths will experience a different relative phase. This has implications for how we should design and operate many different kinds of diffractive optical MEMS, including grating modulators.

Figures 10.4 and 10.5 illustrate that for monochromatic light, there is no difference between having a zero path length difference, a path length difference of one wavelength, corresponding to a relative phase of 2π, or even path length difference of several wavelengths, corresponding to relative phases of $n \cdot 2\pi$ (here n is an integer). For broad band light this is no longer the case, because path length dif-

ferences no longer uniquely determines phase differences. This is explained in Fig. 10.6 that show the phasor representation of the reflections from a grating modulator of three slightly different wavelengths.

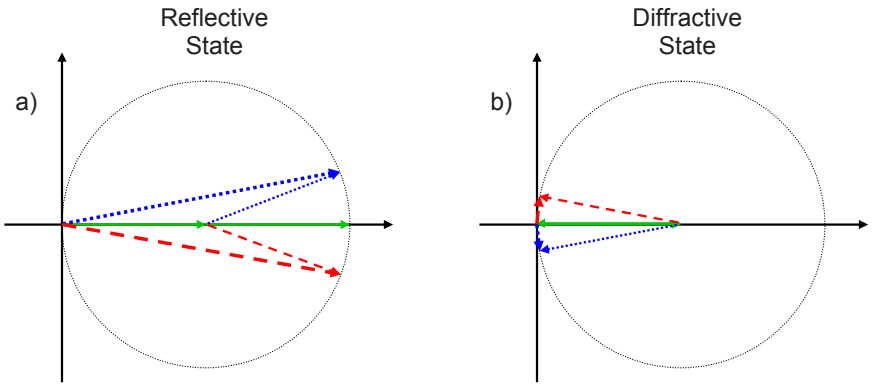


Figure 10.6. Phasor representation of reflected light from a grating light modulator in the reflective (a) and diffractive (b) states at three different wavelengths. At the center wavelength (solid), the phasors add in the reflective state and subtract to zero in the diffractive state. At the wavelengths slightly longer than the center wavelength (dashed), the phasors add slightly out of phase in the reflective state, and do not completely annihilate each other in the diffractive state. The same is true for wavelengths that are slightly shorter than the center wavelength (dotted).

Figure 10.6a represents the grating modulator in the state shown in Fig. 10.2a where the ribbons are suspended over the substrate at a distance equal to one half of the center wavelength, which is the wavelength that the grating modulator is designed for. It can be in any part of the spectrum. The path length difference is then exactly one center wavelength, so the two parts of the reflected light (shown as solid phasors) are exactly in phase and their vector sum achieves its maximum value.

Now let us consider a wavelength that is slightly longer than the center wavelength. At this wavelength the path length difference is slightly less than one wavelength, so the reflection from the substrate (dashed) is not quite in phase with the reflection from the ribbons. The total reflected field (dashed) therefore has a slightly lower amplitude and a slightly different phase than the reflection at the center wavelength.

The situation is very similar for a wavelength that is slightly shorter than the center wavelength. Now the path length difference is slightly more than a wavelength, so the light from the substrate (dotted) has advanced more than 2π radians in phase compared to the light from the ribbons. The total reflected field (dashed)

at this shorter wavelength is therefore advanced in phase and has a slightly lower amplitude than the reflected field at the center wavelength. The net effect of having broad band light incident on a grating modulator that is in the reflective state is therefore to slightly attenuate wavelengths other than the center wavelength and also to give slightly different phase response across the broad-band spectrum.

When the grating modulator is in the diffractive state, the effect on off-center wave lengths is more dramatic. At the center wavelength the path length difference is exactly one half wavelength, so the reflections from the ribbons and from the substrate cancel each other exactly, and all the light is diffracted. At a wavelength slightly longer than the center wavelength, the reflections from the substrate (dashed) have advanced slightly less than π radians, which means that the total reflected field (dashed) is non-zero. Similarly, reflections from the substrate of wavelengths slightly shorter than the center wavelength (dotted) have advanced slightly more than π radians, and again the total reflected field (dotted) is finite.

The net effect is that we get the same type of phase variation across the spectrum as we saw in the reflective state, but the relative amplitude change is much larger in the diffractive state. The absolute variations as a function of wavelength of the fields are also larger in the diffractive state. In most applications, however, we care about variations in optical power, which goes as the square of the optical field, and the absolute power variations are actually larger in the reflective state because of the longer path length difference and the correspondingly stronger phase dependence on wavelength.

Figures 10.4 and 10.6 show the phasor representations of the reflected fields. In a completely analogous fashion, we can depict the total diffracted fields. The structure of the diffracted fields are more complex as we will see in the more detailed treatment later, but for now we will use the fact that the light that is not reflected is diffracted. The total diffracted field is simply the vector difference between the reflections from the ribbons and from the substrate as shown in Fig. 10.7. We see that the phasor representation of the total diffractive field in the reflective state is similar to the reflected fields of the diffractive state and vice versa.

There are, however, important differences. In particular we see from the figure that dispersion is quite different for the reflected and diffracted fields. While the diffractive state has a larger relative amplitude variation for the reflected light, the opposite is true for the diffracted light, i.e. the reflective state is more dispersive for diffracted light. In the specific implementations illustrated in Figs. 10.2, 10.4, 10.6, and 10.7, diffracted light from the reflective state experience the most dispersion, followed by reflected light from the diffractive state, reflected light from the reflective state, and finally diffracted light from the diffractive state.

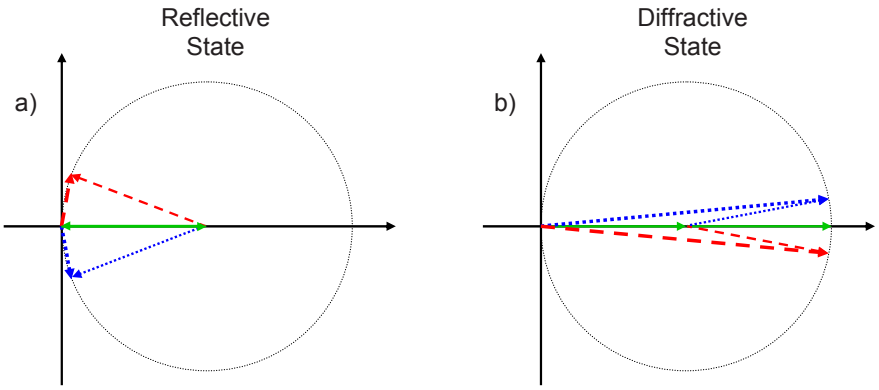


Figure 10.7. Phasor representation of diffracted light from a grating light modulator in the reflective (a) and diffractive (b) states at three different wavelengths. At the center wavelength (solid) the phasors subtract in the reflective state and add in the diffractive state. At longer (dashed) and shorter (dotted) wavelengths the phasors are not perfectly in phase in the diffracted state, and they do not perfectly cancel in the reflective state.

10.4 High Contrast Grating Light Modulator

Based on the observations made in section 10.3, it is clear that the basic projection system of Fig. 10.3 is far from optimal when used with a broad band light source combined with a grating light valve of the construction shown in Fig. 10.1 and 10.2. In any display it is very important to have high contrast, which is defined as the ratio of maximum to minimum optical power per pixel. Even low-quality displays need contrast on the order of 100, while really superior image sharpness requires a contrast of closer to 1,000. When a pixel is in its dark state, it must therefore block light efficiently across the whole spectrum of the illuminating source. In other word, the dark state must have the lowest possible dispersion. The dispersion of the bright state is not nearly as critical, because small variations in the total optical power in the bright state do not lower the contrast significantly.

The problem with the projection system of Fig. 10.3 is that is uses precisely the most dispersive configuration to create dark pixels. As explained above, a dark pixel is formed when the corresponding grating modulator is in it reflective state. Light at the center wavelength will then be reflected back towards the light source by the turning mirror. The relatively strong dispersion of this state will, however, lead to significant amounts of diffracted light at wavelengths other than the center wavelength. With broadband light sources, it is therefore difficult to obtain good contrast in this configuration.

Fortuitously, there is a straight forward solution to this problem. Consider for a moment the modulation characteristics of a grating light modulator as illustrated in Fig. 10.5. So far we have described the operation of grating light modulators as switching from a reflective state with a 2π radians path length difference to a diffractive state with a π radians path length difference. This is of course not the only possibility for switching from reflection to diffraction. We can chose to go from any reflectance maximum to any reflectance minimum. In particular, it is advantageous to switch from a zero path length difference to one that equals π radians. This can be done by modifying the grating light modulator so that the static reflection plane (the substrate in Fig. 10.1) and the movable reflection plane (the ribbons in Fig. 10.1) coincide in the reflective state.

There are many ways that can be implemented, but in MEMS technology it is most straight-forward to create a modulator that consists of uniform ribbons that are separated by the minimum amount that the technology will allow. The operation of such a grating light modulator is shown schematically in Fig. 10.8. When the individual ribbons of this grating are all in the relaxed state, the reflections from the ribbons are all in phase, so they interfere constructively and the reflected field is maximized. To switch to the diffractive state, every other ribbon of the grating are actuated so that the path length difference fro the light that is reflected from the actuated ribbons and from the unactuated ribbons is exactly π radians.

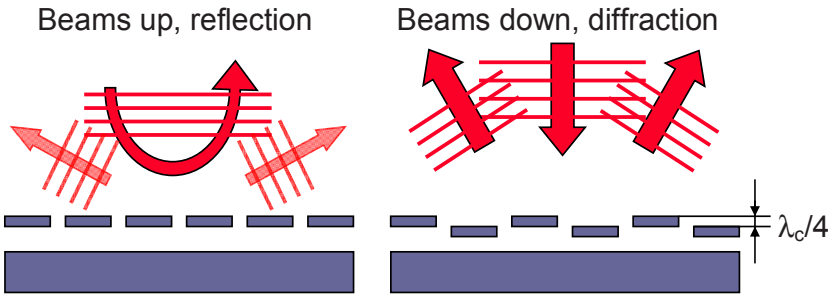


Figure 10.8. Grating modulator designed for high-contrast operation. In the reflective state, the ribbons are all at the same plane, and the reflections from the ribbons are all in phase. In the diffractive state, every other ribbon is actuated to create a path length difference of π radians.

In a practical implementation, there will have to be some space between the ribbons. This creates a weak grating of half the period. Some light will be diffracted from this parasitic grating, but the diffraction angle is twice that of the fundamental diffraction order, so the diffracted light of the reflective state can be separated from the diffracted light of the diffractive state.

The primary advantage of the design shown in Fig. 10.8 is that the reflective state is non-dispersive. There is no path length difference in this state, so all wave-

lengths interfere constructively in reflection. Other than the diffraction from the parasitic grating of half the period, there is no diffracted light from this state, even if broad band illumination is used. This complete absence of diffraction allows us to make a projection system that can achieve very high contrast. The projection principle, which is very similar to the Schieren system we have already introduced, is shown in Fig. 10.9.

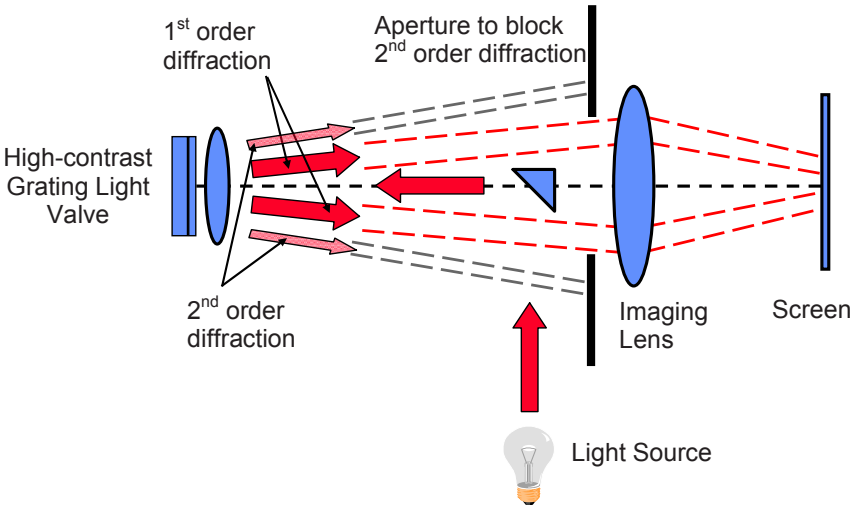


Figure 10.9. Schliern projection system using high-contrast grating light modulators. Grating modulators in the diffractive state create bright pixels on the screen, while reflecting modulators create dark pixels. In the reflecting state the grating produces very little diffracted light, other than the 2nd order diffraction from the parasitic grating, so the pixels are dark over the full spectrum of the illuminating light source, leading to high contrast.

As before, we have that grating light modulators in the reflective state will reflect most the light back towards the illuminating light source. Some light is diffracted by the parasitic grating, but its diffraction angle is twice that of the fundamental grating. This second-order diffracted light can be blocked from the screen by an appropriately placed aperture as shown. The reflective state is therefore for practical purposes non-dispersive, and it creates pixels with very uniform and low illumination cross the spectrum of the light source.

In the diffractive state, which is used to create bright pixels on the screen, the operation of the high-contrast grating modulator is essentially the same as that of the simpler modulator described above. In particular, the pathlength difference is half a wavelength at the center wavelength, so the dispersion characteristics of the

high-contrast grating light modulator in the diffractive state is as shown in Fig. 10.5a and 10.6b. There is some variation in the diffracted light as a function of wavelength, and that will lead to a reduction of the maximum light level of each pixel, but the reduction in brightness is relatively small, and it does not lead to significant decrease of the contrast.

As shown in Fig. 10.9, only the 1st order diffracted light reaches the screen, which means that the effective diffraction efficiency is about 80% as we will see from the more detailed model that we will develop later. In principle, this can be improved by configuring the aperture to block the 2nd order diffracted light, but pass 3rd and higher orders.

In addition to its outstanding optical properties, the high-contrast grating modulator also has some distinct mechanical advantages, stemming from the fact that the distance of the movable ribbons over the substrate is not a critical parameter. In the basic grating light modulator, this distance equals one quarter of the center wavelength, and this small distance makes it difficult to operate in an analog, non-contact mode. Using a distance that is one quarter plus an integer number of half center wavelengths is a possible solution, but that comes at the cost of dramatically increased dispersion. The high-contrast modulator, on the other hand, can be designed with a gap that can be optimized for actuation purposes, making it straight forward to control the ribbon position with high accuracy.

10.5 Diffraction Gratings

The simple phasor model is useful for understanding some aspects of grating light modulator operation. We have used it to explain the differences between the basic grating modulator and the high-contrast grating modulator, and we will use it again to explain the dispersive characteristics of more complex modulators used in fiber optics. The phasor model leaves unanswered a number of critical questions about the grating light modulators, however: Where does the diffracted light go? How long and wide should the ribbons be and how many do we need? To answer these questions we need a detailed mathematical model of diffraction from a phase grating as shown schematically in Figure 10.10. Once we have such a model, we will be able to develop design equations for grating modulators and other diffractive Optical MEMS devices.

In our treatment we will concentrate on reflective gratings because these are more easily implemented than transmission gratings in MEMS technology. The equations we develop are, however, quite general. With minor modifications they can be applied to transmission gratings and other, more complex, periodic structures.

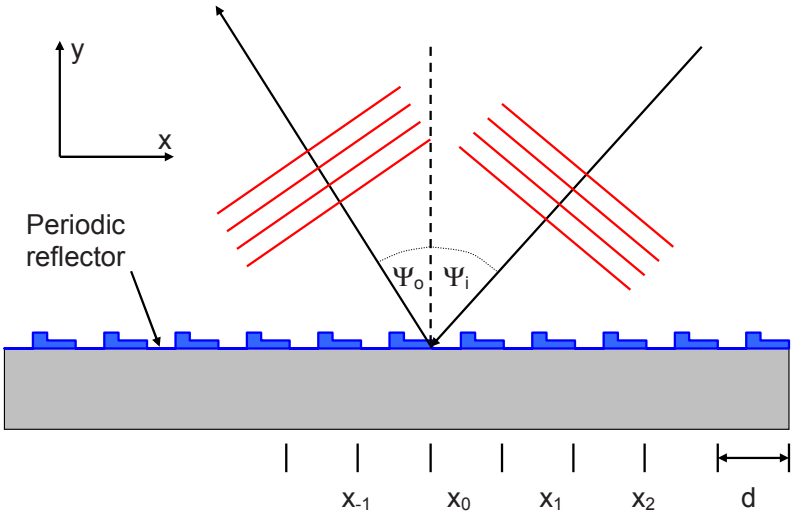


Figure 10.10 *Schematic drawing of a periodic reflection grating. The grating is modeled as uniform in the z-direction (perpendicular to the plane of the drawing) and periodic in the x-direction. Each period has a reflectance with a spatially varying phase delay, θ(x). In general the absolute value of the reflectance also varies spatially, although we will concentrate on phase-only gratings.*

The well-known grating equation gives a partial answer to the question of where the diffracted light goes. This equation gives the directions in which the reflections from each period of the grating add in phase, i.e. the direction in which the path length difference for the reflections from neighboring periods of the grating is an integer number of wave lengths. From Fig. 10.11 we see that this condition of path length difference is met when:

$$DC - AB = m \cdot \lambda \Rightarrow d \cdot \sin \Psi_1 - d \cdot \sin \Psi_0 = m \cdot \lambda \Rightarrow \tag{10.6}$$

$$\sin \Psi_1 - \sin \Psi_0 = \frac{m \cdot \lambda}{d} \tag{10.7}$$

where Ψ_1 and Ψ_0 are the incident and diffracted angles, d is the grating period, λ is the wavelength of the incident monochromatic light, and m is any integer (positive, negative or zero). The incident and reflected/diffracted angles can be combined to a single angle parameter that is defined as

$$p = \sin \Psi_1 - \sin \Psi_0 \tag{10.8}$$

This combination of the incident and diffracted angles into a single parameter is very convenient in that it greatly simplifies the grating equation and related equations that we will derive later.

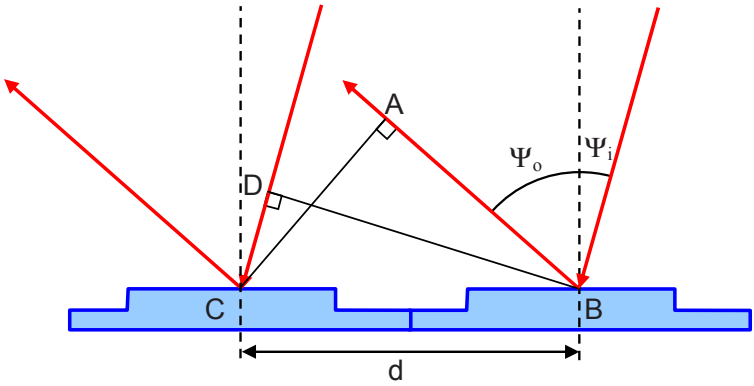


Figure 10.11 *The schematic shows to adjacent periods, or unit cells, of a reflection grating. Reflections from adjacent unit cells interfer constructively for combinations of incident and diffracted angles that make the path length differences (DC-AB) equal to integer number of wavelengths. This principle is expressed in the grating equation (Eq. 10.7).*

The grating equation tells us some of what we need to know to design grating light modulators. In particular, it shows in what directions light will be diffracted, and underscores the importance of the period of the grating. It does not specify how much light goes into the different possible diffraction directions, however, and it therefore gives us no information on how to design and modulate the individual unit cells to achieve a desired angular distribution of light. To get a more complete, but still simple, analytical expression for the diffraction characteristics of such gratings, we use Fraunhofer diffraction theory [7]. This theory is valid in for diffraction from periodic objects with unit cells that are large compared to the wavelength of the incident light.

Fraunhofer diffraction theory treats the optical field as a scalar, i.e. it neglects the vector nature of the electromagnetic fields. Consequently, it does not make predictions about polarization sensitivity of diffraction or other polarization effects of gratings. Such effects become increasingly pronounced as we scale down the unit cells of the grating. It turns out, however, that although exact results require a more complete and complex theory, the simple formulas given by Fraunhofer diffraction are quite accurate and useful even for structures where dimensions of the smallest features approach the wavelength.

In later chapters we will discuss Phonic Crystals that are periodic structures with periods on the order of the wavelength or below. In such structures, the directions of the electromagnetic fields are important, and in addition, there are substantial amounts of coupling of the incident light to optical modes of the grating. Photonic Crystals can therefore not be modeled accurately with Fraunhofer diffraction theory, or any other theory that does not take into consideration polarization effects

and interactions of the incident and diffracted optical fields with optical modes of the grating (or Photonic Crystal) itself.

In our analysis of diffraction from a reflective MEMS grating, we assume monochromatic, plane wave illumination of a grating that is periodic in the *x* dimension, and uniform in the *z* dimension as shown in Fig. 10.10. The Fraunhofer diffraction integral can then be written

$$E(p) = C \cdot \int_{Grating} r(x) \cdot e^{-jkpx} dx \tag{10.9}$$

where *E(p)* is the optical field of a plane wave propagating away from the grating in the direction given by the angle parameter *p*, *C* is a normalizing constant, *r(x)* is the optical field reflection of the grating, $k = 2\pi/\lambda$ is the propagation constant, and λ is the wavelength of the incident light.

Motivated by the grating equation, we use the angle parameter $p = \sin\Psi_1 - \sin\Psi_0$ to specify the incident and reflected/diffracted angles. As before, this combination of the incident and diffracted angels into one parameter greatly simplifies the analytical expressions, but it is an approximation that comes at a price. For many structures of interest the diffraction efficiency, and other important characteristics, will depend on the incident and diffracted angles independently, and not only on the differences in the values of their sine functions. A good example of a device where this must be taken into consideration is the well know blazed grating. This will become important for our discussions of MEMS blazed gratings for optical filtering. For now, however, we will consider gratings operated at near-normal incidence, so that the grating reflection function *r(x)* is independent of the incident angle Ψ_1. Under these circumstances it is very useful to combine the incident and diffracted angles into the parameter *p*.

We now use the fact that the grating transfer function is periodic to rewrite the Fraunhofer integral as the response of a single grating period multiplied by a sum that gives the relative phase response of the individual elements of the grating:

$$E(p) = C \cdot \int_{Period} r(x) \cdot e^{-jkpx} dx \cdot \sum_{n=0}^{N-1} e^{-jkp \cdot nd} \tag{10.10}$$

where *N* is the number of periods in the grating and the transfer function, *r(x),* now is integrated over only one period. The sum has a closed-form solution, so that the expression can be simplified to

$$E(p) = \frac{1 - e^{-jNkpd}}{1 - e^{-jkpd}} \cdot C \cdot \int_{Period} r(x) \cdot e^{-jkpx} dx \tag{10.11}$$

In principle, the reflection can have any spatially varying magnitude and phase across the period of the grating, but for now we will consider an idealized grating with a field reflection of uniform absolute value and a phase that varies spatially in a binary fashion across the unit cell of the grating. This is a good approximation for a MEMS grating consisting of metal reflectors that can be positioned at different height levels. The reflection function of one grating element can then be expressed as

$$r(x)=\begin{cases} |r|\cdot \mathrm{e}^{-j\frac{\theta}{2}} & \text{for} \quad -\dfrac{d}{2}<x<-\dfrac{d}{4} \text{ or } \dfrac{d}{4}<x<\dfrac{d}{2} \\ |r|\cdot \mathrm{e}^{j\frac{\theta}{2}} & \text{for} \quad -\dfrac{d}{4}<x<\dfrac{d}{4} \end{cases} \tag{10.12}$$

Here $|r|$, which represents the uniform field reflectivity from all parts of the grating, will have a value between zero and unity. For typical MEMS implementation with Aluminum-coated mirrors, the reflectivity is about *0.9*. The angle θ is the phase delay created by the height differences on the grating surface. It is of course wavelength dependent, which we have to take into consideration when we treat broad band light. For now we are interested in monochromatic light, so the phase angle θ is a constant.

Given this reflection function, the Fraunhofer integral over one period, *d*, evaluates to

$$E_d(p)=C\cdot \int_{Period} r(x)\cdot e^{-jkpx}dx$$

$$=C\cdot|r|\cdot\left[\int_{-\frac{d}{2}}^{-\frac{d}{4}} e^{-j\frac{\theta}{2}}\cdot e^{-jkpx}dx+\int_{-\frac{d}{4}}^{\frac{d}{4}} e^{j\frac{\theta}{2}}\cdot e^{-jkpx}dx+\int_{\frac{d}{4}}^{\frac{d}{2}} e^{-j\frac{\theta}{2}}\cdot e^{-jkpx}dx\right]=$$

$$C|r|\cdot\left[e^{-j\frac{\theta}{2}}\cdot\frac{e^{-jkp\frac{d}{4}}-e^{-jkp\frac{d}{2}}}{-jkp}+e^{j\frac{\theta}{2}}\cdot\frac{e^{jkp\frac{d}{4}}-e^{-jkp\frac{d}{4}}}{-jkp}+e^{-j\frac{\theta}{2}}\cdot\frac{e^{jkp\frac{d}{2}}-e^{jkp\frac{d}{4}}}{-jkp}\right] \tag{10.13}$$

$$=d\cdot C\cdot|r|\cdot\left[\cos\frac{\theta}{2}\cdot\frac{\sin(kpd/2)}{kpd/2}-j\cdot\sin\frac{\theta}{2}\left(\frac{\sin(kpd/2)}{kpd/2}-\frac{\sin(kpd/4)}{kpd/4}\right)\right]$$

The intensity or irradiance of a uniform optical wave is proportional to the field multiplied by its complex conjugate (or the square of the norm) and can be expressed as

$$I(p) \propto E(p)E^*(p) =$$

$$|E_d(p)|^2 \frac{1-e^{-jNkpd}}{1-e^{-jkpd}} \frac{1-e^{jNkpd}}{1-e^{jkpd}} = |E_d(p)|^2 \left(\frac{\sin \frac{Nkpd}{2}}{\sin \frac{kpd}{2}} \right)^2 \tag{10.14}$$

$$I(p) \propto d^2 \cdot |C|^2 \cdot |r|^2 \cdot$$

$$\left[\cos^2 \frac{\theta}{2} \cdot \frac{\sin^2 \frac{kpd}{2}}{\left(\frac{kpd}{2} \right)^2} + \sin^2 \frac{\theta}{2} \cdot \left(\frac{\sin \frac{kpd}{2}}{\frac{kpd}{2}} - \frac{\sin \frac{kpd}{4}}{\frac{kpd}{4}} \right)^2 \right] \cdot \frac{\sin^2 \frac{Nkpd}{2}}{\sin^2 \frac{kpd}{2}} \tag{10.15}$$

Finally we can write the following equation for *D(p),* the intensity reflectivity, or intensity diffraction, from the grating in the direction given by *p*:[a]

$$D(p) =$$

$$R \cdot \frac{\sin^2 \frac{Nkpd}{2}}{N \cdot \pi \cdot \sin^2 \frac{kpd}{2}} \left[\cos^2 \frac{\theta}{2} \cdot \mathrm{sinc}^2 \frac{kpd}{2} + \sin^2 \frac{\theta}{2} \cdot \left(\mathrm{sinc} \frac{kpd}{2} - \mathrm{sinc} \frac{kpd}{4} \right)^2 \right] \tag{10.16}$$

We note that the equation consists of three parts. The first part, *R*, simply says that the diffracted optical power is scaled by the reflectivity of the grating. This is of course true for any grating with uniform reflectivity.

The second part of the equation is the periodic function $\frac{\sin^2(Nkpd/2)}{N \cdot \pi \cdot \sin^2(kpd/2)}$ that we will call the grating function. It describes the response due to the periodic structure of the grating. It does not contain any information about the unit cells of the grating, so this factor will be the same for all gratings with a given period, *d*, and number of grating elements, *N*.

The grating function is plotted in Fig. 10.12. It is periodic in the parameter $kpd/2$ with a period of π. The local maxima in each period of the grating function are called the diffraction orders of the grating. The maximum at $kpd/2 = 0$,

[a] We are using the following definition of the *sinc* function: $sinc(x) = \sin x / x$. The reader should be aware that the alternative notation, $sinc(x) = \sin(\pi \cdot x)/\pi \cdot x$, is also used in many texts.

which is really the specular reflection from the grating, is also called the 0^{th} diffraction order of the grating. The maxima at $kpd/2 = \pm\pi$ are the $\pm 1^{st}$ diffraction orders, the maxima at $kpd/2 = \pm 2\pi$ are the $\pm 2^{nd}$ diffraction orders, and so on up to the $\pm n^{th}$ orders at $kpd/2 = \pm n \cdot \pi$.

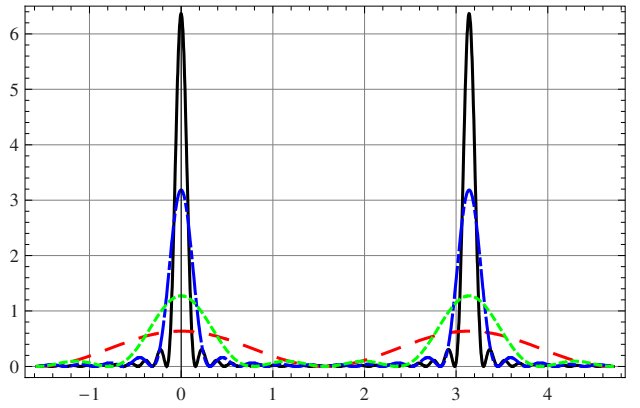


Figure 10.12. Plot of the grating function $\dfrac{\sin^2 Nx}{N \cdot \pi \cdot \sin^2 x}$ *as a function of x for N= 2 (dashed), 4 (dotted), 10 (dot-dashed), and 20 (solid),. The function is periodic with a period of π, and its integral over one period is unity, independent of N. In the limit of large values of N, it approaches a comb function.*

The integral of the grating function over one period has a value of unity independent of N[b]. This means that for large N, the grating function approaches a set of delta functions at the angular positions $kpd = n \cdot 2\pi$, where n is an integer. In other words, the grating function is a comb function for large N. This is what we

[b] The integral of the grating function over one period is simple to evaluate once we realize that the indefinite integral of $\sin^2 Nx / \sin^2 x$ can be written as $Nx + \sum_{i=1}^{2N-2} a_i \cdot \sin(i \cdot 2x)$, where the specific values of the coefficients a_i are functions of N. The exact values of the coefficients are immaterial, because all the elements of the sum integrate to zero over one period. The value of the grating function integrated over one period is then:

$$\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{\sin^2 Nx}{N \cdot \pi \cdot \sin^2 x} dx = \frac{1}{N \cdot \pi} \left[Nx + \sum_{i=1}^{2N-2} a_i \cdot \sin(i \cdot 2x) \right]_{-\frac{\pi}{2}}^{\frac{\pi}{2}} = 1$$

expect from a large grating. We get diffraction orders that are very close to plane waves at well defined angles.

For MEMS implementations, we are more interested in the other extreme. We would like to miniaturize the gratings as much as possible, so it is important to know how few periods we can have and still get good separation of the diffraction orders. We see that the separation of the diffraction orders is $\Delta p_{separation}=\lambda/d$, and the width of an order to the first nulls is $\Delta p_{width}=2\lambda/Nd$, so the ratio of separation to width equals $N/2$. From a miniaturization point of view, this is good news, because a grating with as few as 3 periods is sufficient to create diffraction patterns with well defined and clearly separated diffraction orders.

From this discussion we see that the grating function we have derived contains the same information about the diffraction pattern as the grating equation. It gives us the directions of the diffraction orders, but no information about how much of the incident optical power is diffracted into the different orders. That information is contained in the third part of Eq. 10.16, which we will call the modulation function.

The modulation function, $\cos^2\frac{\theta}{2}\cdot\mathrm{sinc}^2\frac{kpd}{2}+\sin^2\frac{\theta}{2}\cdot\left(\mathrm{sinc}\frac{kpd}{2}-\mathrm{sinc}\frac{kpd}{4}\right)^2$, specifies how the distribution of light between the diffraction orders depends on the phase shift θ. When the phase delay is zero ($\theta=0$), which means that the grating is really nothing more than a flat mirror, the modulation function becomes $sinc^2(kpd/2)$. In this case, the reflectivity in the 0^{th} order is given by the material reflectivity R, and the reflectivity is zero in all other orders, as we would expect from a flat mirror. For $\theta=\pi/2$, the modulation function is zero in the reflected, or 0^{th}, order, so all the light is diffracted into the two 1^{st} order and higher diffraction modes.

This is the same as what we realized using our phasor model of the grating light modulator, but now we are in a position to get a more complete picture of how the light is distributed between diffraction orders, which is something we need to know to understand and design grating light modulators. We start by applying the Fraunhofer grating equation (Eq. 10.16) to the simplest possible case, which is that of a large grating, i.e. a grating with a large number of elements ($N \to \infty$). We know that in this case the ratio in front of the parenthesis is a comb function with delta functions at $\sin(kpd/2)=0$. The diffraction (or more correctly reflection) in the zeroth order diffracted mode ($kp = 0$) is then

$$D^0(kp=0)=I_0\cdot\cos^2\frac{\theta}{2}=\frac{I_0}{2}\cdot(1+\cos\theta) \tag{10.17}$$

One remarkable property of phase gratings is being made very clear by this expression. Unlike amplitude gratings, where the minimum intensity in the zero order is one half of the incident intensity, a phase grating can diffract all the intensity out of the zero order. This is very important in practical applications, because if the diffraction from a grating is less than complete, then any modulator based on diffraction from that device will have reduced optical efficiency.

The intensity in the n^{th} order mode, where n is larger than zero, is given by

$$D^n\left(kp = n\frac{2\pi}{d}\right) = I_0 \cdot \sin^2\frac{\theta}{2} \cdot \mathrm{sinc}^2\frac{n\pi}{2} = \frac{I_0}{2} \cdot (1-\cos\theta) \cdot \mathrm{sinc}^2\frac{n\pi}{2} \qquad (10.18)$$

Notice that only odd order modes exist. Again this is different from the more familiar amplitude grating result. Table 10.1 gives the numerical values of the factor $sinc^2(n\pi/2)$, as well as the accumulated diffracted power in all orders up to the n^{th}.

n	1	3	5	7	9	11
$\mathrm{sinc}^2\frac{n\pi}{2}$	0.405285	0.045032	0.016211	0.008271	0.005004	0.003349
$2\cdot\sum_{i=1}^{n}\mathrm{sinc}^2\frac{i\pi}{2}$	0.810569	0.900633	0.933056	0.949598	0.959605	0.966304

Table 10.1. Relative power in the n^{th} diffracted order and relative power in all modes up the n^{th} for a binary phase grating in its diffractive state.

We see that an optical system that uses the combination of the two first order diffraction modes to create a bright state has at best a through put of 81%. This is the case for the Schlieren projection system of Fig. 10.3. If the system also picks up the two third order diffraction states, the efficiency increases to 90%. To go substantially beyond the 90% given by the combined 1st and 3rd orders takes heroic efforts. The table shows that we have to include all orders up to the ±7th to get 95%, and to get to 99% we have to include all odd orders up to the 27st!

In practical systems we typically use only the first diffraction orders and for some special cases the third orders might be included, but we seldom go beyond that. The exception is high-quality diffractive lenses (which are of course also grating governed by equations similar to the one we have found for rectilinear gratings), where we use multiple phase steps to align several diffraction orders.

Table 10.1 gives the relative diffracted power of our deformable binary phase grating when it is in its diffractive state, i.e. the phase angle is $\theta=\pi/2$. For the general case of $0<\theta<\pi/2$, the sum of the diffraction into all modes (including the 0^{th} order or reflected mode) becomes

$$R_{total} = R \cdot \left(\cos^2 \frac{\theta}{2} + 2 \sin^2 \frac{\theta}{2} \cdot \sum_{n=1}^{\infty} \mathrm{sinc}^2 \frac{n\pi}{2} \right)$$
$$= R \cdot \left(\cos^2 \frac{\theta}{2} + \sin^2 \frac{\theta}{2} \cdot \frac{8}{\pi} \cdot \sum_{n=1}^{\infty} \frac{1}{(2n-1)^2} \right) \tag{10.19}$$

Combined with the well known result from Fourier expansion; $\frac{\pi^2}{8} = \sum_{n=1}^{\infty} \frac{1}{(2n-1)^2}$, we see that R_{total} evaluates to R as it should.

Based on the insight that the modulation function provides on the distribution of the optical power between diffraction orders, we can provide a more accurate graph of diffracted intensity vs. the relative phase shift between the two parts of the grating light modulator as shown in Fig. 10.13. Here we have plotted the relative intensity in reflection and in different diffracted orders relative to the incident optical intensity as a function of the relative phase difference between the reflections from the different parts of the grating. There are three curves for the diffracted light; one for only one 1st order by itself, one for the combination of the two 1st orders, and on for the two 1st orders combined with the two 3rd orders.

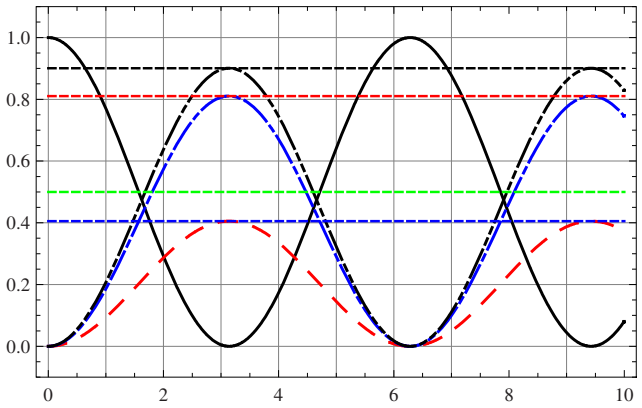


Figure 10.13. Total reflected light (solid), diffracted light in one 1st order (dashed), diffracted light in both 1st orders (dot-dashed), and diffracted light in both 1st and both 3rd orders (double-dot-dashed) from a grating modulator as a function of relative phase difference of the two parts of the reflected light. In all cases the optical powers are harmonic functions of the phase difference.

We see that all the diffracted orders have the same dependence on the phase difference, θ. This means that when the reflected light is at its maximum, all diffraction orders are at zero, and when the reflected light is zero, all diffraction orders are at their maxima. It is therefore indeed possible to create an optical system that

capture several diffraction orders and maintains good contrast with the reflected light. We also realize, however, that the different diffraction orders propagate in different directions, so capturing more diffraction orders requires increasingly complex optics and gives diminishing returns.

The optimum tradeoff between optical efficiency and complexity is of course completely dependent on the application, but as a general rule we can say that capturing only 40% of the light is too little and capturing more than the 3^{rd} orders is too much complexity for the relatively small return. Practical systems therefore either use both 1^{st} orders or combine both 1^{st} and 3^{rd} orders.

Figure 10.13 illustrates what we already have learned from the phasor model, namely that we need a phase shift of $\pi/2$ radians to get high-contrast switching. That is, however, not the only useful way to operate a deformable grating. It is also possible to use the grating as a displacement sensor. Instead of changing the relative position of the two elements of the grating unit cell by an actuator, we design the grating such that this displacement is a function of an external measurand. In principle, the measurand can be anything, e.g. pressure, force, acceleration, rotation, or magnetic fields, or biomolecular associations. In any case, we would like to be able to measure the smallest possible deflection to create a system with good sensitivity. In other words, we would like to create the biggest possible change in optical output, be it reflection of diffraction, for the smallest possible phase shift. In this case, we would like to operate at those points on the curve where the light vs. phase slope is the highest, i.e. at those points where the phase difference is $K \cdot \pi/4$ with K an odd integer. Some of these high-sensitivity points, useful for sensor operation, are marked on Fig. 10.13.

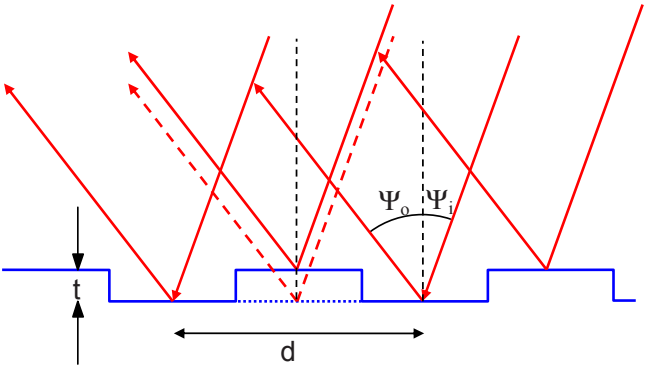


Figure 10.14 Reflections from two adjacent unit cells of a deformable grating modulator with grating amplitude t. The phase modulation is caused by the physical separation of the two reflectors in each unit cell, so it is dependent on the wavelength as well as the incident and refracted angles.

In the above derivation, the phase grating is assumed to be taking place in a plane. In reality the deformable grating modulator is a three dimensional structure, and the phase modulation is obtained by moving one reflective surface with respect to another. The phase modulation will therefore be a function of the wavelength of the incident light and of the angles of incidence and diffraction as shown in Fig. 10.14. Note, however, that it is not dependent on the grating period, *d*.

The phase difference is simply given by

$$\theta = \frac{2\pi \cdot \Delta}{\lambda} \tag{10.20}$$

where λ is the optical wavelength and Δ is the path length difference. At normal incidence, the path-length difference simply equals twice the grating amplitude, i.e. $\Delta_{normal}=2t$. At non-normal incidence, however, we must be careful to take account not only of the vertical difference between the two reflections, but also of the lateral shift of the optical beams.

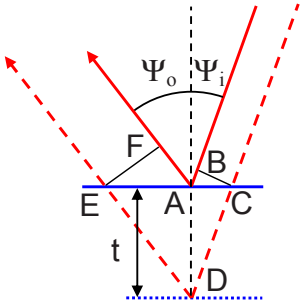


Figure 10.15 *Close up of the deformable grating modulator. Because it is caused by physical displacement, the phase modulation is dependent on the incident and refracted angles.*

From Fig. 10.15 we see that the total path length difference for the light that is reflected from the top and bottom levels of the grating can be expressed as:

$$\Delta = CD + DE - AB - AF = CD + DE - CD \cdot \sin^2 \Psi_i - DE \sin^2 \Psi_o \Rightarrow$$

$$\Delta = \frac{t}{\cos \Psi_i} - \frac{\sin^2 \Psi_i}{\cos \Psi_i} + \frac{t}{\cos \Psi_o} - \frac{\sin^2 \Psi_o}{\cos \Psi_o} = t\left(\cos \Psi_i + \cos \Psi_o\right) \tag{10.21}$$

Combining these equations, we can write the following expression for the phase difference:

$$\theta = \frac{2\pi \cdot t}{\lambda} \cdot \left(\cos \Psi_i + \cos \Psi_o\right) \tag{10.22}$$

Once the lateral shift of the optical beams is properly accounted for, we get the opposite dependence on the cosines than what we might expect! This thickness dependence is ubiquitous in optical interference phenomena, e.g. interference from soap films and other structures with multiple reflections separated by a thin spacer layer, so it is well worth making a note of.

The expression we have found for the phase modulation is valid for relatively small incident and diffraction angles. It is clear from Fig. 10.14 that at larger angles, we must worry about amplitude variations caused by shadowing effects in the three dimensional grating structure. At even larger angles, evanescent modes become important and the Fraunhofer theory breaks down completely. For the relatively small incident and diffraction angles and small grating amplitudes relative to the grating period of practical grating modulators, the Fraunhofer diffraction theory is in good agreement with experiments.

The grating equation and the equation for the phase difference show that both the angular directions of the diffracted light and the ratio of reflected to diffracted light is dependent on wavelength for all finite values of the phase difference θ. In other words, a grating light modulator in any configuration other than the zero path-length-difference state is wavelength dependent. This means that all configurations other than $\theta=0$ will suffer a reduction of contrast if the light source is not monochromatic.

Wavelength dependence is therefore problematic in many GLM applications, but, like most problems, it can be turned into a feature and used to good advantage. Specifically, their wavelength dependence allows grating modulators to be used as tunable optical filter and tunable optical synthesizers. This is illustrated in Fig. 10.16 that shows a grating modulator used as a color filter. The grating is set up to diffract light at a specific center wavelength. At shorter wavelengths there is finite reflection, and the diffraction angles are smaller. Longer wavelengths also have finite reflection, while the diffraction angles are larger. In this chapter and the next we will discuss several instances in which wavelength dependence is a problem requiring a solution, while we will postpone the detailed treatment of designs and characteristics of grating modulators optimized for spectral filtering to Chapter 13.

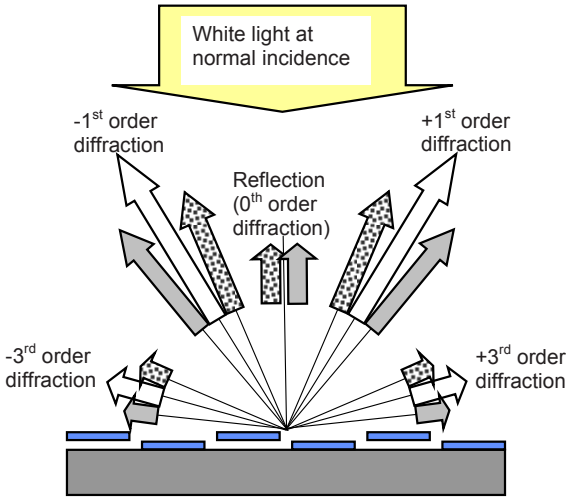


Figure 10.16 *Diffraction from a grating modulator at three different wavelengths. The modulator is set up to diffract all the light at the center wavelength (solid-white). Most of the light at this wavelength comes off the modulator in the ±1st and ±3rd diffraction modes (there is some light in higher order modes). Light at slightly longer wavelengths (solid-gray) is partially diffracted and partially reflected, and the diffraction angles are larger than for the central wavelength. Light at slightly shorter wavelengths (dotted) is also partially diffracted and partially reflected, but the diffraction angles are smaller. By correct configuration of apertures, the dispersive grating can be used to create a color filter.*

We can now summarize what we have learned about grating modulator design from Fraunhofer diffraction theory:

First we verified the insight we had gained from our phasor model that a relative displacement of only a quarter wavelength is sufficient to completely switch the grating modulator between its reflective and diffractive states. We also found that as little as three periods of a grating is sufficient to create a high-quality modulator. This is good news, because it means that grating modulators can be miniaturized in both its in-plane and vertical dimensions!

The third thing we learned is that the diffracted light exits the grating in different directions given by the well-known grating equation: $\sin\Psi_1 - \sin\Psi_0 = \frac{m \cdot \lambda}{d}$, and that any one order do not contain more than about 40% of the incident light. This clearly complicates the systems based on grating light modulators, because we need to control several diffraction orders to achieve high optical efficiency. This shortcoming of the basic grating modulator becomes one of the driving forces behind innovation in grating modulator design.

The grating modulator manipulates light by changing the path length for part of the light that is being diffracted. This dependence on path length can be turned around and used as a sensing mechanism. The Fraunhofer model shows how the grating should be constructed to achieve the highest possible sensitivity to displacement. We will use this to analyze and design deformable grating sensors in Chapter 12.

The last thing Fraunhofer theory taught us about grating modulators is that they are dispersive, i.e. they are wavelength dependent. This follows from the fact that we create phase modulation in the grating modulator by establishing path length differences for different parts of the incident light. A constant path length, or time delay, leads to a wavelength dependent phase difference. This dispersion, just as the existence of multiple diffraction orders, complicates systems design and becomes a driver for innovation. The phase difference is also angle dependent, but that is usually not of practical significance.

10.6 Projection Displays Based on Grating Modulators

Now we have covered the basics of grating modulators in sufficient detail that we can turn our attention to practical implementations. The detailed MEMS designs are of course strongly dependent on applications. Here we will first consider grating modulators for displays, and then go on to discuss how to optimize the modulators design for fiber optical applications.

10.6.1 Actuator Design

Our grating analysis led to the development of two important equations for reflectivity and phase shift:

$$R_{total} = R \cdot \left(\cos^2 \frac{\theta}{2} + \sin^2 \frac{\theta}{2} \cdot \frac{8}{\pi} \cdot \sum_{n=1}^{\infty} \frac{1}{(2n-1)^2} \right) \tag{10.23}$$

$$\theta = \frac{2\pi \cdot t}{\lambda} \cdot \left(\cos \Psi_i + \cos \Psi_o \right) \tag{10.24}$$

These equations show that a displacement of a quarter wavelength (corresponding in reflection to a phase shift $\frac{2\pi}{\lambda} \cdot \frac{\lambda}{4} = \pi$ radians) is sufficient to complete switch a grating modulator between its reflective and its diffractive states. For visible wavelengths centered around *500 nm*, this means a displacement on the order of

125 nm, and for fiber-optic communication wavelengths around *1550 nm*, a displacement of about *400 nm*. The lateral dimensions of the individual ribbons of the grating must be slightly larger than the wavelength of the light for the equations to be valid, but we need only three period (3 ribbon pairs or 6 ribbons) to create a grating modulator with good separation of the diffracted and reflected light, so as a rule of thumb the lateral dimensions should be on the order of, or larger, than ten wavelengths.

Their small in-plane, and even smaller vertical dimensions make grating modulators at visible and near-IR wavelengths very compatible with IC and MEMS technology. Most materials used in MEMS can be deposited as thin films, ranging in thickness from hundreds of nanometer to several micron. The thickness control and uniformity, both across any given wafer and between wafers, are excellent over this range. Furthermore, the surface roughness of most as-deposited thin films is also sufficient to yield high-quality optics. The exceptions here are films of some polycrystalline materials, like poly-Si and poly-Ge, that might create problems due to their surface roughness. The bottom line is that IC and MEMS technology can easily meet the demands on vertical structural accuracy set by grating modulators.

Likewise, the lateral definition of the grating represents no challenge for modern lithography, which at present is pushing below the *100 nm* barrier. If we use our rule of thumb of ten wavelengths, we see that an array of a one million grating modulators only occupies *5 by 5 mm* in the visible and *15 by 15 mm* at fiber-optic wavelengths. Large arrays can therefore readily be accommodated on chips of modest size, which not only reduce chip price, but also the cost of packaging.

The excellent compatibility with standard IC technology let us have our pick of materials without having to worry about problems with dimensional control, so we are free to design our ribbons structures based on functional considerations. Clearly the small displacement that is required favors a simple parallel-plate electrostatic actuator. Electrostatic actuation does not require special materials beyond standard conductors and insulators to control the electrostatic potentials, and they are therefore material compatible with IC and MEMS technology. The small required displacements also mean that the electrode gap in the electrostatic actuators can be small, so that large electrostatic forces can be created with relatively small voltages. There is therefore no need for more complex and difficult-to-fabricate structures, e.g. combdrives, whose primary purpose is to extend the travel range of electrostatic actuators.

To get an idea of how a simple parallel-plate actuator will perform in a grating light modulator we will use a simple spring mass model as shown in Fig. 10.17. The ribbon is modeled as a conducting plate of mass $m=\rho \cdot A \cdot t$, where ρ is the density of the ribbon material, A is the area of the ribbon and t is its thickness.

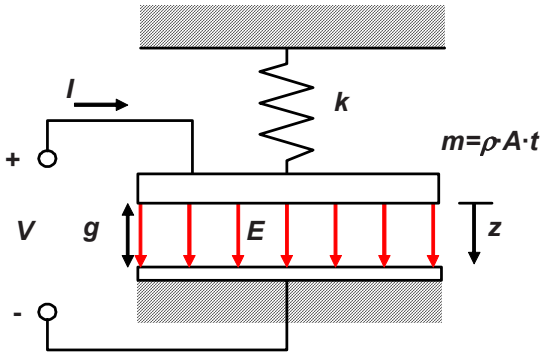


Figure 10.17 *Spring-mass model of a grating-modulator ribbon actuated by a parallel-plate electrostatic actuator. The lower electrode is fixed to the substrate, while the upper electrode, the ribbon, is suspended on a spring and can move in response to the applied electrostatic force.*

The purpose here is to get an overview of the design space, so we will use any reasonable approximation that simplifies the derivations. We start with the approximation that the fields are uniform between the electrodes and zero outside, so we can write the following expression for the electrostatic force:

$$F = \frac{Q^2}{2\varepsilon A} = \frac{V^2 A\varepsilon}{2g^2} \tag{10.25}$$

where ε=8.85·10^{-12} Fm^{-1} is the dielectric constant (the exact value of the dielectric constant is ε= 8.854,187,817·10^{-12} Fm^{-1} by definition), A is the ribbon area, g is the ribbon-substrate gap, and V is the applied voltage.

The resonance frequency of the spring-mass system is

$$f = \frac{1}{2\pi}\sqrt{\frac{k}{m}} \tag{10.26}$$

where k is the spring constant of the ribbon suspension, and m is the ribbon mass. We assume that the flexibility and precision of our MEMS technology allow us to give the spring constant any value, so we chose to set it equal to the maximum value that enables the required $\lambda/4$ displacement with the maximum voltage, V_{max}, applied.

$$f = \frac{1}{2\pi}\sqrt{\frac{F_{\max}}{\frac{\lambda}{4}\cdot m}} = \frac{1}{2\pi}\sqrt{\frac{4}{\lambda\cdot\rho\cdot A\cdot t}\frac{V_{\max}^2 A\varepsilon}{2g^2}} = \frac{V_{\max}}{2\pi}\sqrt{\frac{2\varepsilon}{\rho\cdot\lambda\cdot g^2\cdot t}} \tag{10.27}$$

The ribbon-substrate gap is a strong function of the wavelength, because it must at a minimum equal one quarter of the wavelength. We will distinguish between two cases: In what we will call the contact mode, the gap equals one quarter of the wavelength, which means that the ribbons touch the substrate in their maximum-deflection state. Clearly it is not practical to allow the two electrodes to touch, but they can be separated by a thin insulator. In the second mode of operation, which we will call continuous mode, the ribbon-substrate gap is more than three times the maximum deflection, so that the ribbon can be electrostatically positioned at any deflection between zero and one quarter wavelength without problems with electrostatic instability.

To build in a safety margin and to keep the formulas simple, we will say that the gap equals the wavelength, i.e. it is four times the maximum deflection, in the continuous mode. The resonance frequency can then be expressed as:

$$f = \frac{V_{\max} \cdot D_{\text{mode}}}{2\pi \cdot \lambda^{3/2}} \sqrt{\frac{2\varepsilon}{\rho \cdot t}} \tag{10.28}$$

where the constant D_{mode} is 4 for contact mode and 1 for continuous mode. Assuming that the ribbons are close to critically damped, the 10%-to-90% switching time for the ribbons is related to the resonance frequency through the following expression:

$$t_{10-90} = \frac{0.35}{f} = \frac{2\pi \cdot \lambda^{3/2}}{V_{\max} \cdot D_{\text{mode}}} \sqrt{\frac{\rho \cdot t}{2\varepsilon}} \tag{10.29}$$

This expression is plotted in Fig. 10.18 for silicon nitride ribbons. The density of silicon nitride is about *3,000 kg/m*^{3} (the exact value depends on whether the film is stoichiometric or if it is silicon rich). Other possible ribbon materials like silicon, silicon dioxide and aluminum, have lower densities ranging from *2,200* to *2,700 kg/m*^{3}, but that has only a minor effect on the results. The ribbon thickness is chosen to be *100 nm*, which is reasonable for many practical applications and well within the capabilities of modern thin-film deposition processes.

The range of voltages in Fig. 10.18 is representative of voltages used in practical devices. Voltages below *10V* tend to make the devices too sensitive to shock and vibrations during fabrication, packaging, and installation, so only rarely is an electrostatic MEMS device designed for such low maximum voltage. At the other end, *200V* represents an upper limit for the voltage that can be applied to electrostatic actuators with micron sized gaps. From a systems point of view, much lower voltages are preferred, so practical devices tend to have maximum voltages of a few tens of volts.

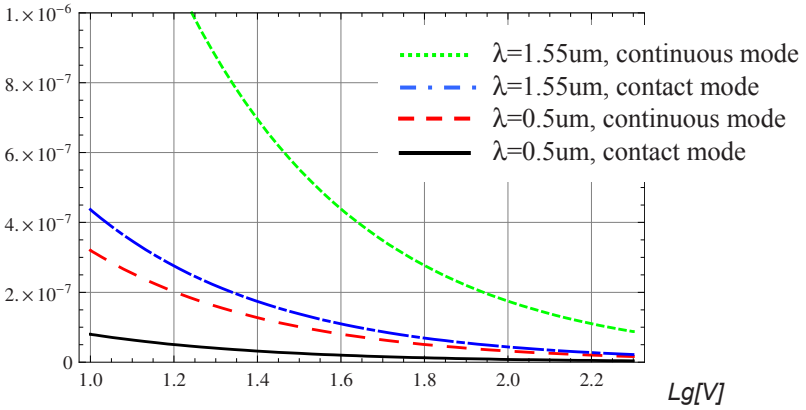


Figure 10.18 *Switching times in seconds as a function of maximum applied voltage (logarithmic scale) for grating light modulators driven by parallel-plate actuators. The ribbons are 100nm thick and made of silicon nitride with a density of 3,000 kg/m^3.*

The graphs of switching times demonstrate that even though the grating modulator is a mechanical structure, it performs at speeds that are unheard of for any other type of machinery! We see that using relatively modest actuation voltages, we can design grating modulators with switching times well below *1 us*. Even in the most difficult case, which is that of continuous operation at *1.55.um* wavelength, a voltage of less than 30V is enough to break the *1 us* barrier.

For the opposite extreme, which is that of contact mode operation at *500 nm* wavelength, the graph predicts that it is possible to get below *10 ns*. The *20 ns* switching times referenced above were indeed obtained in contact mode. These early results are still among the fastest MEMS ever reported, although it seems quite obvious that even faster devices can be designed and fabricated.

10.6.2 Ribbon Mechanics

The speed analysis that led to the graphs of Fig. 10.18 is based on the assumption that we can design the ribbons to have exactly the right spring constant that will give the correct maximum deflection of $\lambda/4$ for a given maximum applied voltage, while at the same time making the ribbons quite thin. The validity of this assumption depends on the available technology and on the exact structure we are trying to fabricate. For example, if the spring force is mainly due to bending of the ribbons, then it becomes difficult to make sufficiently stiff springs as we reduce the ribbon thickness, because the bending stiffness goes as the cube of the thickness.

It is therefore very convenient, even critical, to be able to use tensile stress to stiffen the ribbons. A complete and instructive description of the effect of stress

on beams (or ribbon) deflection can be found in [8]. Here we will consider the two limiting cases of bending beams and vibrating strings.

In a bending beam, the potential energy storage is dominated by stress in the material causes by bending. In a vibrating string, on the other hand, the potential energy is mostly stored as stress caused by elongation. The ribbons of a grating modulator will in principle store potential energy both as bending stress and elongation stress, but typically one or the other will be dominant.

To get simple, closed form expressions for the displacement along the length of a beam we will assume uniform loads. This is not exact, because the electrostatic forces will typically vary along the length of the ribbon, but it is a convenient approximation, and it gives results that are in good agreement with observations.

For bending beams, i.e. low-stress ribbons dominated by bending, the deflection curve takes the form [9]

$$y = \frac{q}{2E \cdot t^3} \cdot x^2 (L - x)^2 \tag{10.30}$$

where L is the ribbon length, t is the ribbon thickness, E is the Young's modulus of the ribbon material, and q is the applied spatially uniform and temporally harmonically varying force per unit of area (or applied pressure). Note that any symmetric load will give a deflection curve with the same boundary conditions (zero deflection and zero slope at the ends) and same symmetry (mirror symmetry about the midpoint of the ribbon), so the shape, if not the magnitude, given by the equation is a good approximation to the deflection resulting from any realistic loading condition.

If, due to the particular deposition conditions of the material, the ribbon has a large tensile stress, then the stored potential energy in the ribbon is mostly due to elongation, and its deflection can be expressed as [10,11,12,13]:

$$y = \frac{q}{2 \cdot t \cdot \sigma} \cdot x(L - x) \tag{10.31}$$

where t is the ribbon thickness, and σ is the tensile stress in the ribbon.

The deflection shapes given by two equations above, together with the shape of a vibrating string [14], are plotted in Fig. 10.19. As expected, the stress dominated ribbon behaves very much like a vibrating spring, and it has a much larger flat part in the center than the bending dominated beam. This is of great advantage in grating light modulators, because we would like all the light reflected from the ribbon to get the same phase delay. This "flattening" effect can be further enhanced by removing the electrostatic force from the center of the beam, i.e. arrange the substrate electrodes such that the electrostatic force is not applied at the center.

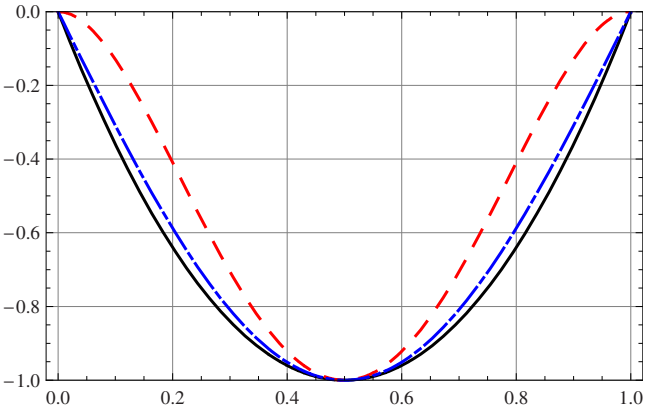


Figure 10.19 *Deflection curves for of bending-dominated ribbons (dashed), elongation-dominated (solid) ribbons. The well-know vibrating string is included for comparison (dot-dashed). The elongation-dominated, or stress-dominated ribbon, which deflects very much like a vibrating string, is a significantly flatter at the center than the bending dominated beam.*

The resonant frequency of the fundamental vibration mode of both bending-dominated and stress-dominated ribbons can conveniently be found by the Rayleigh method [[15]]. This method is based on the fact that a vibrating mechanical oscillator has a constant, or near constant, stored energy over the vibration cycle. The kinetic energy at the time of maximum vibration velocity, (i.e. no deflection) must therefore equal the maximum potential energy at the time of maximum deflection (i.e. no velocity). The Rayleigh method for determining the resonant frequency is very insensitive to errors in the actual deflection curve. To get useful results, all we need are reasonable approximations to the real deflection curves.

To apply the Rayleigh method, we need to calculate the stored kinetic and potential energy in the ribbons. The maximum ***kinetic energy*** of a ribbon element dx undergoing harmonic vibration is given by

$$E_{kin} = \frac{1}{2} \cdot \omega^2 y^2 \cdot \rho \cdot b \cdot t \cdot dx \tag{10.32}$$

where ω is the natural frequency of the harmonic oscillations, y is the harmonic vibration amplitude, and ρ, b, and t are the density, width, and thickness of the ribbons, respectively. The ***potential energy*** due to bending and elongation are given by [16]:

$$E_{bending} = \frac{EI}{2} \cdot \left(\frac{d^2 y}{dx^2} \right)^2 \cdot dx = \frac{E \cdot b \cdot t^3}{24} \cdot \left(\frac{d^2 y}{dx^2} \right)^2 \cdot dx \tag{10.33}$$

$$E_{stress} = \frac{\sigma \cdot A}{2} \cdot \left(\frac{dy}{dx}\right)^2 \cdot dx = \frac{\sigma \cdot b \cdot t}{2} \cdot \left(\frac{dy}{dx}\right)^2 \cdot dx \tag{10.34}$$

where E is the Young's modulus of the ribbon material, I is the ribbon moment of inertia, and σ is the tensile stress in the ribbons. Other energy terms like rotational kinetic energy and shear potential energy [17] are negligible in practical grating modulators.

Now we apply the Rayleigh method, i.e. we integrate the maximum kinetic energy (Eq. 5.33) over the ribbon, and set it equal to the integral of the maximum potential energy. First we carry out the calculation for the bending-dominated case. Of course we are using a symbolic-math software package to do the integrals.

$$\int_{ribbon} E_{kin} = \int_{ribbon} E_{bending} \Rightarrow \int_0^L \frac{1}{2} \cdot \omega^2 y^2 \cdot \rho bt \cdot dx = \int_0^L \frac{EI}{2} \cdot \left(\frac{d^2 y}{dx^2}\right)^2 \cdot dx \Rightarrow \tag{10.35}$$

$$\begin{aligned} \omega^2 &= \frac{\frac{EI}{2} \cdot \int_0^L \left(\frac{d^2 y}{dx^2}\right)^2 \cdot dx}{\frac{1}{2} \cdot \rho bt \cdot \int_0^L y^2 \cdot dx} = \frac{EI}{\rho bt} \frac{\int_0^L \left(\frac{d^2\left(x^2(L-x)^2\right)}{dx^2}\right)^2 \cdot dx}{\int_0^L x^4 (L-x)^4 \cdot dx} \\ &= \frac{E \cdot bt^3}{12 \cdot \rho bt} \frac{\frac{4L^5}{5}}{\frac{L^9}{630}} = \frac{42 \cdot E \cdot t^2}{\rho \cdot L^4} \end{aligned} \tag{10.36}$$

$$f_{bending} = \frac{t}{2\pi \cdot L^2} \sqrt{\frac{42 \cdot E}{\rho}} \tag{10.37}$$

Similarly, we find for the stress dominated case:

$$\int_{ribbon} E_{kin} = \int_{ribbon} E_{stress} \Rightarrow \int_0^L \frac{1}{2} \cdot \omega^2 y^2 \cdot \rho bt \cdot dx = \int_0^L \frac{\sigma \cdot b \cdot t}{2} \cdot \left(\frac{dy}{dx}\right)^2 \cdot dx \Rightarrow \tag{10.38}$$

$$\omega^2 = \frac{\frac{\sigma \cdot b \cdot t}{2} \int_0^L \left(\frac{dy}{dx}\right)^2 \cdot dx}{\frac{\rho bt}{2} \cdot \int_0^L y^2 \cdot dx} = \frac{\frac{\sigma}{2} \int_0^L \left(\frac{d\left(x(L-x)\right)}{dx}\right)^2 \cdot dx}{\frac{\rho}{2} \cdot \int_0^L \left(x(L-x)\right)^2 \cdot dx} = \frac{\frac{\sigma}{2} \cdot \frac{L^3}{3}}{\frac{\rho}{2} \cdot \frac{L^5}{30}} = \frac{10 \cdot \sigma}{\rho \cdot L^2} \tag{10.39}$$

$$f_{stress} = \frac{1}{2\pi \cdot L}\sqrt{\frac{10 \cdot \sigma}{\rho}} \tag{10.40}$$

This formula is virtually identical to the expression for the resonance frequency of the fundamental mode of a simple vibrating string. (The vibrating string has a deflection given by $y = y_{\max} \cdot \sin(\pi \cdot x/L)$, and the resonant frequency is $\frac{1}{2\pi \cdot 2L}\sqrt{\sigma/\rho}$, which differs from our expression by the factor $\sqrt{10}/\pi \approx 1.007$ [18]).

Comparison of the above equations shows that the resonance frequencies of both the bending-dominated and stress-dominated ribbons are dependent on ribbon geometry and material constants. The main difference is that the bending beam is characterized by its Young's modulus, while the stress-dominated ribbons depend on material stress. The materials that have the durability, reliability, and IC compatibility required for grating modulator fabrication have only a very limited range of elastic-modulus values. Material stress on the other hand is a strong function of fabrication parameters like deposition temperature, pressure and chemical composition and can be controlled with precision over several orders of magnitude. For example, the stress in silicon nitride films can be accurately controlled from the MPa level to several hundreds of MPa [19].

An additional advantage of the stress dominated ribbon is that its resonance frequency is independent of ribbon thickness, while the bending-dominated beam has a diminishing resonance frequency as the thickness is reduced. This means that stress-dominated ribbons can be made much thinner and therefore faster than bending beams. For example, if we design a *20 um* long and *100 nm* thick ribbon, it will be *100* times faster if it has a stress comparable in value to its Young's modulus.

The benefit of using stress as a design parameter in ribbon design is therefore twofold. First we get the added flexibility of being able to chose exactly the stress we ant to achieve the desired ribbon stiffness. Secondly, we can make the ribbons as thin as our technology allows without compromising switching speed. Most applications that critically depend on fast switching will therefore incorporate tensile stress in the ribbons. The conclusion is that it is indeed possible to achieve switching speeds that are approaching the fundamental limit set by mass and spring constant if we are willing and able to control the stress in the ribbons.

10.6.3 Linear Display Architecture

The switching speeds demonstrated in Fig. 10.18 are impressive, particularly considering the fact that these are mechanical devices. The question remains, how-

ever, whether this speed can be turned into useful practical advantages. For fiber optic applications we can make the argument that faster is always better. The switching speeds of grating modulators cannot compare to those of electro-optic devices, but switching times in the 100 ns range are still very useful for network reconfiguration in response to changing traffic patterns and to link failures.

For displays it is not so obvious that the capability to switch in the tens or hundreds of nanoseconds is useful in practical systems. Video refresh rates even for the highest quality displays are less than one hundred Hertz, so what use are speeds that several orders of magnitude faster? The answer is that the ability to switch at high speed allows flexibility in the addressing of the individual pixels in the displays, and, more importantly, it allows more efficient display designs, like the linear-modulator-array, or 1-D, architecture [20] shown in Fig. 10.20.

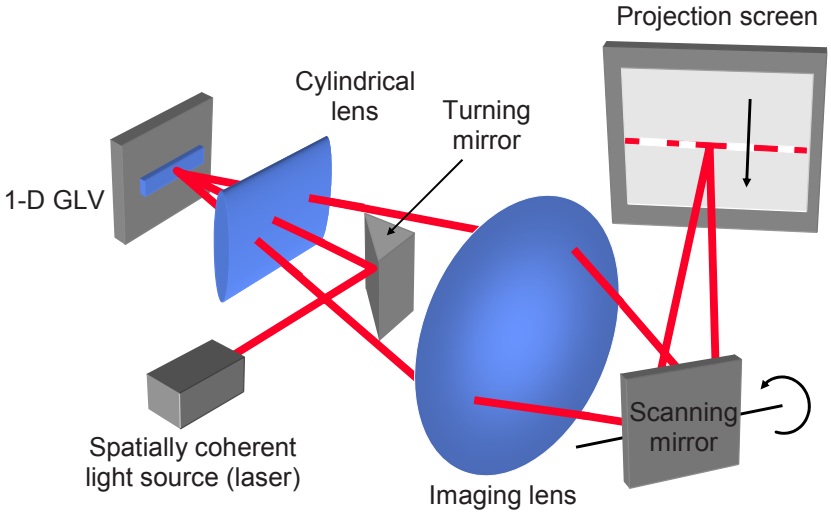


Figure 10.20. Schlieren-type display system based on a linear array of GLMs. Shown schematically is the path of light for one grating modulator in the display. The spatially coherent beam from the illuminating laser reflects off the turning mirror and is focused by the cylindrical lens to a stripe on the linear grating modulator array. The diffracted light from each grating modulator is imaged onto a screen through the scanning mirror.

The optical system using a linear array is similar to the system we studied earlier in that it is a Schlieren-type display with a turning mirror that serves the dual purposes of directing the illumination to the modulator array and blocking reflected (but not diffracted) light from each modulator from reaching the projection screen. Where the 2-D system can use any light source, the 1-D system requires a source with substantial spatial coherence, so that the illuminating light can be effectively

focused to a narrow stripe on the modulator array. In practice this means that the light source must be a laser. The reflected light from the modulators is blocked by the turning mirror and sent back towards the source. (Care is taken not to couple the reflected light into the source, because that type of feedback into a laser will cause instability and optical power fluctuations.) The diffracted light, on the other hand, misses the turning mirror and is imaged onto the projection screen through a scanning mirror. The function of the scanning mirror is to translate the modulated stripe of light from the grating-modulator array across the screen to form a 2-D image. Each modulator in the array therefore does not correspond to a single pixel on the screen, but rather a column of pixels.

The operation of the linear-array display require that each modulator can be reconfigured not at the image refresh rate, but at a mush higher rate that equals the refresh rate multiplied by the number of pixels in a column. Typical high-quality displays might have a refresh rate of 100 Hz and 1,000 pixels in each column, so the modulator refresh rate is 100 kHz. The switching time must then be substantially shorter than 10 us, so that the switching can take place in a small part of the grating refresh rate.

It is clear from the graphs in Fig. 10.18 that grating modulators can be designed to have more than sufficient speed for this type of operation. This is true both for the contact mode and the continuous mode. MEMS displays based on arrays of rotating micromirrors, e.g. TI's DLP technology, require much larger displacement to achieve high contrast, and therefore do not have sufficient speed for linear-array operation.

Another consequence of the one-modulator-per-column-of-pixels design shown in Fig. 10.20 is that each modulator must be able to handle the optical power of one whole column. In practice, this does not present problems for image projection. Regular Aluminum-coated mirrors that absorb about 10 % of the incident light can handle the power levels required for even the most bright projection displays.

Given that GLMs have the required speed and power handling, the swept-linear-array system has numerous advantages. Most obviously it requires much fewer modulators than two-dimensional displays. For a 1,000 by 1,000 pixel display, the linear array only has one thousand elements, while the 2-D array has one million. Fewer elements translate into simpler fabrication and better yield.

More importantly, the geometry of the linear array allows drive and multiplexing circuitry to be placed next to, rather than underneath, the modulators. This allows more flexibility in integration of the modulators and their driving and multiplexing circuitry. For example, the circuits can be made on the same substrate before OR after the modulator fabrication is completed, OR the circuitry can be made on separate chips and flip-chip bonded to the modulator substrate. With 2-D arrays it is practically impossible to avoid placing multiplexing transistors underneath the

array[c]. This means that the circuits must be made first and be subject to the thermal loading of going through the modulator fabrication process, which again mean that the transistor technology must be developed especially for integration with grating modulators.

The bottom line is that for 2-D arrays the transistor technology must be custom developed and must be re-optimized each time the modulator fabrication process is changed. 1-D arrays, on the other hand, can be integrated with standard, state-of-the-art circuitry, which means that a number of providers can be found. The development cost of modulator arrays AND electronics is therefore much less for 1-D than for 2-D arrays.

From a MEMS design point of view, the 1-D array is simpler because the fill factor does not depend on the termination of the ribbons. We can therefore use any convenient structure that suspends the ribbons with correct spring constant without having to worry about how light is reflected from the terminations. To achieve high fill factor in 2-D arrays, the suspending spring must be created on a separate layer underneath the reflecting surfaces of the ribbons.

The final advantage of the linear array is that it leaves more room for the driving circuitry so that it is more practical to operate the grating modulators in the continuous, or non-contact, mode. Typical display applications need 12 bits of grey scale (12 bits of resolution of the optical power), so holding a grey-scale value requires a larger number of transistors. 2-D arrays typically are operated in a binary fashion, with grey scale created though temporal multiplexing. This type of operation fits well with contact-mode modulators and requires fewer transistors in the multiplexing circuitry of each pixel. Contact mode is, however, more difficult to sustain repeatedly and without wear over a large number of switching cycles, so non-contact mode is preferable from a reliability point of view.

10.6.4 1-D Modulator Array Fabrication

A grating modulator design suitable for 1-D arrays is shown in Fig. 10.21. It has three periods of fixed and movable ribbon pairs. The movable ribbons can be pulled towards the substrate by electrostatic attraction. In the relaxed state with no deflection of the movable ribbons, the grating is in its reflective state, which is the dark state of the imaging system of Fig. 10.20. When the ribbons are at their maximum deflection of $\lambda/4$, the grating is in its diffractive, or bright, state. Me-

[c] The instability of the parallel-plate electrostatic actuator creates a deflection vs. voltage curve with significant hysteresis. It is possible to use this bi-stability as a set-and-hold mechanism so that any modulator in a 2-D array can be switched, by addressing the row and column to which it belongs, without changing the states of the other modulators in the row or column. This type of mechanical memory is, however, difficult to implement in large arrays, so this approach has not yet been adopted in commercial products.

chanically the fixed and movable ribbons are identical. The only difference between them is that the fixed ribbons are electrically shorted to the substrate so that there are no electrostatic forces pulling them downwards.

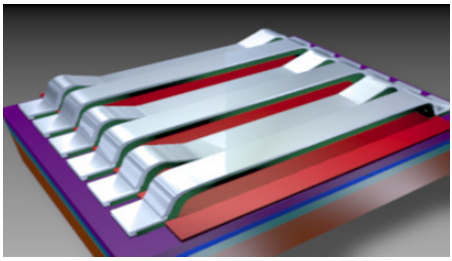


Figure 10.21 Schematic drawing of a single pixel in Silicon Light Machine's high-contrast grating light modulator. The grating has three pairs of silicon nitride ribbons with an overlayer of aluminum to enhance reflectivity. Every other ribbon can be actuated by electrostatic attraction towards the substrate to switch from the reflective (relaxed) state to the diffractive (actuated) state. (Courtesy Silicon Light Machines)

The tools and processes needed to fabricate the grating light modulator of Fig. 10.21 are relatively simple by IC standards. There are many variations on the process flow, depending on the available technology, but a typical fabrication sequence goes something like this:

1. Thermal oxidation (SiO_2) of standard <100> wafers to a thickness of *1μm*. This step creates an electrically insulating layer. A process with a lower thermal budget can be substituted if necessary.
2. Deposition of *0.5 um* of polycrystalline silicon. This film will become the lower electrode. Metals cannot be used because of the subsequent high-temperature fabrication steps. The conductivity of this layer is not critical, so poly-Si films with about *10 Ω/sq* are sufficient. Attention has to be paid to the surface roughness of this layer, however, because the topography of this film will translate into films that are deposited later.
3. *Mask 1* is used to define the lower electrodes and the wiring in the Poly-Si film through photolithography and dry etching.
4. The Poly-Si layer is oxidized protect it from the subsequent sacrificial etch. A thin silicon dioxide layer of about *30 nm* is sufficient for this purpose.
5. A layer of amorphous silicon is deposited. This layer will be removed before the process is complete, and its purpose is to define the spacing between the substrate electrode and the ribbons. Its thickness is therefore an important design parameter. For contact mode operation in the visible it should be about *600 nm* to roughly match the longest wavelength to be modulated. The cross section of a ribbon after the completion of this process step is shown in Fig. 10.22a.

6. *Mask 2* is used to define openings in the a-Si layer, through which the silicon nitride ribbons will be attached to the poly-Si wiring layer.
7. A layer of silicon nitride is deposited with low-pressure Chemical Vapor Deposition (LP-CVD) at relatively elevated temperatures *(~ 600 ºC*). The thickness of this layer is not critical from an operational point of view, so the thickness and the stress of the nitride film will be chosen to provide the right spring constant for the ribbons, while staying within the bounds of the available technology. It is important to control the stress in the film, so the nitride is not necessarily stoichiometric (i.e. the ratio of Si to N in the film is not exactly as given by the formula Si_3N_4). A typical thickness is about *100 nm*. Figure 10.22b shows a cross section of a ribbon at this stage.
8. *Mask 3* is used to define via holes in the nitride film to allow the aluminum overlayer on the ribbons to contact the poly-Si wiring.
9. Aluminum is deposited and patterned using *Mask 4* and *5*. A thin layer of about *50 nm* is sufficient to give the ribbons close to the reflectivity of bulk aluminum, while thicker layers are needed to form bond pads so two separate deposition and masking steps will typically be required
10. The ribbons and their supporting springs are defined in the Silicon Nitride film by lithography with *Mask 6* followed by dry etching of the Nitride.
11. The a-Si sacrificial layer is removed in a gas-phase Xenon-difluoride (XeF_2) etch. This etch is unique in that it etches silicon (and Poly-Si and A-Si with high preferentiality over almost any other material used in IC technology. In particular, it is difficult to find another sacrificial layer-etch combination that does not damage thin Aluminum films. After this step, wafer processing is finished and the ribbon cross section is as shown in Figure 10.22c.

Figure 10.22 shows schematically the cross section of a single ribbon at various stages of the process sequence, while Fig. 10.23 shows Scanning Electron Micrographs of part of a linear or 1-D modulator array. The process as described here uses only 6 lithography steps. In practice we might choose to use a couple of extra masks for purposes of integration and packaging, but still the process is by all standards very simple.

The ribbon material of the modulators shown in Figs. 10.22 and 10.23 is Silicon Nitride (Si_3N_4), which is a very stable and robust insulator. Typically, all grating modulators are operating at a small fraction of the tensile yield stress, even it their most deflected state. This is possible because the operational principle of the grating modulator requires such small maximum deflections. In combination with non-contact operation, this leads to negligible material fatigue and good reliability over the expected life time of the modulators.

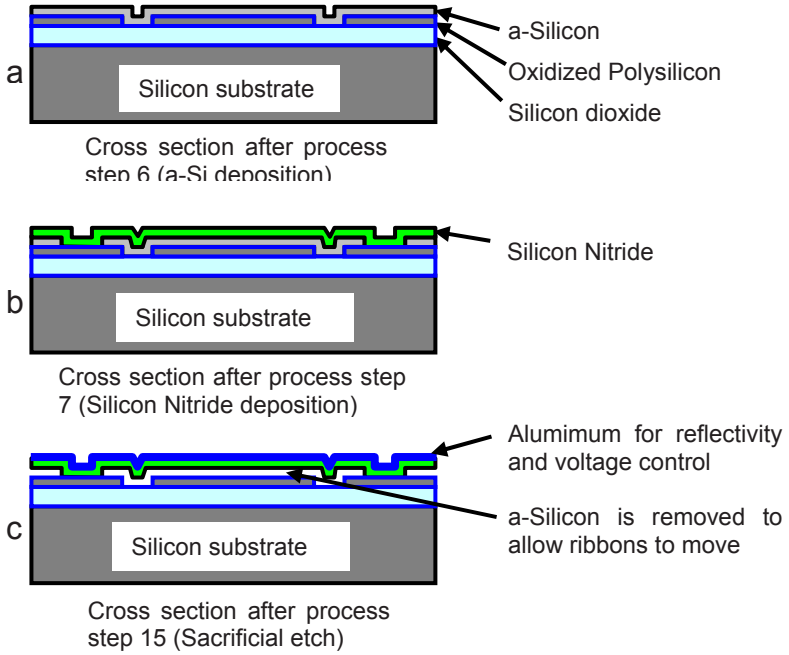


Figure 10.22 *Schematic cross section of a single ribbon (the movable and fixed ribbons have identical cross sections) in a high-contrast grating light modulator at three different stages of the fabrication process (not to scale). At the completion of the process, the silicon nitride ribbons are suspended by springs over the substrate electrode, so that they will move towards the substrate when actuated by electrostatic attraction.*

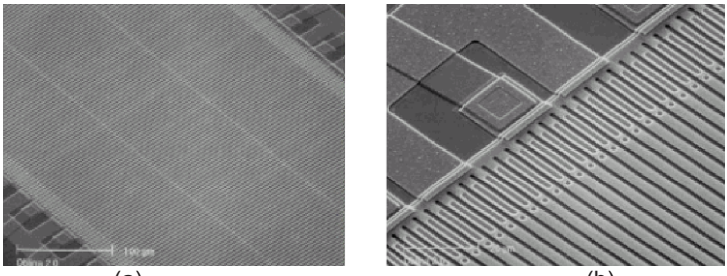


Figure 10.23 *Scanning Electron Micrograph of a part of a Silicon Light Machines grating light modulator array. Figure a) shows about 15 individual modulators, each with 3 ribbon pairs, of an array of 1080, and figure b) shows a close up of the ribbon suspensions. Each modulator is made up of three pairs of one movable and one fixed ribbon. The array is addressed from both sides, so the period of the addressing lines in b) corresponds to two modulators or 12 ribbons. (Courtesy Silicon Light Machines)*

10.6.5 Light Sources for Swept-Line Projection Displays

The tensile stressed ribbons meet the speed and power handling requirements and work very well in the swept-line projection display. The overall system is also quite power efficient. The fill factor in a 1-D grating modulator array is determined by the spacing between the movable and fixed ribbons. With ribbon widths on the order of a few micron and using modern photo lithography, this fill factor is on the order of 95% or better. Combine that with an Aluminum reflectivity of 91% at visible wavelengths, and a diffraction efficiency into the two first orders of 81%, and the overall maximum throughput of the modulator is about 70%.

Of the 30% that is lost, about 20% is scattered into higher diffraction orders. This scattered light must not be permitted to reach the screen, because that will reduce the contrast. The other 10% is absorbed in the metal and the substrate of the modulator chip and is only problematic if it results in overheating of the ribbons.

The swept-line architecture also has the advantage of projecting an image with no pixilation. The linear grating-modulator array is pixilated, but only in terms of their electronic addressing. Mechanically there is no difference going between ribbons of the same or neighboring modulators. The pixelation of the modulators is therefore not discernible in the projected line on the screen. Likewise there are no discernible pixel boundaries in the sweep direction on the screen. The result is a very clean projected image without the disturbing higher spatial frequencies common to pixel-based projection systems.

Overall, the swept-line projection display with linear grating arrays works very well. It exhibits high contrast, good efficiency, simple and inexpensive chip technology, and it is robust and reliable. It has been demonstrated in form factors ranging from handheld devices to the enormous 10 meters by 50 meters display exhibited by Sony Corporation at the 2005 World Exposition in Aichi, Japan.

Comparing the swept-line display to the traditional Schliern projector with a 2-D modulator array, we see that there are two major differences between the two optical systems. The most obvious is the scanning mirror that enables the linear modulator array to create a 2-D image. The scanner does not represent a significant increase in complexity. High-quality scanners of all sizes, ranging from MEMS scanners for the smaller systems to galvanometric scanners for the larger displays, are inexpensive and readily available.

The second difference is less obvious, but more difficult to deal with. The scanned-line projector requires a spatially coherent light source, i.e. a laser, to allow most of the light to be focused onto a single, narrow line at the modulator array. Traditional light sources are spatially incoherent and their output light can therefore not be focused to a narrow line.

This difference in focusing between traditional and spatially coherent sources is illustrated in Fig. 10.24. The light bulb, like any traditional light source of finite area, can be considered as a collection of individual light emitters with no fixed phase relationship between their emitted optical fields. That means that each individual source creates an illuminated spot of an area equal to the point spread function of the lens system. The total illuminated area is then the sum of the illuminated spots from each of the individual sources. For typical light emitters, particularly the high-power sources required for projection displays, the illumination area is many times larger than the point spread function.

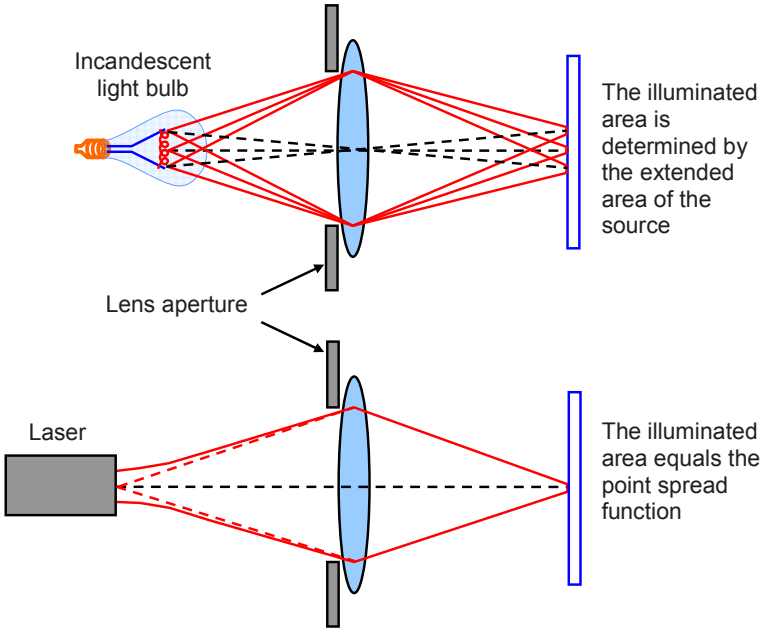


Figure 10.24. Comparison of illumination with traditional and spatially-coherent light sources. The light bulb acts as a large number of independent light sources, so the illuminated area is determined by the spatial extent of the source. The spatially coherent laser beam appears to originate in a point, so it illuminates an area equal to the point spread function of the lens system.

Spatially coherent lasers behave very differently. The optical fields from different parts of their output apertures have a well-defined phase relationship (typically they are all in phase), so that the light appears to be, or can be made to appear to be, originating from a single point. The illuminated area in the image plane is then equal to the point spread function of the lens system.

The consequence of the short length of the GLM ribbons and the large ratio of length (perpendicularly to the ribbons) to width (along the ribbons) of the 1-D

grating modulator is that practical swept-line displays need lasers for illumination. This creates both challenges and opportunities. The first consideration is availability. A fully functioning display requires three colors, red, green, and blue, that are individually modulated and projected. Fortuitously, lasers and optical parametric oscillators are now available at all wavelengths in the visible spectrum. There are, however, big differences in cost and complexity of lasers of different colors. Semiconductor lasers emitting in the red are powerful, efficient, reliable, and inexpensive. They are also available as laser arrays that eliminate the appearance of speckle in swept-line displays (see below). Semiconductor lasers are also commercially available in the green and blue, but they are less mature, so their specifications and diversity of designs cannot match those of red laser diodes. The technology is advancing rapidly, however, and new products with improved characteristics are continuously being developed. It is a safe bet that high-power, low-cost blue and green laser-diode arrays will become available in the near future. When that happens Grating Light Modulator systems for a wide range of applications will become competitive in the market place. These systems will befit from the very high brightness, as well as the efficient and reliable operation and very long lifetimes of semiconductor lasers.

Beyond availability and price, another issue that we must consider when using laser for projection is speckle. Speckle is the spatial pattern a laser spots makes when reflected off a scattering surface. To observe it, simply shine a laser pointer on a piece of paper and note that the illuminated spot is not uniform, but has a well-developed amplitude modulation. If the laser is single-mode, i.e. its output is a single optical beam with a well-defined phase front, then the contrast in the speckle pattern is unity.

The speckle is, however, not inherent to the laser beam itself. If we were to take a picture of the laser beam directly (never look directly into a laser, not even a low-power laser pointer!), the picture would not show a speckle pattern, but rather a smoothly varying intensity distribution as we would expect from a Gaussian beam. The speckle we see on the projection screen is a consequence of the irregularity of the scattering surface, as illustrated in Fig. 10.25. If we consider an area of the projection-screen corresponding to the minimum resolvable spot for the viewer, then the perceived brightness of the area depends on the direction of the viewer and on the details of the scattering surface. If the light from different parts of the area interferes destructively, then the area appears dark, while if the light interferes constructively, it appears bright. The observed intensity pattern therefore has the same randomness as the scattering surface.

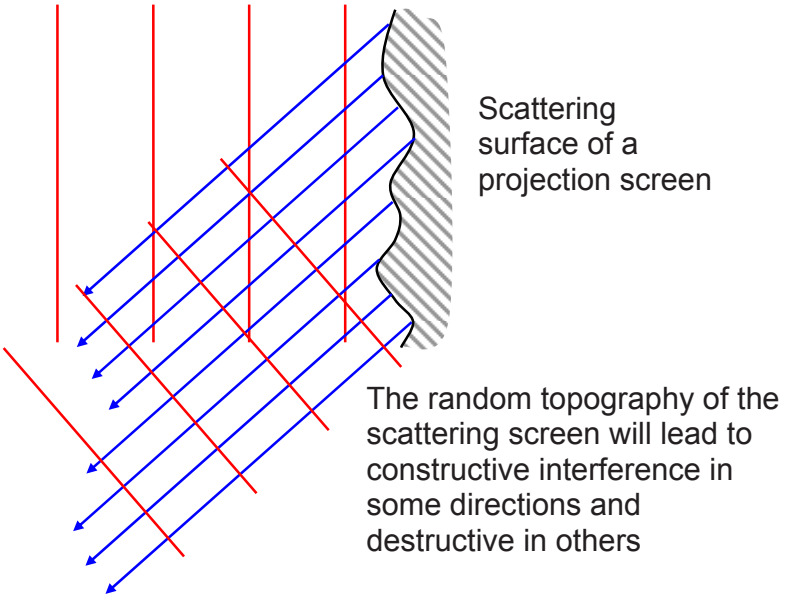


Figure 10.25. Laser speckle is caused by the irregularities of the surface that the laser illuminates. The height differences of the scattering surface create phase differences so that light from different parts of the surface interfere differently in different directions. The result is a speckle pattern with 100 % contrast if the laser has a single spatial mode.

Laser speckle is very noticeable and degrades the image quality of laser projections. Fortunately, it is relatively straightforward to remove it or decrease it to negligible levels. The trick is to mimic a traditional light source, which, as described above, really should be thought of as several independent light sources. By overlaying several speckle patterns they average so that their contrast is reduced.

A convenient way to this is to use a laser array as shown in Fig. 10.26. The lasers in the arrays do not have a fixed phase relationship, so they are indeed independent light sources. The coherence of the individual lasers allows the output of the array to be focused to a tight line on the GLM pixels, but each pixel along the line is illuminated by several neighboring lasers. Once the light is projected on a screen, the each pixel will show light coming from several lasers from slightly different directions. The speckle patterns created by each laser are independent and will average out to an almost uniform illumination.

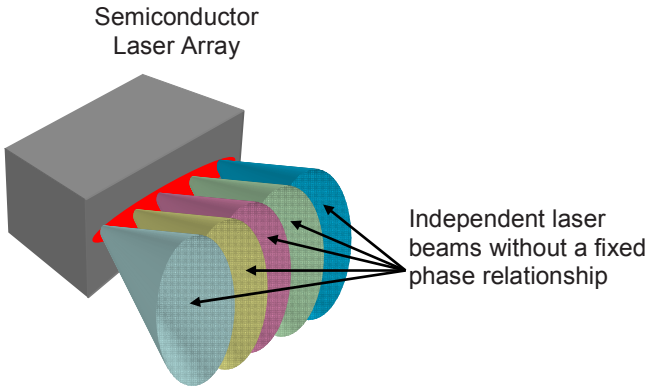


Figure 10.26. A semiconductor laser array used for swept-line projection displays consist of laser that operate independently[d], so that the array is spatially coherent on the short axis and in-coherent on the long axis. The outputs from the different laser are shown in different shades for clarity. The array output can be focused to a narrow line in which each point is illuminated by several overlapping laser beams coming from slightly different directions. The effect is to average out the speckle pattern from each laser and produce a uniform projection.

10.7 Summary of Grating Light Modulators

Most optical detectors are square-law devices, i.e. they are sensitive to optical power or intensity, but not directly to optical phase. That does not mean that it is useless to modulate the phase of optical signals. On the contrary, phase modulation is often more efficient than direct amplitude modulation. In Chapter 9 we demonstrated that phase singularities result in images with intensity minima that are smaller and have better contrast than those produced by any amplitude modulation. In this chapter we show that phase modulated gratings can be switched between reflecting and diffracting states, and that this switching can be converted to amplitude modulation by simple optical systems. The main conclusion of the phenomenological description in the first part of this chapter is that phase modulation followed by PM to AM conversion in diffraction gratings is straightforward to implement using MEMS technology, and that this approach leads to Grating Light Modulators of simple mechanical and optical designs.

The phasor representation of optical fields, that was first introduced in Chapter 2, is used here to model the optical characteristics of GLMs. It gives us an accurate

[d] There are also phased arrays of semiconductor lasers, and those are very useful for many applications, but in project displays we prefer arrays of independent lasers.

dependence of reflected and diffracted light on grating amplitude, and allows us to establish the analog and digital modulation properties of the GLM. The phasor representation is also a very useful tool for investigation of dispersive properties, and it leads us to the high-contrast GLM design.

The basic phasor description cannot answer questions about GLM scaling, however. To understand how far we can miniaturize the GLMs, we model it as a periodic array of phase delays and adapt the standard Fraunhofer diffraction formula to this structure. The resulting equations tell us the required size and number of grating elements in practical GLMs, and guide the design of MEMS implementations.

The last part of the Chapter is devoted to a detailed description of GLM projection displays. This example illustrates how well phase-modulating MEMS arrays are suited to this application. The small required displacement of only $\lambda/4$, makes the projected arrays compact, reliable, and inexpensive. The small inertia and efficient actuation also gives the GLMs high speed, which enable the swept-line architecture and further reduces complexity and cost.

Exercises

Problem 10.1 - Phasor Representation

Use the phasor representation to describe the following optical components (see Chapter 6 for their description):

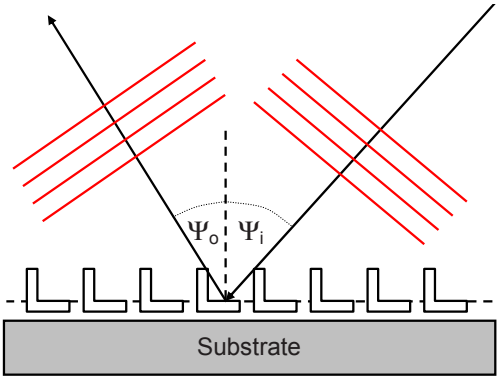
a. Mach-Zender Interferometer
b. 3-dB coupler
c. Fabry-Perot Interferometer
d. Ring filter
e. Diffraction grating
f. What are the advantages, if any, of the phasor representation in each case? If you find the phasor representation contrived and therefore useless in any of these cases, explain how.

Problem 10.2 - Holographic Display

a. How can the high-contrast GLM be modified to give both amplitude and phase modulation?
b. Show how this in principle can be used to create a holographic display.
c. Comment on the practicality of such a holographic display. What are the potential commercial uses of holographic displays?

Problem 10.3 - Corner Cube Grating

Consider the corner-cube grating shown in the figure. Each element of the grating is a corner cube in two dimensions (i.e. there is no variation in the dimension perpendicular to the plane). Note that each corner cube retro reflects the light that is incident on it.



Corner Cube Grating. Each corner retro reflects the light, so that a diffraction pattern is formed by the interference of the retro reflections.

a. Follow the procedure of Chapter 10.5 and write an equation for the retro reflected diffraction pattern. Make any reasonable simplifying assumption, but justify their use.
b. What are the advantages and disadvantages of this structure compared to traditional gratings?
c. What functions could you implement if each corner cube can be moved vertically or horizontally by MEMS actuators?
d. Describe how you could implement the corner-cube grating using IC and MEMS fabrication techniques.

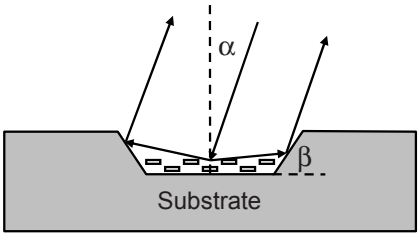
Problem 10.4 - GLM Corner Cube

a. How can you combine a GLM with a corner cube (i.e. a single element of the corner-cube grating of Problem 10.3) to create a retro reflecting modulator?
b. How is the efficiency of the modulator affected by the incident angle? Analyze the modulator in 2-D to simplify the expressions.

Problem 10.5 - Diffractive Corner

Consider the 2-D diffractive corner in the figure. The purpose of this structure is to retro reflect the light in the first order diffraction modes of the grating. Assume

that the truncated pyramidical hole is anisotropically etched in <100> silicon, so that the angle β is given by $\tan^{-1}\left(\sqrt{2}\right) \approx 54.7°$.

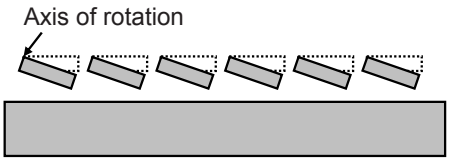


Diffractive Retro Reflector. The first order diffracted modes of the grating are reflected back in the direction of the incident light.

a. Show that in 2-D and under the small-angle approximation, all light that hits the diffractive retro reflector and that is diffracted into the first order diffraction modes of the grating, is sent back along the incident path.
b. How can the principle of the diffractive retro reflector be extended to 3-D?
c. What are the advantages and disadvantages of the diffractive retro reflector compared to the corner cube?

Problem 10.6 - Tilting GLM

The GLM below have tilting, rather than piston-motion, micromirrors.



Grating Light Modulator with tilting mirrors.

a. Draw a schematic of an optical system that uses the tilting-mirror GLM.
b. What is the minimum distance that the edges of the mirror have to move?
c. Compare and contrast the tilting-mirror GLM to the standard GLM. In what types of application would you prefer one over the other?

References

1 O. Solgaard, F. S. A. Sandejas, D. M. Bloom, "A deformable grating optical modulator", Optics Letters, vol. 17, no. 9, pp. 688-690, 1 May 1992.

2 R.B. Apte, F.S.A. Sandejas, W.C. Banyai, D.M. Bloom, "Deformable Grating Light Valves for High Resolution Displays," Society for Information Display'93 Symposium, Seattle, Washington, May 17th, 1993.
3 F.S.A. Sandejas, R.B. Apte, W.C. Banyai, D.M. Bloom, "Surface Microfabrication of Deformable Grating Light Valves for High Resolution Displays," Proceedings of the Seventh International Conference on Solid-Stare Sensors and Actuators (Transducers '93), 1993, pp. 6-8.
4 E. Labin, S. M. P. T. E. Journal, April 1950.
5 Earl I. Sponable, S.M.P.T.E. Journal, April 1953.
6 E. Baumann, S. M. P. T. E. Journal, April 1953.
7 M. Born, E. Wolf, "Principles of Optics", 6th edition, Pergamon Press, Oxford, 1986, pp.401-407.
8 Stephen D. Senturia, Microsystems Design, Kluwer Academic Publishers, 2001, Chapter 9.6.3, pp. 231-235.
9 Warren C. Young, "Roark's Formulas for Stress & Strain", McGraw-Hill, Sixth edition, 1989, pp. 102-104.
10 Warren C. Young, "Roark's Formulas for Stress & Strain", McGraw-Hill, Sixth edition, 1989, pp. 162-166.
11 Stephen D. Senturia, Microsystems Design, Kluwer Academic Publishers, 2001, Chapter 9.6.3, pp. 231-235.
12 O. Solgaard, "Integrated Semiconductor Light Modulators for Fiber-Optic and Display Applications", Stanford University Ph.D. Thesis, 1992, pp. 152-162.
13 S. P. Timoshenko, "Strength of materials, Part II, Advanced Theory and Problems", Second Edition, third printing, D. Van Nostrand Company Inc., New York, July 1943.
14 E. Kreyzig, "Advanced Engineering Mathematics", Third edition, pp. 417-425, Wiley International Edition, John Wiley and Sons Inc., New York, 1972
15 S. Timoshenko, "Vibration problems in engineering", Second printing, D. Van Nostrand Company Inc., New York, 1928.
16 R. T. Howe, "Resonant Microsensors", Transducers '87, Rec. of the 4th Int. Conf. on Solid-State Sensors and Actuators, pp. 843-848, 1987.
17 W. C. Albert, "Vibrating Quartz Crystal Beam Accelerometer", Proceedings, 28th ISA International Instrumentation Symposium, 1982, pp.33-44.
18 E. Kreyzig, "Advanced Engineering Mathematics", Third edition, pp. 417-425, Wiley International Edition, John Wiley and Sons Inc., New York, 1972
19 P.A. Beck, S.M. Taylor, J.P. McVittie, S.A. Ahn, "Low Stress nitride and polysilicon films for micromachining applications", Mat. Res. Symp. Proc. , vol 182, 1990, pp. 207-212.
20 S.R. Kubota, "The Grating Light Valve Projector", Optics & Photonics News, vol. 13, no. 9, September 2002, pp. 50-53.

11: Grating Light Modulators for Fiber Optics

11.1 Fiber Optic Modulators

Chapter 10 covers the basics of Grating Light Modulators (GLMs) and presents a set of tools and equations for their design and modeling. GLM system integration is illustrated through the example of a projections system with Schliern optics. In this chapter we extend the treatment to GLMs for fiber optics.

Grating modulators are among the fastest micromechanical systems with demonstrated response times as low as 20 ns [1]. Still, they are not nearly fast enough to be used to encode data on fiber optic communication channels where data rates now exceed 40 Gb/s. The roles for GLMs in fiber optics are therefore as variable attenuators, variable spectral filters, channel equalizers etc. In these applications, the GLM changes the signal strength of a one or more fiber channels at a rate that is low compared to the signal frequency. The signal itself is typically digital on-off keying or some more complex data format, but the GLM operates as an analog modulator, i.e. it controls the signal strength in a continuous fashion.

Optical modulators for fiber optics are quite different from the projection-display modulators described in Chapter 10. Ideally, all optical modulators should have perfect contrast[a], and no wavelength dependence or polarization dependence. In addition to these first order operational characteristics, we would also like our modulators to have high power handling, linear operation (i.e. the transmission should be a linear function of the control signal, so that analog signals can be reproduced with perfect fidelity), low power consumption, insensitivity to environmental influences including temperature variations, long-term stability, small size, and low cost. No real modulator will have all these properties. The best we can do is to prioritize the characteristics that are the most important for the target application.

---

[a] Perfect contrast means that in the on state, the modulator transmits all the light without attenuation, while in the off state, the modulator transmits no light. Perfect contrast therefore also implies zero insertion loss, but loss is not nearly as important as leakage in the off-state.

Fiber optic data modulators do typically not need as high contrast as display modulators. Most systems use digital modulation formats, and small differences of the off-state are no more important than similar differences of the on state. On the other hand, modulators that are designed for analog operations like variable attenuation, channel equalization, and traffic grooming, i.e. exactly the types of applications where optical MEMS plays a significant role, do require high contrast. The extreme example is modulators that are used as a part of a switch to block a signal so that another can be introduced. In such systems, the cross talk (ratio of unwanted to wanted signal power) should be better than -40 dB [2].

Dispersion is also very important in fiber-optic systems, because spectral phase distortion leads to temporal pulse distortion that compromise signal fidelity. The effects of linear dispersion can in principle be reversed or mitigated through predistortion. When combined with optical non-linearities of the fiber, however, distortion will lead to non-reversible signal degradation. Dispersion must therefore be tightly controlled, but the total dispersion is dominated by dispersion of the fibers, so dispersion in discreet modulators typically can be ignored.

Lastly, polarization dependence is much more of a challenge in fiber optics than in displays. In projection displays we can control the polarization of the illumination, so that polarization dependence can be ignored. Controlling input polarization is impractical in fiber optics. As we learned in Chapter 5, standard, single-mode fiber is not really single mode at all, but rather two-mode due to the presence of two orthogonal polarization modes. On perfect fibers, these two modes are degenerate (i.e. they have the same effective index of propagation constant) and uncoupled. The irregularities of real fiber will, however, lift the degeneracy and introduce time-varying coupling between the modes. Consequently, real fiber exhibit polarization-mode dispersion (pulse spreading due to differences of propagation constant for the two polarizations) and time-varying output polarization.

Polarization-mode dispersion is only important in very high capacity systems, but the time varying output polarization means that the state of polarization is unknown in practical fiber systems. Some specialized systems use polarization maintaining fiber or truly single-mode fiber to avoid polarization variations, but the standard solution is to use polarization independent components. Polarization dependence can be achieved by separating the polarizations and treating them independently before recombining, but the simpler, more practical, and less expensive solution is to construct the optical components to be inherently polarization independent.

The focus of this short chapter is the design of GLMs to meet fiber optic system requirements. In section 11.2 we use the phasor representation to model the behavior and optimize the dispersion characteristics of the three-level GLM. Then, in section 11.3, we consider geometries that minimize the polarization dependence. We show that the linear geometry of Chapter 10 is inferior to two-dimensional gratings with square and hexagonal unit cells.

11.2 Low Dispersion Grating Light Modulators

The high-contrast grating light modulator described in Chapter 10 is ideal for displays and a number of other applications that require very high contrast, because it produces a near perfect dark state. We achieve this in a Schlieren projection system, with a high-contrast grating light modulator that has a reflective state with very low dispersion. The dark state is therefore dark over a broad band of optical wavelengths. Bright pixels are created by setting the modulator to its diffractive state. This state has some dispersion that leads to coloring of the bright pixels, but that can be ignored in most cases, or compensated for if necessary.

On the screen of the Schlieren display the bright pixels are created by combining two (or more) diffracted orders. This combination means that the fields of the two diffracted orders interfere. The resulting interference pattern average over the pixel and is therefore on no consequence to the viewer. (This should not be confused with the speckle pattern that we must take care to remove from any projection display with laser illumination.) In display applications we can therefore think of the combination of the diffracted orders as a simple addition of the optical powers in each order.

If we try to use a similar Schlieren system in a fiber optical modulator or switch, we can no longer ignore the interference between the diffracted orders. If we place a fiber in the position of a bright pixel, the incident angles of the diffracted orders and the interference between them will lead to very inefficient coupling of the light into the fiber. In principle we can combine the two diffracted orders with high efficiency, but it requires another grating or other optical device, and the phase relationships of the two diffracted beams must be controlled with subwavelength accuracy (interfereometric precision). One possible way around these problems is to use only one diffracted order and block the other, but that means we are throwing away more than half the light, so it is not an efficient design.

The conclusion is that Schlieren type optics is impractical for fiber optics. What we need is a modulator that uses the reflective state as the bright state (on state). The diffractive state will then be the dark state (off state) and must be engineered for low dispersion so we can achieve high system contrast.

11.2.1 Three-level Grating Light Modulator

To see how to design a modulator with a low-dispersion reflective state, consider the three-level grating of Fig. 11.1. Here the movable and fixed ribbons are of different widths, and the space between them is large enough that we also get significant reflections from the substrate. There are therefore three reflected-field components (phasors) that are added to give the total reflected field. The relative strengths of these field components are given by the relative widths of the ribbons and the space that separates them.

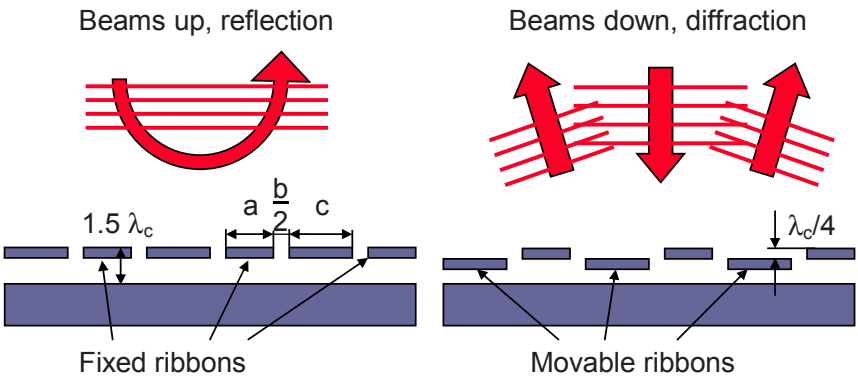


Figure 11.1. Three-level grating modulator designed for low dispersion of the diffractive state. The fixed ribbons are an integer number of half wavelengths above the substrate, so the reflections from the fixed ribbons and the substrate are always in phase at the center wavelength. In the reflective state, all reflections are in phase at the center wavelength, but there is some dispersion due to the ribbon-substrate gap. In the diffractive state, the movable ribbons are moved down so that their reflections are out of phase with the reflections from the fixed ribbons and from the substrate. This creates a low-dispersion diffractive state, because at off-center wavelengths, the phase error of the fixed-ribbon-to-substrate gap compensates for the phase error of the fixed-to-movable ribbon gap.

The reflections from the fixed ribbons and from the substrate are in-phase at the center wavelength. In the structure shown in Fig. 11.1, this is achieved by setting the height of the fixed ribbons over the substrate to *1.5* times the center wavelength, so that the total phase difference between the reflections off the fixed ribbons and the substrate is 6π radians. In principle we could have chosen the path-length difference to be any integer number of half wavelengths, corresponding to an integer number of 2π phase difference.

In the reflective state, the movable ribbons are in the same plane as the fixed ribbons, so that all the reflections are in phase at the center wavelength. The phasors representing this state are shown in green in Fig. 11.2a). When the incident light is at a slightly longer wavelength, the substrate reflections are delayed by a little less than 6π radians, so the overall reflected light is also lagging behind in phase as demonstrated by the dashed phasors in Fig. 11.2a). Associated with this phase lag, there is also a slight amplitude reduction a shown.

The substrate reflections of wavelengths that are a little shorter than the center wavelength, on the other hand, accumulate slightly more than 6π radians phase de-

lay relative to the ribbon reflections. The result is that the overall reflected field is advanced in phase relative to fields at the center wavelength, and again there is an associated reduction in amplitude.

Figure 11.2a shows that the three-level grating modulator has significant dispersion in its reflective state. The magnitude of the dispersion is similar to that of the diffractive state of the basic modulator and the high contrast modulator in the diffractive state. This is acceptable, however, because in the three-level grating modulator the reflective state will be used as the on state, so the dispersion will only slightly color the output light without significantly reducing contrast.

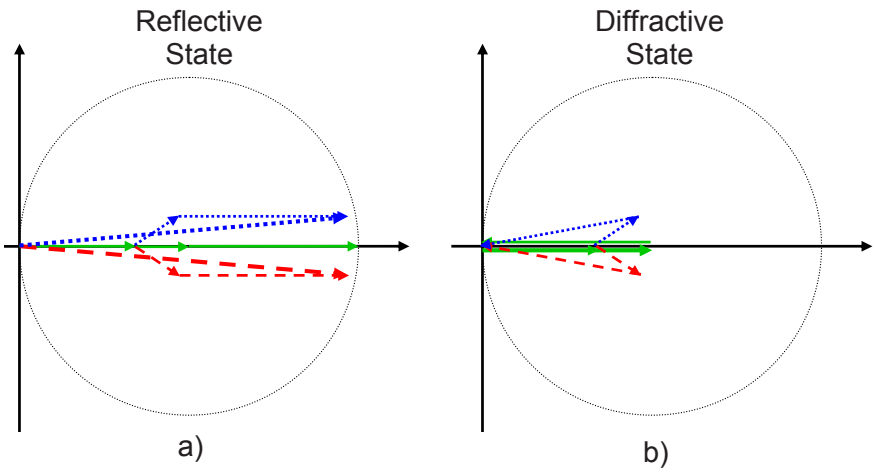


Figure 11.2. Phasor representation of three different wavelengths reflected from a three-level grating light modulator designed for low dispersion of the reflected dark state. At the center wavelength (solid), all the phasors add in phase, while at wavelengths slightly longer (dashed) or shorter (dotted) than the center wavelength, phase errors lead to significant dispersion in the reflective state. The diffractive state, on the other hand, has very low dispersion, because the phase errors of the substrate and movable-ribbon reflections compensate each other.

The usefulness of the three-level grating modulator becomes clear when we consider dispersion from its diffractive state. In this state, the movable ribbons are positioned one quarter of a center wavelength below the fixed ribbons, so the movable-ribbon reflections lag behind the fixed-ribbon reflection by π radians. At the center wavelength, the substrate and fixed-ribbon reflections are in phase (or more correctly 6π radians out of phase), and their sum is exactly cancelled by the movable-ribbon reflections as shown in Fig. 11.2b).

To get perfect cancellation, the combined width of the movable ribbons must equal the combined widths of the fixed ribbons and the separating gaps. In other words, perfect cancellation requires:

$$a + b = c \tag{11.1}$$

where a is the fixed ribbon width, b is the combined widths of the two gaps in each period of the grating, and c is the width of the movable ribbons.

At wavelengths slightly longer than the center wavelength, the phase difference of the reflections from the fixed gratings and from the substrate is slightly less than 6π radians, so their sum also lags in phase relative to the fixed-ribbon reflections. Similarly, the movable-ribbon reflections are not quite π radians behind the fixed ribbon reflections. The net result, shown by dashed phasors in Fig. 11.2b is that the total reflected field is very close to zero even for wavelengths longer than the center wavelength.

We can go through an analogous argument for the wavelengths shorter than the center wavelength. Now the substrate reflections are more than 6π radians, and the movable-ribbon reflections more than π radians, behind the fixed ribbon reflections. Again the result is that the total reflected field is very close to zero also for wavelengths that are shorter than the center wavelength. This is shown in dotted phasors in Fig. 11.2b.

11.2.2 Optimum Design of Three-Level Grating Modulator

The cancellation of the reflected fields from the diffractive state of the three-level grating modulator is not exact. To find the optimum design and to be able to calculate how much residual dispersion we are left with, we must consider the vector sum for non-center wavelengths in detail. The goal is to have the three vectors representing the reflections from the fixed ribbons, the movable ribbons, and the substrate form a triangle as shown in Fig. 11.3. In other words, we would like to find a combination of lengths (a,b,c) and angle ratio (α/β), that allow the three vectors to form a triangle.

The ideal solution would be independent of the angle β, because that would mean that the solution is valid for all wavelengths. Unfortunately, it is immediately clear that such a solution do not exist. If we set $\beta=0$, we see that the solution is $a+b=c$, as noted above. It is obvious, however, that this solution does not work for $\beta\neq 0$. The best we can hope for is therefore to find a solution that minimizes the dispersion across the wavelength range of interest.

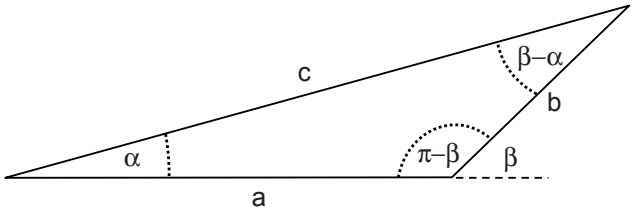


Figure 11.3. Phasors representing the reflections from the fixed ribbons (a), substrate (b), and the movable ribbons (c) of the three-level grating light modulator in the diffractive state. The angles α to β represent the phase lag relative to the fixed ribbons of the movable ribbons and the substrate respectively.

In Figure 11.3, the vector lengths, *a, b*, and *c* represent the areas of the fixed ribbons, gaps, and movable ribbons, while the angles α and β are the phase of the reflections from the movable ribbons and the substrate relative to these same reflections at the center wavelength. The ratio $k=\beta/\alpha$ is therefore a constant equal to the ratio of the distance of the movable ribbons and the substrate from the fixed ribbons at the center wavelength. The numerical value of the constant $k=\beta/\alpha$ must be an even integer, because the operation of the three-level grating modulator requires that the substrate must be an integer number of half center wavelengths below the fixed ribbons, while the movable ribbons must be a quarter of a center wavelength below the fixed ribbons in the diffractive state. For the specific modulator design of Fig. 11.1, the ratio is 6, but it could in principle just as well have been 2 or 4 or some other even integer.

We start our analysis of the vector, or phasor, sum representing the reflections from the three-level grating modulator by using simple geometrical identities to write:

$$c^2 = b^2 + a^2 - 2ba \cdot \cos(\pi - \beta) \tag{11.2}$$

$$\begin{aligned} &\frac{c^2}{b^2} = 1 + \frac{a^2}{b^2} + 2\frac{a}{b} \cdot \cos\beta \Rightarrow \\ &\frac{c^2}{b^2} = 1 + \frac{\sin^2(\beta - \alpha)}{\sin^2(\alpha)} + 2\frac{\sin(\beta - \alpha)}{\sin(\alpha)} \cdot \cos(\beta) \end{aligned} \tag{11.3}$$

For small angles, this simplifies to:

$$\frac{c^2}{b^2} = 1 + \frac{(\beta - \alpha)^2}{\alpha^2} + 2\frac{\beta - \alpha}{\alpha} = \frac{\beta^2}{\alpha^2} \Rightarrow \frac{c}{b} = \frac{\beta}{\alpha} = k \tag{11.4}$$

In the small-angle limit, we find the following solution that is indeed independent of wavelength:

$$\frac{c}{b} = \frac{\beta}{\alpha} = k \tag{11.5}$$

and

$$c = a + b \tag{11.6}$$

The second part of this solution is what we earlier found to be valid at the center wavelength. It simply says that the vector must add up to zero when they are parallel or close to parallel (small angels). We also find that the ratio of the movable-ribbon reflections to the gap reflections must equal the ratio, k, of the phase lags. To first order, the value of k does not matter, so we are free to choose it so that we can simplify the actuation or the fabrication of the modulator, or we can use some other criteria that is important for a given application or technology.

The modulator of Fig. 11.1 fulfills both conditions (Eqs. 11.5 and 11.6), so it does indeed have zero dispersion in the limit of small optical bandwidths (small wavelength variations leading to small phase angles). In fiber optics, the fractional bandwidths are small, typically on the order of a few percent, so the small-angle solutions are often sufficient.

11.2.3 Contrast in the Three-level Grating Modulator

As noted above, the solution is not perfect for finite optical bandwidths. To see how well the three-level modulator extinguishes broadband light, consider a situation where the spectrum of the incident light is centered at a wavelength λ_c and has a fractional bandwidth of $\Delta\lambda/\lambda_c$. For the shortest wavelength in the incident spectrum, $\lambda_c+\Delta\lambda$, the phase lag of the reflections from the substrate is $0.5 \cdot k \cdot \pi \cdot \Delta\lambda/\lambda_c$, and for the longest wavelength it is $-0.5 \cdot k \cdot \pi \cdot \Delta\lambda/\lambda_c$. Similarly, the phase lags for the reflections from the movable ribbons in the diffractive state are $\pm 0.5\ \pi \cdot \Delta\lambda/\lambda_c$ for the extreme wavelengths. We can then use vector summation to find the power reflectivity at the shortest wavelengths in the diffractive state:

$$R_{\lambda+\Delta\lambda} = \frac{\left|\vec{E}_{fixed} + \vec{E}_{substrate} + \vec{E}_{movable}\right|^2}{\left(\left|\vec{E}_{fixed}\right| + \left|\vec{E}_{substrate}\right| + \left|\vec{E}_{movable}\right|\right)^2}$$
$$= \frac{\left[a + b\cdot\cos\left(\frac{k\pi\cdot\Delta\lambda}{2\lambda_c}\right) - c\cdot\cos\left(\frac{\pi\cdot\Delta\lambda}{2\lambda_c}\right)\right]^2}{(a+b+c)^2} \tag{11.7}$$
$$+ \frac{\left[b\cdot\sin\left(\frac{k\pi\cdot\Delta\lambda}{2\lambda_c}\right) - c\cdot\cos\left(\frac{\pi\cdot\Delta\lambda}{2\lambda_c}\right)\right]^2}{(a+b+c)^2}$$

Combined with the condition $c = a + b$, this gives

$$R_{\lambda+\Delta\lambda} = \frac{1}{4}\left\{\begin{array}{l}\left[\left(1-\frac{1}{k}\right) + \frac{1}{k}\cdot\cos\left(\frac{k\pi\cdot\Delta\lambda}{2\lambda_c}\right) - \cos\left(\frac{\pi\cdot\Delta\lambda}{2\lambda_c}\right)\right]^2 \\ + \left[\frac{1}{k}\cdot\sin\left(\frac{k\pi\cdot\Delta\lambda}{2\lambda_c}\right) - \sin\left(\frac{\pi\cdot\Delta\lambda}{2\lambda_c}\right)\right]^2\end{array}\right\} \tag{11.8}$$

This is a relatively complex-looking formula, but it only has two free parameters, the factor k and the fractional bandwidth $\Delta\lambda/\lambda$, and it is further simplified by the fact that k can only take even integer values. To gain an understanding of how low the reflectivity is at extreme wavelengths, we plot the reflectivity as a function of fractional bandwidth for a few values of the factor k. This is done in Fig. 11.4 for fractional bandwidths up to *0.5*, which is roughly the fractional bandwidth of white light.

It is immediately clear from Fig. 11.4 that the reflections at a fractional bandwidth of *0.5* are too high for all three designs. Even in the best case of $k=2$, the reflections at the extreme wavelengths are just a little less than -10 dB, leading to contrast values on the same order. This is not sufficient for most practical white-light imaging applications that require as high as *30 dB* contrast for best image quality as discussed before.

For fiber optics, on the other hand, the situation looks very good. The Conventional band, or C-band for short, of single-mode fiber optical communications ranges from *1530 nm* to *1565 nm*, so $\Delta\lambda/\lambda_c \approx 0.02$. Over this narrow wavelength range we see from the figure that the attenuation of the reflected light is better than *-50dB* even in the worst case. If we also include S-band (for short-wavelength band), that goes from *1460 nm* up to C-band, and L-band, that goes from C-band to *1625 nm*, we have a total fractional bandwidth of about *0.1*. In

this case the maximum attenuation is about *-24 dB* for *k=6*, *-29 dB* for *k=4*, and *-38 dB* for *k=2*. These numbers are much worse than for C-band, but still useful for many applications in fiber optics.

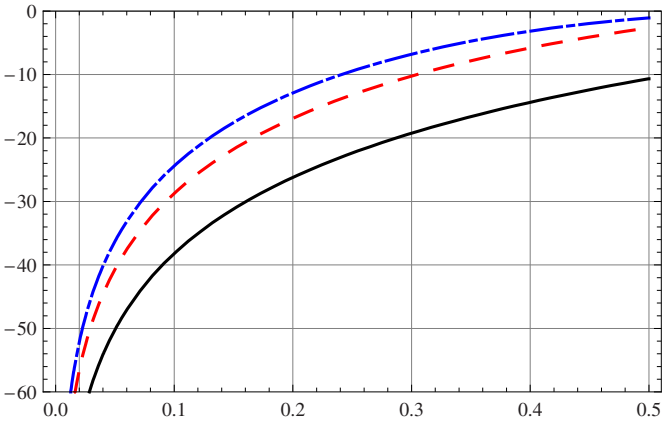


Figure 11.4. Reflection on a dB scale as a function of relative bandwidth from a three-level modulator in the diffractive state. The three curves represent the reflection for three different values of the phase difference between the two fixed levels of the grating. The solid line is for a phase difference of 2π, the dashed for 4π, and the dot-dashed for 6π.

It is possible to improve the contrast across a broad input spectrum by allowing some reflection at the center wavelength. That allows us to create two reflection nulls at wavelengths on either side of the center wavelength. The key here is the symmetry of Eq. 11.3. Consider the situation where we have an input spectrum with a width $\Delta\lambda$ centered at λ_c. By choosing the ratio *c/b* so that Eq. 11.3 is fulfilled at $\beta = 0.25 \cdot k \cdot \pi \cdot \Delta\lambda/\lambda_c$, it is also automatically fulfilled at $\beta = -0.25 \cdot k \cdot \pi \cdot \Delta\lambda/\lambda_c$. This value of *c/b* and the corresponding value of *a/b* found from Eq. 11.3, creates two nulls in the reflection spectrum at $\lambda_c \pm 0.25 \cdot \Delta\lambda$. There will be finite reflections at λ_c, but the maximum reflection for any wavelength in the spectrum will be lower than when we use the small angle approximation to Eq. 11.3.

We will not pursue this approach in detail here, but rather leave it as an exercise for the reader. The conclusions of the analysis will not change substantially, however: The three level grating is excellent for relatively narrow band applications, for example fiber optic modulators in C-band, but not for applications that require truly wide-band operation like imaging with a white-light source.

11.2.4 Wavelength Dependence of Attenuation

Figures 11.2 and 11.3 do not tell the whole story about the three-level grating modulator. In addition to the purely reflective state explained in Fig. 11.2a, and

the purely diffractive state of Fig. 11.2b and Fig. 11.3, we must also consider intermediate states as shown in Fig. 11.5. We see that if the modulator is set to give a small, but finite reflection at the center wavelength, then the longest wavelength will get a significantly larger reflection, while the shortest wavelength will get less. The result is that there is significant variation of attenuation across the optical input spectrum.

The variation of attenuation cross the spectrum from the three-level modulator as illustrated in Fig. 11.5 is of little consequence in some applications, but creates problems in others. In particular, Voltage-controlled Optical Attenuators (VOAs) used to flatten the spectrum in WDM fiber optic systems must meet stringent requirements on spectral variation at different levels of attenuation. VOAs based on a single three-level grating modulator has been commercialized for C-band operation, but operation across the combined S, C, and L bans require more complex solutions incorporating several modulators.

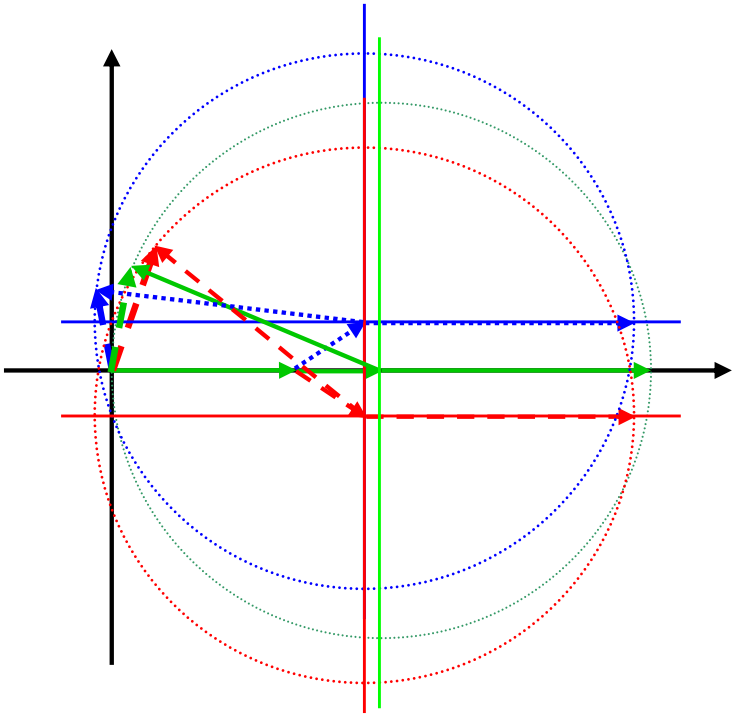


Figure 11.5. Phasor representation of three different wavelengths reflected from a three-level grating light modulator in an intermediate state. At the center wavelength (solid), the phasors add to the selected magnitude, but at longer wavelength (dashed) the magnitude is larger than selected, and at shorter wavelengths (dotted) it is smaller. The result is significant variation of attenuation across the spectrum.

11.2.5 Alternative Modulator Architectures

The three-level grating modulator of Fig. 11.1 is a simple and ingenious solution to the problem of making a modulator with low dispersion in its reflective state. The structure of the three-level modulator is practically speaking not any more complicated than the basic or high-contrast modulators, because in all three cases we must fabricate one and only one layer of ribbons above a substrate, and all the movable ribbons move uniformly, i.e. only one control voltage is needed.

Clearly there are more complex variations of the grating modulator that can be very useful for specific applications. For example, the three-level modulator of Fig. 11.6 has the same ribbon structure as the three-level modulator we have discussed up to now, but both of the two sets of ribbons are movable. This simple modification allow us to optimize the modulator for a given center wavelength. We move all ribbons such that their reflections are in phase with the substrate reflections (or more correctly out of phase by some integer of 2π radians) at the chosen wavelength. Around this center wavelength we can then modulate the reflectance spectrum by moving the set of wider ribbons from the reflective state in which all ribbons are in the same plane, to a diffractive state in which the reflections from the two set of ribbons are out of phase (i.e. the path-length difference is π radians).

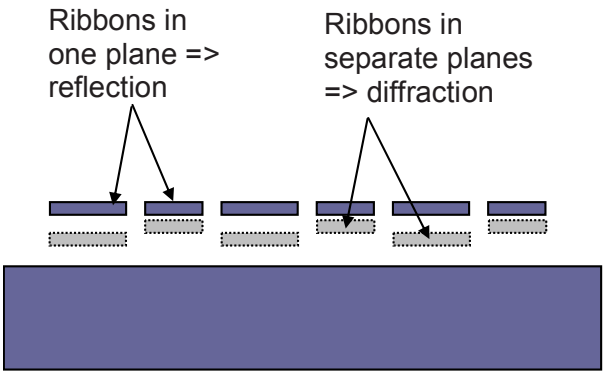


Figure 11.6 *The figure shows two overlaid states of a three-level grating modulator with two sets of movable ribbons. In the reflective state (dark colored ribbons) the ribbons are all in the same plane, and the reflections from the ribbons are all in phase. In the diffractive state (light colored ribbons with dashed outlines), every other ribbon is actuated to create a path length difference of π radians at the chosen center wavelength.*

Being able to optimize the path-length differences for operation at a chosen wavelength does not give the modulator more bandwidth, but we can choose where in the spectrum that bandwidth is applied. For example, we can modulate three different light sources in the visible wavelength range, let's say Red, Green, and

Blue, sequentially with the same modulator, and achieve high contrast for each color. So even though the three-level modulator cannot directly modulate white light with high contrast in the reflective state, it can indeed be used to create a high-contrast RGB image using three different light sources.

The three-level grating modulator of Fig. 11.6 is only marginally more complex to fabricate than the standard three-level modulator. Clearly there are many variations on this basic structure that allows optimization of one important figure of merit or other. For example, we can envision making four-level modulators that have very good broad band contrast, and that combine good color filtering in one state with very high optical efficiency in another. For almost any application that can be performed as a sequence of relatively narrow band modulation functions, there will be a grating light modulator that performs very well and that is simple to fabricate and operate in practice. This illustrates the point we made in the introduction to this chapter. The flexibility and precise dimensional control that we get from integrated-circuit fabrication technology, makes MEMS the preferred technology for implementations of grating modulators.

11.3 Polarization Independent Grating Light Modulators

In our discussions so far we have modeled the grating light modulator in all its variations as one-dimensional in the sense that the optical interaction between the incident light and the grating is confined to a plane, in which the grating is periodic in one dimension and uniform in the other. In reality, of course, the gratings are three-dimensional objects as shown in Fig. 11.7. A complete model of the optical interaction would have to include the fact that the gratings have depth that leads to shadowing and other effects, and that they are uniform only over a finite length leading to additional diffraction effects from the terminations of the ribbons. In practical grating modulators, these effects have only insignificant influence on operation so they can, for most practical purposes, be ignored.

Polarization dependence, on the other hand, is an issue that is important for many practical implementations for the grating light modulator. It becomes progressively worse as device dimensions are scaled down. This is unfortunate, because the most important strength of MEMS technology is that it enables miniaturization. Both the lateral and vertical dimensions of the gratings can be controlled with high precision. Vertical precision is of course necessary for phase control, while lateral accuracy allows us to create very narrow ribbons. As for any integrated-circuit device, the tendency is to shrink it to the minimum dimensions that can support the desired functionality. For the grating light modulator, this means that the period of the structure will be not much larger than a wavelength. It is exactly at this length scale that periodic structures exhibit the strongest polarization effects.

For many applications, polarization dependency is simple to deal with. For example, in a laser-based imaging system, the input polarization to the modulators can be held constant so that any polarization sensitivity is immaterial. Unfortunately, there are applications for which this approach will not work. Notably in fiber optics, the polarization state of the incident light is randomized by temperature dependent and time-varying birefringence on the fiber. For fiber optic applications, it is therefore necessary to modify the grating modulator to remove or reduce its polarization sensitivity.

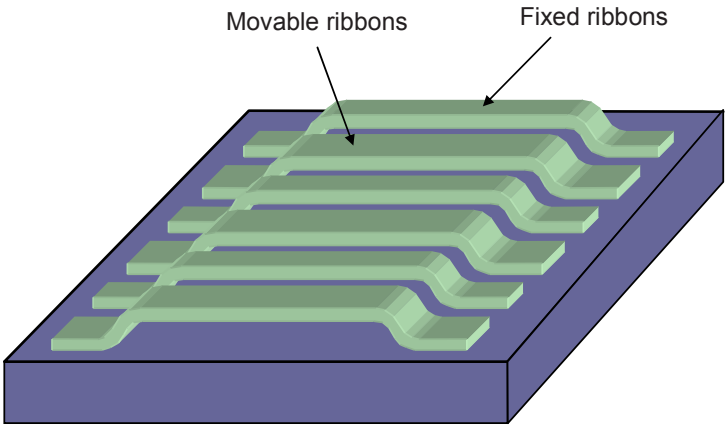


Figure 11.7 *The structure of the grating modulator lead to a number of second order effects, including shadowing of the substrate reflections by the ribbons, diffraction from the ribbon terminations, and polarization dependence caused by the electric field of the incident optical beam sees a different structure whether the electric field is oriented along or perpendicular to the ribbons. The polarization effects increase as the dimensions of the grating ribbons approach the wavelength of the incident light.*

The cause of polarization sensitivity in gratings is their rectilinear geometry. The ribbons of the grating light modulator of Fig. 11.7 interact differently with light whether its electric field is pointing along or perpendicular to the ribbons. This realization points the way to the solution to the polarization dependency problem. It is caused by the geometry, so we must change the geometry to remove the problem. Specifically, we must make a grating that is the same for all polarizations.

One solution is a quadratic lattice with a four-fold symmetric unit cell aligned with the lattice as shown in Fig. 11.8. By symmetry, a grating of this design will have the same response to light at normal incidence whether the light is polarized along the <10> or the <01> direction of the lattice. Any linear optical response of light at normal incidence must then be completely polarization independent, because the incident optical field can be expressed as a superposition of fields in the

<10> and <01> directions, and the response to each part of the superposition then add linearly to give the total response.

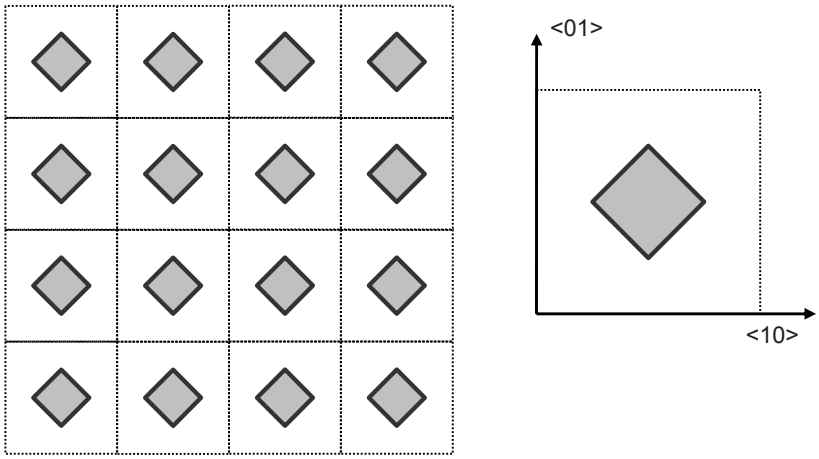


Figure 11.8 Square lattice with four-fold rotation symmetric unit cell. Optical devices with this symmetry are polarization insensitive in all linear response to light at normal incidence.

The symmetry of Fig. 11.8 guarantees polarization independence only for optical fields at normal incidence. Any tilt between the optical axis of the excitation light and the lattice will break the symmetry and allow, but not necessitate, polarization dependence. To minimize polarization effects on oblique incidence and for focused light (which can be considered a superposition of plane waves at different incidence), optical designers therefore often use structures of higher symmetry in an attempt to minimize polarization effects.

An example of such a highly–symmetric grating is the Lightconnect grating modulator [3] shown in Fig. 11.9. The design has a square lattice and a circularly symmetric pedestal reflector.

The square lattice is a good design for reducing polarization dependence of linear optics, but under certain conditions, the hexagonal lattice, shown in Fig. 11.10 is better. To see that the hexagonal symmetry also leads to polarization independence, consider light polarized in the horizontal and vertical directions (shown as solid and dashed vectors in Fig. 11.10). First we see that by symmetry, the response of normally-incident light polarized along the vertical direction and along the directions at ±30 degrees to the horizontal must be the same. The linear response to horizontally polarized light, which can be expressed as a superposition of fields along the directions at ±30 degrees to the horizontal, is therefore equal to the response to vertically polarized light. As for the square lattice, we have that

two orthogonal polarizations have the same response, which means that the structure is polarization independent at normal incidence.

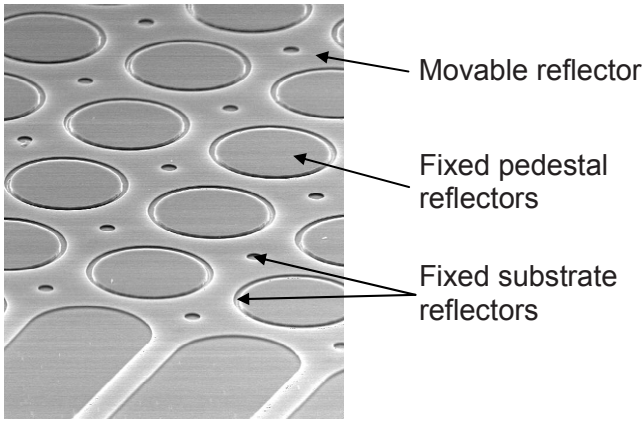


Figure 11.9 *The Lightconnect grating modulator is optimized for fiber-optic applications. It has a three-level reflector design to minimize dispersion, and it is four-fold rotation symmetric to minimize polarization sensitivity. Reprinted with permission.*

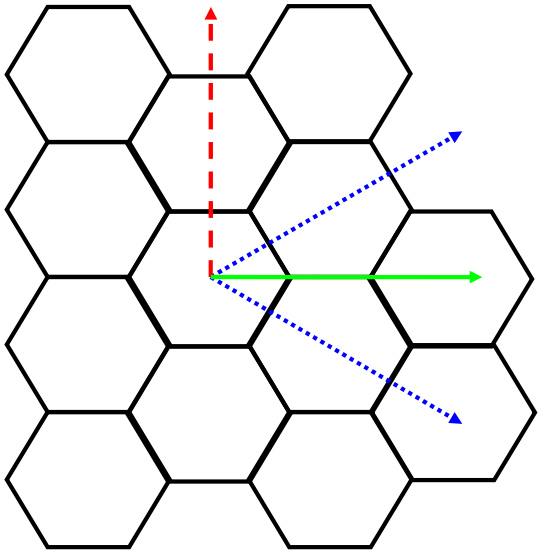


Figure 11.10 *Hexagonal lattice and unit cell, whose responses are independent of polarization for light at normal incidence and uniform in its polarization dependence for off-normal incidence.*

Both quadratic and hexagonal lattices will in general be polarization dependent to light at off-axis incidence. What makes the hexagonal lattice preferable for certain device structures is that the polarization dependence is the same for all directions of the projection of the optical axis on the lattice. In other words, the polarization dependence at a given incident angle is the same no matter what direction the optical axis is tilted. This is not the case for the square lattice, which can have different polarization dependencies for light tilted along different directions. This difference is, however, typically not sufficiently important for the optical designer to choose the hexagonal over the square lattice. The simpler layout and more straight-forward wiring of the multiplexing circuitry of quadratic lattices make them the choice for many practical array implementations.

11.4 Summary of GLMS for Fiber Optics

Fiber optic modulators are quite different from display modulators, because the phase sensitivity of coupling to single mode fibers makes Schliern type optics impractical for fiber optics. To get the required throughput, fiber-optic systems need modulators that use the reflective state as the on state and the diffractive state as the off state.

The most important finding of this chapter is that a three-level GLM can be designed to have very low dispersion in the diffracting state (off state), so that high system contrast can be achieved. The key to the low dispersion of the three-level GLM is that in the diffractive state the reflections from the movable ribbons are out of phase with the reflections from the fixed ribbons and from the substrate, such that the phase error of the fixed-ribbon-to-substrate gap compensates for the phase error of the fixed-to-movable ribbon gap at off-center wavelengths. The three-level grating modulator has significant dispersion in its on state. This dispersion will slightly color the output light without significantly reducing contrast.

The most important finding of the last section of this chapter is that the polarization dependence of the GLMs is due to their geometry, and that gratings with square or hexagonal unit cells have no polarization dependence at normal incidence. At off-normal incidence, hexagonal gratings perform better than ones with square unit cells, but the later has the seemingly minor, but often determining, advantage that it is simpler to lay out and wire.

Further Reading

Fiber-Optic Communication Systems, Third Edition, Govind P. Agrawal, Wiley ISBN 0-471-21571-6, 2002.

Exercises

Problem 11.1 - Two-Moving-Levels Grating Light Modulator

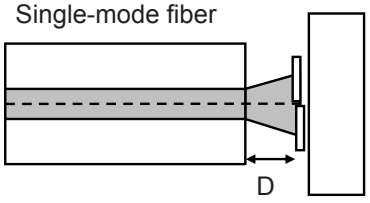
The text describes a three-level GLM with one movable level and two fixed levels. Consider now a three-level GLM with two movable levels.

a. Using phasor diagrams, describe the dispersion characteristics of this GLM. What are the advantages of two moving levels?

b. What type of applications would benefit from a two-moving-levels GLM?

Problem 11.2 - Single Phase Step Modulator

The mode selective properties of single-mode fiber make it possible to create a fiber modulator that has only two mirrors (or even only one if the substrate is used as a reflector). Such a modulator can be made with a focusing lens that creates a well defined beam waist on the modulator and that captures all specularly reflected light, or it can be made by simply sticking the fiber up against the mirrors without the use of a lens as shown in the figure.

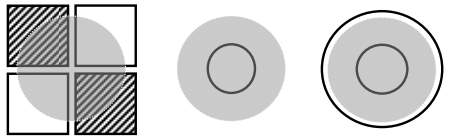


Single Phase Step Modulator. The two mirrors create a phase step along the dividing line between the mirrors.

a. Explain qualitatively how the modulator works.
b. Calculate the maximum back reflection as a function of the fiber to-to-modulator distance D.
c. Using phasors, show how the back reflection depends on the height difference of the mirrors, when the phase step is off set from the center position by δ.
d. The modulator can also be made with multiple phase steps. What are the advantages and difficulties of single phase step modulators?

Problem 11.3 - Polarization Insensitive Fiber Modulator

In Problem 11.2 the phase step is along a straight line. This leads to polarization dependence of the modulated output. The figure shows two attempts at removing or reducing the polarization dependence.



Designs of polarization-insensitive phase step modulators. In the four mirror design, pairs of diagonally opposite mirrors move as a unit.

a. Will each of thee designs prohibit polarization dependence? Explain your answer.
b. What should be the radius of the circular mirror?
c. Which one of these solutions is more practical?

Problem 11.4 - Fabry-Perot GLM

In the GLMs we have described each phase modulator is a simple reflector, and phase modulation is created exclusively by changes in path length. Now consider a GLM in which each phase modulator is a Fabry-Perot.

a. How could you realize such a modulator?
b. What would be the advantages in terms of required mechanical motion?

References

1 R.B. Apte, F.S.A. Sandejas, W.C. Banyai, D.M. Bloom, "Deformable Grating Light Valves for High Resolution Displays," Society for Information Display'93 Symposium, Seattle, Washington, May 17th, 1993.
2 E. L. Goldstein, L. Eskildsen, A.F. Elrefaie, "Performance Implications of Component Crosstalk in Transparent Lightwave Networks", Photonics Technology Letters, Vol. 6, No. 5, May 1994, pp. 657-660.
3 A. Godil, "Diffractive MEMS technology Offers a New Platform for Optical Networks", Laser Focus World, May 2002.

12. Optical Displacement Sensors

12.1 Introduction to Optical Displacement Sensors

The main theme of this book is manipulation of light, i.e. how to use amplitude and phase modulation to shape and direct the optical field. Mostly we have concentrated on positioning and deforming of reflecting surfaces to achieve the desired optical effect. In this chapter we will consider the opposite; How to use measurements of the optical power to deduce the position, deflection, or rotation of a mechanical structure. As in the rest of the book, we will investigate systems that, in whole or in part, can be implemented on the chip scale. We will occasionally mention macroscopic measurement systems, but our focus will be on microoptics.

The conceptually simplest optical method for determining the displacement of a mechanical object is to reflect a light beam off the object and record the position of the reflected beam on a position sensitive detector (PSD). If the system is set up properly, then measurements of the position of the optical beam on the PSD allow us to deduce the position of the reflecting object. One much used displacement sensor of this sort is the so called "optical lever" used in Atomic Force Microscopes (AFMs) and shown in Fig. 12.1

An alternative method for measuring the position of the AFM cantilever is the fiber interferometer, as shown in Fig. 12.2. Here the reflected field on the fiber consists of two interfering parts; the reflection from the fiber facet and the reflection from the cantilever. These parts interfere to set up the back reflected light on the fiber and to give the interferometer its position sensitivity.

The theoretical limits for the lever and interferometer measurements are similar and exhibit similar wavelength dependence. In practice, interferometry typically achieves better sensitivity, because it is easier to obtain close-to-theoretical performance with the interferometer. Its sensitivity is practically speaking independent of the transversal beam size. The optical beam of the lever has to be focused on the detector to achieve good sensitivity, so it becomes impractical in the theoretical limit. Nevertheless, the lever is more used because its sensitivity is good

enough for many applications even with large beams, and with a large beam, and therefore a long collimated region, it is very simple and practical.

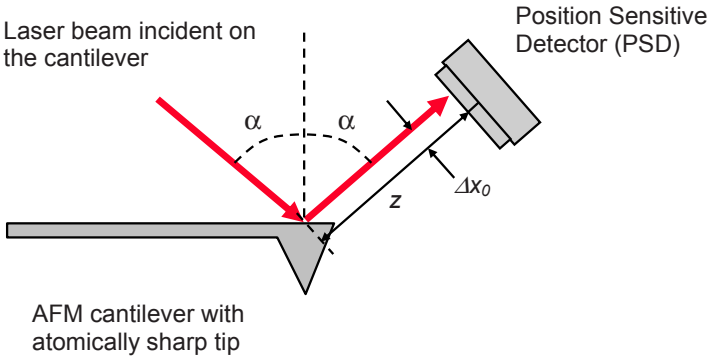


Figure 12.1 *The Optical Lever is a combination of a moving mirror, here placed on an AFM cantilever, and a position-sensitive detector. The position of the mirror is found from the position of the optical beam on the PSD.*

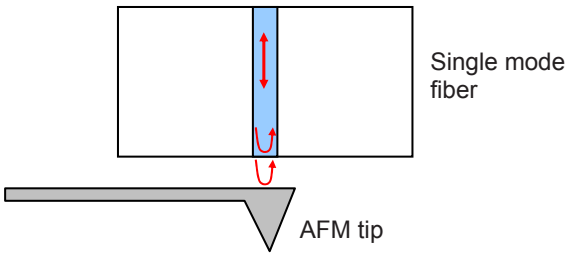


Figure 12.2 *The distance from the optical fiber facet to the AFM cantilever is measured by observing the back-reflected light on the fiber. The back-reflected light consists of two parts: the reflection from the fiber end and the reflection from the cantilever. The interference between these two parts gives the fiber interferometer its sensitivity to position of the cantilever.*

The literature describes a large number of different interferometer designs and applications [1]. We will concentrate on a few types that are well suited to chip-scale integration. Large scale interferometers are typically creating a spatial pattern of fringes, i.e. alternating band of constructive and destructive interference that is called an interferogram. An interferogram resembles a topographical map of a surface with the fringe period corresponding to a height difference of one half of a wavelength (in reflection). Most microscopical interferometers do not produce spatial interferograms, but rather have only a single output channel with a

temporal "fringe pattern", i.e. a time varying signal that reflects changes in phase between two or more signal paths. We will study several such interferometers, including Fabry-Perot interferometers and Michelson interferometers in this chapter.

Another effect that can be used to construct optical position sensors is photon tunneling, i.e. transmission mediated by evanescent fields. The sensitivity of such sensors arises from the exponential decay of the evanescent fields and the corresponding exponential reduction in transmission with tunneling distance. An example of such a sensor is shown in Fig. 12.3, where Total Internal Reflection (TIR) of the incident optical beam sets of evanescent fields in the tunneling gap. The resulting transmission through the tunneling gap is a strong function of the width of the tunneling gap (see section 3.5.3).

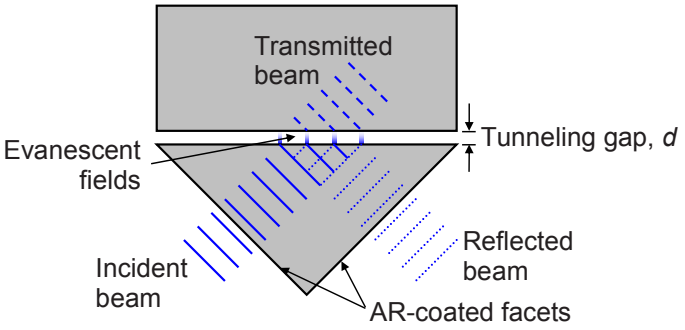


Figure 12.3 *Optical-tunneling displacement sensors based on frustrated Total Internal Reflection (TIR). The ratio of transmitted to reflected optical power is exponentially dependent on the width of the tunneling gap.*

The sensor of Fig. 12.3 is only one of many possible architectures that utilize photon tunneling. Other configurations include optical tweezers used for ultrasensitive displacement and force measurement, and tunneling between coupled states in Photonic Crystals. The former requires large laser systems and it does not lend itself to miniaturization, so we won't consider it further. PC tunneling sensors are covered in Chapter 15.

Any measurement is fundamentally limited by noise, so we must consider noise to give a comprehensive account of optical displacement sensors. An important part of this chapter is therefore devoted to the description of noise in optical sensor systems. Armed with models for noise contributions, we compare the fundamental limits of optical and other displacement sensors and draw conclusions about their relative strengths and weaknesses.

Both the optical lever and several interferometers will be described in detail in this chapter, but the emphasis is on interferometers because they function well on the chip scale and therefore in MEMS implementations.

12.2 Interferometers as Displacement Sensors

12.2.1 The Michelson Interferometer

Optical interferometry is a traditional technique that is used in a number of high-accuracy position measurements. The output of an interferometer depends strongly on the relative distance between two (or more) reflectors that create interfering optical fields. When used as a sensor, the interferometer is set up such that the quantity we want to measure, e.g. pressure or acceleration, displaces one reflector with respect to the other. By measuring the output of the interferometer, we can deduce the separation between the two reflectors, and from that the magnitude of the measurand. Optical interferometry is a very versatile technique; any signal that we somehow can make influence the position of a mirror can be measured.

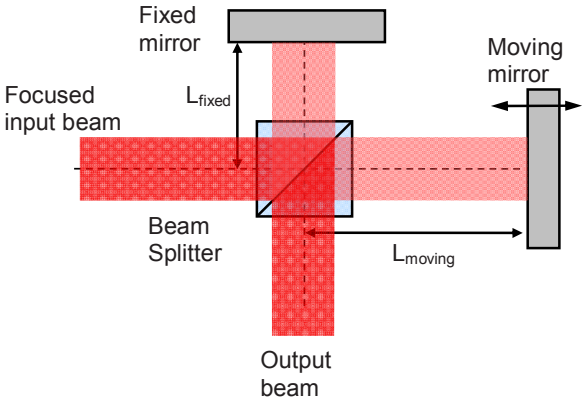


Figure 12.4 *The Michelson interferometer consists of a beam splitter and two mirrors; one fixed reference mirror and one movable target mirror. The incident light is split into the fixed and variable arms of the interferometer and recombined in the beam splitter after reflecting off the mirrors. The phase difference between the two beams upon recombination determines how much light is transferred to the output and how much is coupled to the backwards propagating beam.*

To understand the basics and develop design equations for interferometric displacement sensors, we will first consider the basic Michelson interferometer of Fig. 12.4. The incident optical beam is split into two beams in the (non-polarizing) beam splitter. One beam is directed to a stationary reference mirror and the other to a moving target mirror. The two beams are then reflected back from the two mirrors and recombined in the beam splitter so that some of the light

is sent back towards the input and the rest of the light is deflected down to the output.

Let us now consider the idealized case, i.e. we assume that the monochromatic optical field behaves like a plane wave, and that mirrors are perfectly reflecting and perfectly parallel to the phase fronts of the plane wave. Further we assume that the beam splitter divides the incident light into two identical beams, each with exactly half of the incident optical power. The power[1] transfer function of the Michelson interferometer is then:

$$\frac{P_{out}}{P_{in}} = 0.5 \cdot \left[1 + \cos\left(\frac{2\pi \cdot 2\left(L_{fixed} - L_{moving}\right)}{\lambda} \right) \right] = 0.5 \cdot \left[1 + \cos\left(\frac{2\pi \cdot 2\Delta L}{\lambda} \right) \right] \qquad (12.1)$$

where P_{out} and P_{in} are the output and input optical powers, λ is the wavelength, and $2\Delta L = 2\left(L_{fixed} - L_{moving}\right)$ is the total path length difference for the light reflected from the fixed and moving mirrors. (The factor of 2 results from the fact that the beams propagate back AND forth over the distances from the beam splitter to the mirrors.) The transfer function of the Michelson interferometer is shown in Fig. 12.5.

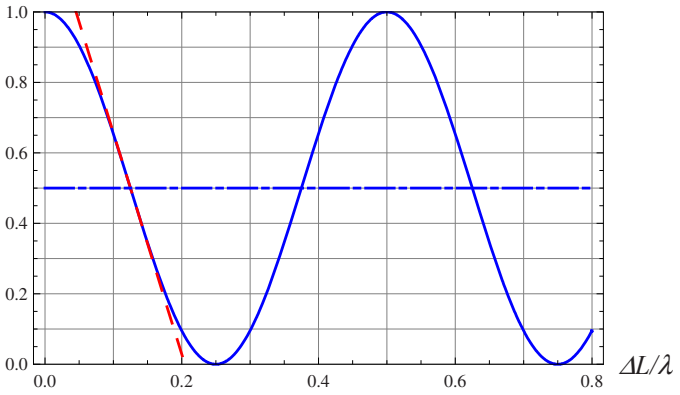


Figure 12.5 *The optical power transfer (solid curve) through a Michelson interferometer as a function of the length difference between the interferometer arms (ΔL) normalized to the wavelength. The dot-dashed line intersects the transfer functions at its highest-slope points, and the dashed curve shows the slope at one of these points.*

[1] Our modeling is based on plane waves, but in practical systems all beams have finite cross sections and finite power, so we use power rather than intensity (power per area), because power is what we actually measure, and this usage simplifies the noise calculation in later sections of this chapter.

The power that isn't transferred to the input in the beam splitter[2] will have to propagate back towards the source, so the transfer function into the backwards propagating beam is the complement of the transfer function to the output:

$$\frac{I_{back}}{I_{in}} = 0.5 \cdot \left[1 - \cos\left(\frac{2\pi \cdot 2\Delta L}{\lambda} \right) \right] \tag{12.2}$$

We should note here that in practice most Michelson interferometers are not operated with their mirrors perfectly normal to the optical axes. In fact, it is often beneficial to tilt one or both mirrors, because that means that the path-length difference is spatially varying over the output beam. This creates a spatially varying pattern of alternating bright and dark lines in the output. These are the well-know interference fringes that are very useful for observing aberrations on the surface of the target mirror. Also neglected are a number of other effects discussed in Chapter 4. These include additional phase shift due to focusing and changes associated with spatial mode filtering.

The periodicity of the interferometer transfer function leads to ambiguity in the measurements of mirror separation if the range over which the target mirror moves is larger than half a wavelength. This ambiguity has led to three very different ways of using an interferometer. For very sensitive measurements, the reflectors are positioned such that the arm-length difference nominally equals an odd integer number of eights of a wavelength ($\Delta L=(2n-1)\lambda/8$ where n is an integer). These are the points, marked in Fig. 12.5, where the transfer function has its maximum slope and therefore its maximum sensitivity to displacement. There will be no ambiguity in the measurement as long as the target mirror moves less than $\pm\lambda/8$.

Another common way to use an interferometer is to "count fringes". While the mirror is moving, we determine how many maxima we see in the output. The number of maxima is then multiplied by half the wavelength to get a measure of the mirror motion with an accuracy on the order of the wavelength.

The third mode of operation is to combine accurate power measurements with fringe counting to lift the ambiguity caused by the periodicity. This method gives both good accuracy and extended dynamic range, but it requires automatic fringe counting, which in turn requires a mechanism for determining the direction the

[2] We will not consider the detailed operation of the beam splitter here, but note that the description of the interferometer indicates that it must have a $\pi/2$ radians phase difference between the transmitted and the deflected light, just like the fiber directional coupler discussed in Chapter 6. The key insight is that the transmitted and back reflected optical fields must be exactly π radians out of phase at the beam splitter. To reach the output, each beam must undergo one transmission and one deflection in the beam splitter, so the total phase difference between transmission and deflection must be $\pi/2$ radians.

mirror moves. Many different techniques involving multiple wavelengths (either simultaneously by having a broad band source or multiple sources, or sequentially by tuning the wavelength of the source) have been developed for a wide variety of applications [2].

A very useful consequence of the periodic transfer function of optical interferometers is that the sensitivity to displacement is independent of the length of the arms and of the difference in arm length. In other word, it does not matter if the interferometer is 10 um long or 10 km long provided that we operate at a maximum-sensitivity point. This means that optical interferometers are extremely well suited for measurements of phenomena that accumulate displacement over a long distance. A good example is gravity wave detection. In the LIGO project [3], optical interferometers are used to try to detect displacement of mirrors that are spaced tens of km apart. Gravity waves, that stretch the whole arm of the interferometer (hopefully), have a chance of being detected, even though the relative elongation of the interferometer arm is very low.

Another type of application that benefit from the periodic transfer function of optical interferometers is measurements of small deflections that might take place over a relatively short distance, but at an unknown location. We can then put the reflectors far apart and deduce from measurements that a small displacement took place somewhere between the mirrors. This is very useful in seismic studies where accurate displacements are important, but their exact location is not.

So optical interferometers are good for measurements of small differential displacements that take place over long distances, but those are of course not typical MEMS or microsystem applications. As we will see, however, the excellent displacement sensitivity of optical interferometers also makes them very useful for chip-scale sensing. That is particularly true for applications that benefit from remote measurements of displacements.

12.2.2 Displacement Sensitivity

To find a mathematical expression for the sensitivity of an interferometer, we take the derivative of the power transfer function with respect to the interferometer arm length difference:

$$\frac{d\left(I_{out}/I_{in}\right)}{d\left(\Delta L\right)} = \frac{2\pi}{\lambda}\cdot\sin\left(\frac{4\pi\cdot\Delta L}{\lambda}\right) \tag{12.3}$$

This expression achieves its maximum absolute value when

$$\sin\left(\frac{4\pi\cdot\Delta L}{\lambda}\right) = 1 \Rightarrow \Delta L = \left(2n-1\right)\frac{\lambda}{8} \tag{12.4}$$

These points are marked as the intersections between the transfer function (blue curve) and the straight line at $I_{out}/I_{in} = 0.5$. The absolute value of the slope at these maximum points is:

$$\left.\frac{d(I_{out}/I_{in})}{d(\Delta L)}\right|_{\max} = \frac{2\pi}{\lambda} \tag{12.5}$$

We will use this expression later to determine the minimum detectable displacement (Noise-Equivalent-Displacement) and dynamic range of the interferometer.

12.2.3 Implementations of Interferometric Displacement Sensors

The Michelson is much used in spectroscopy where it forms the basis of traditional Transform spectrometers described in Chapter 13. Another popular application is biomedical imaging where the Michelson interferometer is used in Optical Coherence Tomography. A variation of the Michelson is the Mach-Zehnder interferometer shown in Fig. 12.6.

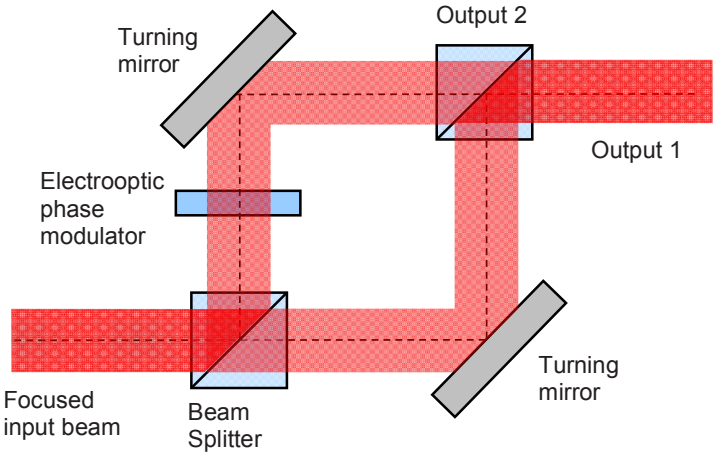


Figure 12.6 *The Mach-Zehnder Interferometer is conceptually very similar to the Michelson interferometer. It consists of two beam splitters and two turning mirrors. The first beam splitter separates the incident optical beam into two parts that travel through the two arms of the interferometer before they are recombined in the second beam splitter. As in the Michelson, the phase difference between the two beams upon recombination determines how much light is transferred to each output. The electrooptic phase modulator allows us to change the phase in one arm of the interferometer. This phase modulation is converted to amplitude modulation during recombination in the second beam splitter.*

The figure makes it clear that the Mach-Zehnder can be viewed as an "unfolded" Michelson. Instead of sending the light back to be recombined in the first beam splitter, the turning mirrors relay the optical beams to a second beam splitter where the two parts are recombined. The transfer function and the sensitivity to path-length differences is the same in the Mach-Zehnder as in the Michelson interferometer, and like the Michelson, the Mach-Zehnder is used in a number of applications.

The Mach-Zehnder lends itself particularly well to optical fiber and optical waveguide implementations, and is therefore the structure of choice for interferometeric amplitude modulators. The basic principle of interferometeric amplitude modulation is illustrated in Fig. 12.6. By inserting an electrooptic phase modulator in one arm of the interferometer, the relative phase of the two recombining beams can be controlled, and therefore the transfer of optical power to the two outputs.

The Michelson and Mach-Zehnder interferometers are very flexible devices that can be implemented in a variety of ways and successfully be applied in a larger number of systems for spectroscopy, optical-phase sensing, and optical modulation. For chip-scale displacement sensing, however, the Michelson and Mach-Zehnder are too complex. A much simpler displacement sensor is the grating interferometer of Fig. 12.7.

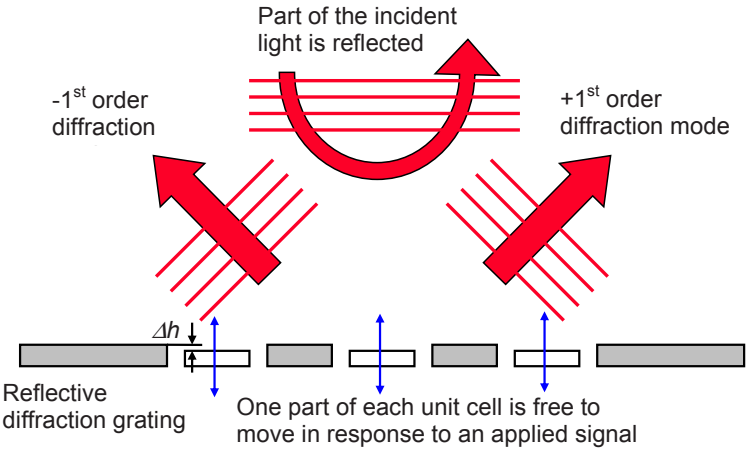


Figure 12.7 *Conceptual drawing of the operation and structure of a grating-interferometer displacement sensor. The output of the interferometer is a function of the offset, or height difference (Δh), between the fixed (grey) and movable (white) sets of reflectors. By measuring the distribution of diffracted and reflected light from the grating, the offset can be found with interferometric precision. This type of interferometer can be implemented in a single thin film and is well suited for MEMS implementations.*

As the Michelson and Mach-Zehnder, the grating interferometer has a collimated optical beam as its input. The input beam is incident on and reflected off a binary phase grating, i.e. a grating that has a unit cell consisting of two offset reflectors. The incident light is partly reflected and partly diffracted, with the ratio of reflected to diffracted power depending on the lateral offset between the reflectors in the unit cell of the grating. The reflected and diffracted optical fields constitute the two outputs of the interferometer. These fields radiate from the grating in different directions, so they are easily separated.

Figure 12.8 shows a variation of the grating interferometer. In this structure, the reflections are coming from two planes that are defined by different layers of the MEMS structure. This variation is useful for applications where it is beneficial that one of the reflecting surfaces is part of a continuous solid body, e.g. a diaphragm in a pressure sensor or a proof mass in an accelerometer. The operation and sensitivity of this interferometer is the same as the single-layer grating of Fig. 12.7, but the wavelength dependence is larger, because the static height difference between the reflectors is larger.

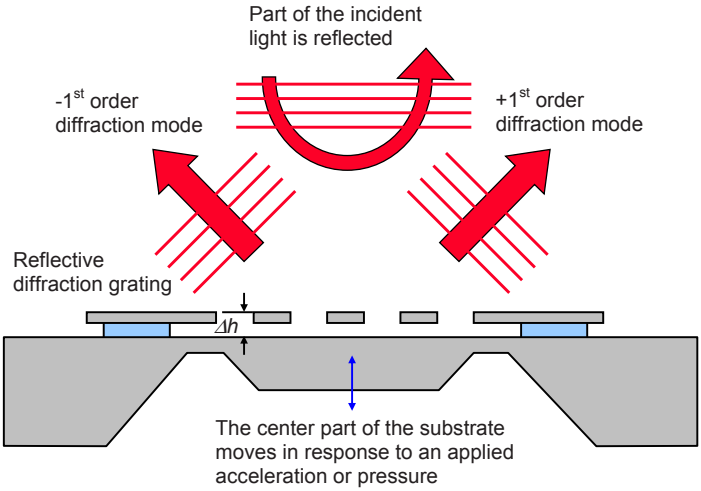


Figure 12.8 *Grating-interferometer used for pressure sensors or accelerometers. In these applications it is beneficial that one reflecting surface is a continuous body, so the two reflectors are placed on two different layers. The operation and sensitivity of this sensor is the same as the basic grating interferometer of Fig. 12.7, but it is typically more sensitive to variations in wavelength, because the nominal height difference (Δh), between the fixed grating and the movable underlying substrate is larger.*

The two-layer-grating implementation of Fig. 12.8 puts less demand on the resolution, or minimum feature size, that the lithography technology must provide. In

this structure, the upper grating can be pattern to have openings with dimensions down to the minimum feature size, because the underlying reflector is solid so that there are no gaps in the overall grating. This feature of the two-layer grating is particularly useful for applications where the light that penetrates the grating can be detrimental to the operation of underlying MEMS structures or electronic circuitry.

From the explanation of the operation of the grating interferometer, we see that the grating interferometer is indeed very similar to the Michelson and Mach-Zehnder. All these interferometers have a single input[3] that is split into two parts and then recombined to create the two outputs. The difference is that in the Michelson and Mach-Zehnder there is a clearly identifiable initial beam splitter, followed by propagation paths for the two beams, and finally a beam recombiner (or recombining beam splitter). In the grating interferometer, all these functions are combined into a single element; the grating itself. It performs the initial beam splitting and propagation-path differentiation by having the beam spatially be separated into two parts that hits the two parts of the grating unit cells. The recombination also happens right at the grating where the two parts interfere to create the two outputs.

In deriving the mathematical description of the grating interferometer, we will for simplicity neglect any gaps between the reflectors of the grating. From Chapter 10 we know that the power of the reflection (or zero-order diffraction) from a binary diffraction grating can be expressed

$$D^0 = \frac{P_0}{2} \cdot \left(1 + \cos\theta\right) \tag{12.6}$$

where P_0 is the power of the incident beam, and θ is the total phase shift difference between the optical fields that are reflected from the two parts of binary unit cell of the grating. In reflection, this total phase shift difference can be expressed as

$$\theta = \frac{2\pi \cdot 2\Delta h}{\lambda} \tag{12.7}$$

where Δh is the height difference or offset between the two parts of the unit cell, and λ is the wavelength of the light. The power transfer function of the grating interferometer is then

[3] Each of these interferometers are linear systems with two outputs, so we know from Chapter 2 they also must have two inputs. The unused input of the Michelson is simply a beam that enters the device on the output port, and for the Mach-Zehnder the extra input enters the first beam splitter vertically from below. In the grating interferometer, the extra input is the reverse of the diffracted field. In practice we sometimes use the Michelson and Mach-Zehnder with two inputs, but not the grating interferometer, because setting up the reverse of the diffracted field is cumbersome.

$$\frac{I_{reflection}}{I_{in}} = 0.5 \cdot \left[1 + \cos\left(\frac{2\pi \cdot 2\Delta h}{\lambda} \right) \right] \tag{12.8}$$

where I_{out} and I_{in} are the output and input optical intensities of the interferometer. Similarly, we find for the power transfer into the n^{th} order diffraction mode, where n is larger than zero:

$$\frac{P_{nth-order}}{P_{in}} = 0.5 \cdot \left[1 - \cos\left(\frac{2\pi \cdot 2\Delta h}{\lambda} \right) \right] \cdot sinc^2 \frac{n\pi}{2} \tag{12.9}$$

We also showed in Chapter 10 that the total diffraction, i.e. the sum over all diffraction orders larger than the 0^{th}, is the complement of the reflection, so the transfer into all diffraction modes is:

$$\frac{P_{all-orders}}{P_{in}} = 0.5 \cdot \left[1 - \cos\left(\frac{2\pi \cdot 2\Delta h}{\lambda} \right) \right] \tag{12.10}$$

We see from Eqs. 12.8 and 12.10 that the transfer functions of the grating interferometer are the same as the for the Michelson and Mach-Zehnder. This is of course what we expect given the similarities between these types of interferometers. In fact, all interferometers with exactly two interfering fields have a harmonic response of the form give by Eq. 12.1, provided that ΔL is correctly interpreted. All these types of interferometers therefore have the same sensitivity to displacements, given by Eqs. 12.3 and 12.5.

The grating interferometers of Figs. 12.7 and 12.8 are very simple. They can be made in a single layer or single thin film, yet they achieve the same displacements sensitivity as much more complex interferometers that contain many more optical components. This simplicity makes the grating interferometer well suited for single-chip integration. Using IC deposition and etching technology together with lithography, we can make on-chip grating interferometers with a few simple fabrication steps. One of the very useful features of IC technology for this purpose is that is allows very accurate thickness control of deposited films and of material removed by etching. It is therefore straightforward to create the static offset of $\lambda/8$ shown in Fig. 12.7. The purpose of this static offset is to operate the interferometer at a point of maximum displacement sensitivity as indicated in Fig. 12.5.

A short coming of the grating interferometer is that it has a limited measurement range, over which it will give unambiguous results. If we are considering only a single measurement point, then the periodicity of the response means that displacements that are different by an integer number of half wavelengths cannot be distinguished. There are several different methods to lift this ambiguity. One way is to continuously record and digitize the optical signal and do a signal-vs-time to position-vs-time conversion during post-processing of the interferometer data. This is unfortunately not as simple as it sounds and may lead to erroneous results

on adverse input data. Another method is the simultaneous use of multiple wavelengths to read out the displacement. Combined with post processing, this method yields good results, but the hardware and software are complicated, so it's used mostly in high-end systems applied to critical applications that can justify the high complexity and cost.

12.2.4 Improved Sensitivity of High-Finesse Interferometers

The interferometers we have studied so far are all two-beam interferometers, i.e. the output is created by the interference of exactly two optical beams. As we have seen, all these interferometers have an output with a harmonic dependence on some characteristic path-length difference. Another class of interferometers is based on recirculating beams and their outputs can be viewed as resulting from interference between a large or even infinite (at least in theory), number of optical beams. One such device is the Fabry-Perot interferometer of Fig. 12.9.

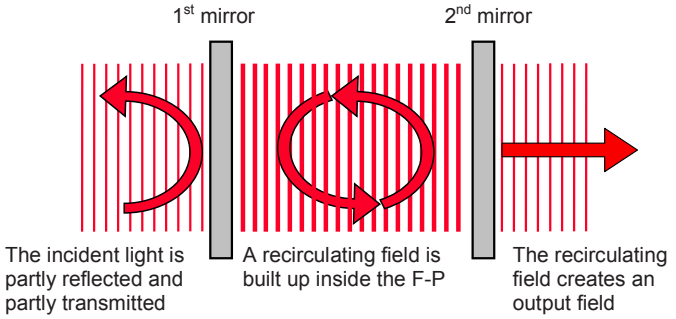


Figure 12.9 *Schematic drawing of a Fabry-Perot interferometer with a plane wave incident from the left. The incident light is partially transmitted through the first mirror. The part that is transmitted is reflected back and forth between the two mirrors to build up a recirculating field between the mirrors. If the wavelength of the incident field is chosen so that the total round trip between the mirrors equals an integer number of wavelength, then the recirculating field builds up to a maximum value, and the transmission through the F-P is unity.*

In the Fabry-Perot interferometer, like the Michelson and Mach-Zehnder, we have one input[4] and two outputs; the transmitted light and the reflected light. For simplicity, we model the input as a plane wave at normal incidence from the left on the left mirror. Some of the incident light is transmitted through the first mirror and is reflected back and forth between the two mirrors to build up a recirculating field in the resonator. The steady-state transmitted field is simply the recirculating field multiplied by the field-transmission of the right mirror, while the reflected

[4] As for two-beam interferometers, the F-P also has in principle two inputs. Here the two inputs are incident fields from the left and right. In our treatment we are only considering the incident field from the left.

field is created as an interference between the portion of the recirculating field that escapes through the left mirror and the part of the incident field that is directly reflected from the same mirror.

The reflected field can be calculated by summing up the field that is reflected from the first mirror with all the field components that have been transmitted through the first mirror and then passed back and forth through the cavity for a number of times before it is transmitted back trough the first mirror

$$E_r = E_i \cdot r_1 - E_i \cdot \frac{t_1^2}{r_1} \sum_{n=1}^{\infty} \left(r_2 \cdot r_1 \cdot e^{-j2kL} \right)^n \tag{12.11}$$

where E_i is the incident field on the first mirror, t is the field transmission trough the first mirror, and r_1 and r_2 are the field reflectivities seen from inside the F-P. The minus sign before the summation is due to the fact that the reflectivities from opposite sides of a mirror when referred to the same reference plane are of opposite signs (see chapter 2).

For simplicity we now assume that the mirrors are lossless, which means that $r_1^2 + t_1^2 = 1$. Using this assumption and carrying out the summation using the standard formula $s = a \cdot r + a^2 \cdot r + a^3 \cdot r + ... + a^N \cdot r = \frac{a^{N+1} \cdot r - a \cdot r}{a-1}$, we find the following expression for the reflected field:

$$E_r = E_i \cdot r_1 - E_i \cdot \frac{1 - r_1^2}{r_1} \cdot \frac{r_2 \cdot r_1 \cdot e^{-j2kL}}{1 - r_2 \cdot r_1 \cdot e^{-j2kL}} = E_i \cdot \frac{r_1 - r_2 \cdot e^{-j2kL}}{1 - r_2 r_1 \cdot e^{-j2kL}} \tag{12.12}$$

Likewise we can find the recirculating field right inside the first mirror by summing all field contributing components at this point in the cavity:

$$E_{circ} = E_i \cdot t_1 + E_i \cdot t_1 \sum_{n=1}^{\infty} \left(r_2 \cdot r_1 \cdot e^{-j2kL} \right)^n = E_i \cdot t_1 \left(\frac{1}{1 - r_2 r_1 \cdot e^{-j2kL}} \right) \tag{12.13}$$

The transmitted field through the cavity is then

$$E_{trans} = E_{circ} \cdot e^{-jkL} \cdot t_2 = E_i \cdot t_1 t_2 \left(\frac{e^{-jkL}}{1 - r_2 r_1 \cdot e^{-j2kL}} \right) \tag{12.14}$$

From these field relationships, we can calculate the reflectance and transmittance spectra of the Fabry-Perot. We start with the reflectance

$$R = \frac{E_r \cdot E_r^*}{E_i \cdot E_i^*} = \frac{\left(r_1 - r_2 \cdot e^{-j2kL}\right)\left(r_1 - r_2 \cdot e^{j2kL}\right)}{\left(1 - r_2 r_1 \cdot e^{-j2kL}\right)\left(1 - r_2 r_1 \cdot e^{j2kL}\right)}$$

$$= \frac{r_1^2 - 2r_2 r_1 \cdot \cos\left(2L\frac{2\pi}{\lambda}\right) + r_2^2}{1 - 2r_2 r_1 \cdot \cos\left(2L\frac{2\pi}{\lambda}\right) + r_1^2 r_2^2} = \frac{\dfrac{\left(r_1 - r_2\right)^2 + 4\sin^2\left(L\dfrac{2\pi}{\lambda}\right)}{\left(1 - r_2 r_1\right)^2}}{1 + \dfrac{4r_2 r_1}{\left(1 - r_2 r_1\right)^2} \cdot \sin^2\left(L\dfrac{2\pi}{\lambda}\right)} \tag{12.15}$$

where we have used the geometrical identity $\cos\theta = 1 - 2\sin^2(\theta/2)$. Similarly we find the transmittance

$$T = \frac{E_t \cdot E_t^*}{E_i \cdot E_i^*} = \left(1 - r_1^2\right)\left(1 - r_2^2\right)\frac{1}{\left(1 - r_2 r_1 \cdot e^{-j2kL}\right)\left(1 - r_2 r_1 \cdot e^{j2kL}\right)}$$

$$= \frac{\left(1 - r_1^2\right)\left(1 - r_2^2\right)}{1 - 2r_2 r_1 \cdot \cos\left(2L\frac{2\pi}{\lambda}\right) + r_1^2 r_2^2} = \frac{\dfrac{\left(1 - r_1^2\right)\left(1 - r_2^2\right)}{\left(1 - r_2 r_1\right)^2}}{1 + \dfrac{4r_2 r_1}{\left(1 - r_2 r_1\right)^2} \cdot \sin^2\left(L\dfrac{2\pi}{\lambda}\right)} \tag{12.16}$$

We see that the reflectance and transmittance add up to unity, as they should given that we have made the assumption that the mirrors are lossless.

$$T + R = \frac{\dfrac{\left(1 - r_1^2\right)\left(1 - r_2^2\right)}{\left(1 - r_2 r_1\right)^2} + \dfrac{\left(r_1 - r_2\right)^2 + 4\sin^2\left(L\dfrac{2\pi}{\lambda}\right)}{\left(1 - r_2 r_1\right)^2} +}{1 + \dfrac{4r_2 r_1}{\left(1 - r_2 r_1\right)^2} \cdot \sin^2\left(L\dfrac{2\pi}{\lambda}\right)} = 1 \tag{12.17}$$

The reflectance and transmittance of the F-P are shown in Fig. 12.10. We see that the response have many similarities to the response of a two-beam interferometer shown in Fig. 12.5. As for the two-beam interferometer, the response as a function of L is periodic with a period of $\lambda/2$. In contrast to the harmonic response of the two-beam interferometer, the F-P response exhibits sharp maxima in the transmittance and sharp minima in the reflectance. These "resonant" peaks occur at wavelengths given by

$$\lambda = n \cdot \frac{2\pi}{L} \tag{12.18}$$

where n is an integer.

The graphs show that as the mirror reflectivities are increased, the transmittance peaks and the reflectance minima become narrower with increasingly steep slopes. This means that by increasing the mirror reflectivities, we increase the sensitivity of the F-P to length changes. By setting up and using a F-P correctly we will therefore be able to create a more sensitive displacement sensor than what we get with a two-beam interferometer.

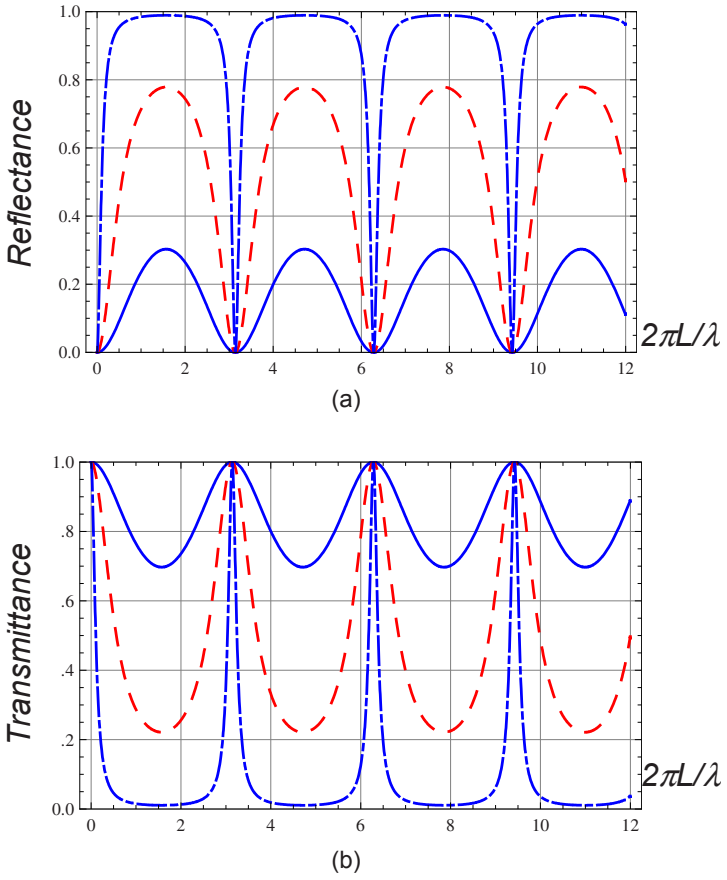


Figure 12.10 *Reflectance (a) and Transmittance (b) of a Fabry-Perot interferometer with identical mirrors with a reflectivity of 0.3 (solid), 0.6 (dashed), and 0.9 (dot-dashed). As the two-beam interferometer of Fig. 12.5, the Fabry-Perot has a periodic response as a function of the characteristic length L. One important difference is that the response of Fabry-Perots with high reflectivities has a much larger maximum slope with respect to length, and therefore can be made into more sensitive displacement sensors.*

To quantify the increase in sensitivity, we simplify the formulas and derive analytical expressions. We have already seen that the reflectance and transmittance are complementary for loss-less mirrors, so we will carry out the derivations for transmittance only. If we make the simplifying assumption that the F-P is symmetric, i.e. the two mirrors have the same reflectivity $r=r_1=r_2$, we find the following expression for the transmittance:

$$T = \frac{\frac{\left(1-r^2\right)\left(1-r^2\right)}{\left(1-r^2\right)^2}}{1+\frac{4r^2}{\left(1-r^2{}_1\right)^2}\cdot\sin^2\left(L\frac{2\pi}{\lambda}\right)} = \frac{1}{1+C_F\cdot\sin^2\left(L\frac{2\pi}{\lambda}\right)} \tag{12.19}$$

where the parameter $C_F = \frac{4r^2}{\left(1-r^2\right)^2}$ is the coefficient of finesse for the F-P. This simple formula shows that the transmittance of symmetric F-Ps is unity at the resonant wavelengths.

Based on Eq. 12.19, we find a simple expression for the full-width-at-half-maximum of the transmittance peaks of symmetric F-Ps:

$$\frac{1}{1+C_F\cdot\sin^2\left(\frac{\Delta L_{FWHM}}{2}\cdot\frac{2\pi}{\lambda}\right)} = \frac{1}{2} \Rightarrow \Delta L_{FWHM} = \frac{\lambda}{\pi}\sin^{-1}\left(\frac{1}{\sqrt{C_F}}\right) \tag{12.20}$$

The Finesse[5] of a F-P, or other optical device with a periodic transmittance function, is defined as the ratio of the period to the FWHM of the transmittance peak:

$$F \equiv \frac{\Delta L_{period}}{\Delta L_{FWHM}} = \frac{\frac{\lambda}{2}}{\frac{\lambda}{\pi}\cdot\sin^{-1}\left(1/\sqrt{C_F}\right)} = \frac{\pi}{2\cdot\sin^{-1}\left(1/\sqrt{C_F}\right)} \tag{12.21}$$

The Finesse and coefficient of finesse are typically only used for F-P with high mirror reflectivities, so $\sin^{-1}\left(\frac{1}{\sqrt{C_F}}\right) \approx \frac{1}{\sqrt{C_F}}$, and

$$F \approx \frac{\pi\cdot\sqrt{C_F}}{2} = \frac{\pi\cdot r}{\left(1-r^2\right)} \tag{12.22}$$

[5] The related quantities of Finesse and coefficient of finesse should not be confused.

We see that the Finesse, just like the coefficient of finesse, is a function of the mirror reflectivities and nothing else.

The sensitivity of the F-P transmitted power can now be found by differentiating Eq. 12.19 with respect to the cavity length:

$$\frac{dT}{dL} = \frac{-C_F \cdot 2\sin\left(L\frac{2\pi}{\lambda}\right)\cdot\cos\left(L\frac{2\pi}{\lambda}\right)\cdot\frac{2\pi}{\lambda}}{\left[1+C_F\cdot\sin^2\left(L\frac{2\pi}{\lambda}\right)\right]^2} \tag{12.23}$$

In many situations, the most practical way to operate a F-P displacement sensor is to position the mirrors such that the transmittance is close to 0.5 with no signal applied. This is not the point of maximum sensitivity for all values of the Finnesse, but it is close, so we can write:

$$\left.\frac{dT}{dL}\right|_{\max} \approx \frac{C_F \cdot 2\frac{1}{\sqrt{C_F}}\frac{2\pi}{\lambda}}{[1+1]^2} = \sqrt{C_F}\cdot\frac{\pi}{\lambda} = F\cdot\frac{2}{\lambda} \tag{12.24}$$

Compared to the sensitivity of $2\pi/\lambda$ of the two-beam interferometer given in Eq. 12.5, the sensitivity of a F-P is higher by a factor of F/π[6].

Equation 12.24 shows that the sensitivity of a F-P displacement sensor is only dependent on the finesse, i.e. the mirror reflectivities, and not on the cavity length. This seems somewhat counter intuitive, because the transmission peaks when plotted vs. wavelength or optical frequency becomes narrower as the cavity length increases. It is therefore natural to assume that a longer cavity with narrower transmission peaks will be more sensitive to displacement of one of the mirrors. This is, however, not the case, because in the longer cavity, the same absolute change in cavity length leads to a smaller shift of the position of the transmission peak. The narrower peak and the smaller shift exactly cancel so that the sensitivity stays the same.

This is illustrated in Fig. 12.11 that shows plots of Eq. 12.19 for different F-P cavity lengths. The coefficient of finesse is 89.75 (corresponding to mirror reflectivities of 0.9) for both cavities, but the resonator represented by three narrow peaks is four times longer than the other. As expected, the longer cavity has much sharper transmittance peaks. It is also much less sensitive to absolute changes in the cavity length as illustrated by the dashed lines. The net effect is that the sensitivity to

[6] Finesse is not usually used to describe the two-beam interferometer, but according to the definition its finesse is 2, so we see that even though Eq. 12.24 does not give the exact value for the sensitivity of the two-beam interferometer, it is close.

cavity length changes is independent of the cavity length[7], as predicted by Eq. 12.24.

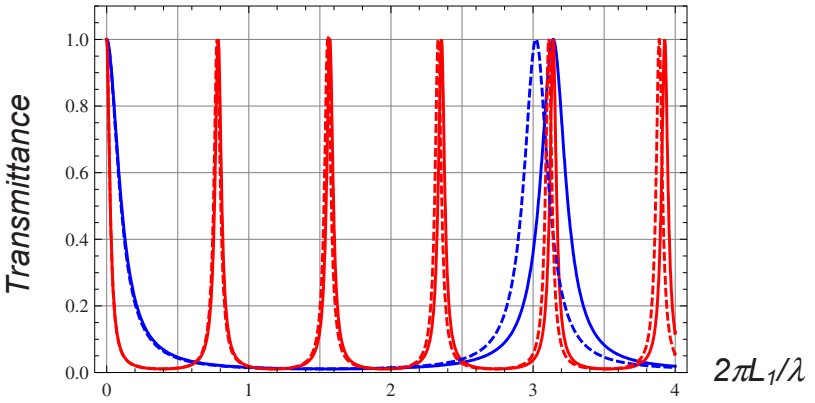


Figure 12.11 *Comparison of transmittance, and variations in transmittance, through symmetric Fabry-Perot interferometers of different lengths, but identical field reflectivities (r_1=r_2=0.9). The solid line with broader transmittance peaks represents a cavity of length L_1, and the line with narrower peaks represents a four times longer cavity. The dashed lines shows the transmittance after the mirrors are moved apart by the same distance $\Delta L=0.04 \cdot L_1$. The broader transmittance peaks of the shorter F-P translate four times further in response to the change in the cavity length. The net effect is that the sensitivity to changes in the cavity length is the same for the two F-Ps.*

The independence of F-P sensitivity on cavity length is significant for MEMS implementations. In MEMS it is most often advantageous to scale devices down to the smallest sizes, over which we have sufficient dimensional control. Equation 12.24 tells us that we don't have to sacrifice sensitivity when scaling the lengths of F-Ps down to dimensions that are practical for MEMS.

12.2.5 Effect of Apertures in Interferometers

In the preceding discussions, we have been mostly concerned with the length of F-P resonators. Practical MEMS devices must be small in all dimensions, however,

[7] The reader might ask that if the sensitivity of the interferometer is independent of its length, then what is the reason for the long lengths of many high-precision interferometers? The answer is that long interferometers are very good at picking up distributed displacements that accumulate over long distances. For measurements of phenomena of that type, long-arm interferometers are clearly superior to short cavities.

so we must now ask the question of how the transverse dimensions influence the characteristics of interferometers. Specifically, we need to know what the limits on scaling of transversal dimensions are, and what penalties, if any, we pay when we scale interferometers to smaller transversal dimensions.

Reducing the transverse dimensions of optical devices means that the optical beams must be focused or collimated to smaller sizes. We know from diffraction theory that smaller beam cross sections lead to more rapid divergence of the beams, so the required length of the optical beams within the device sets hard limits on how tightly the beams can be focused. This is discussed at length in Chapter 4 that covers the basics of diffraction theory and Gaussian Beam propagation, and Chapters 7 and 8 that describe the optimization of longitudinal and transverse dimensions of microscanners and fiber-optical switches.

These same considerations discussed in Chapters 4, 7, and 8 are also valid for interferometers, but in addition there are subtle phase effects associated with focusing that must be taken into account under certain conditions. To see how, we start by considering the mathematical description of the field of a fundamental Gaussian beam:

$$u = \frac{\omega_0}{\omega} \exp\left[-j\left(\phi + \frac{k}{2} r^2 \left(\frac{1}{R} - j\frac{\lambda}{\pi \cdot \omega^2} \right) \right) \right] \tag{12.25}$$

where $k=2\pi/\lambda$ is the wavevector, r is the transversal coordinate, ω is the beam radius, ω_0 is the beam radius at the waist, and R is the radius of curvature. This equation shows that the fundamental Gaussian has an extra phase shift in addition to the well-known $2\pi z/\lambda$ of a plane wave. The extra phase delay, called the Gouy phase, is given by:

$$\phi = \arctan\left(\frac{\lambda \cdot z}{\pi \cdot \omega_0^2} \right) = \arctan\left(\frac{z}{\pi \cdot \omega_0^2 / \lambda} \right) = \arctan\left(\frac{z}{z_R} \right) \tag{12.26}$$

where

$$z_R = \frac{\pi \cdot \omega_0^2}{\lambda} \tag{12.27}$$

is the Rayleigh length.

The Gouy phase shift ϕ is plotted in Fig. 12.12, showing that the fundamental Gaussian has an extra phase shift of π radians over a distance of about 10 Rayleigh lengths around the beam waist. Over this same propagation distance, the standard plane wave phase shift is:

$$\frac{2\pi \cdot 10 z_R}{\lambda} = \frac{2\pi \cdot 10 \cdot \pi \cdot \omega_0^2}{\lambda^2} \approx 200 \frac{\omega_0^2}{\lambda^2} \tag{12.28}$$

We see that the contribution of the Gouy phase to the overall phase shift through focus is less than 1% for even the most tightly focused beams, and much smaller for typical collimated beams. In two-beam and other low-Finesse interferometers, the Gouy phase therefore has but a small effect on measurement calibration, and can be ignored in most practical situations.

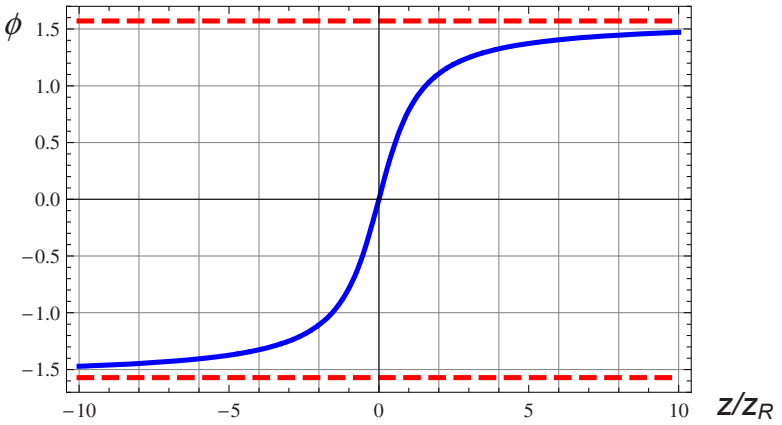


Figure 12.12 Extra phase shift due to focusing of a fundamental Gaussian plotted against the Rayleigh length of the beam. Over a distance of about 10 Rayleigh lengths around focus, there is an extra π phase shift.

In high-Finesse optical resonators, on the other hand, the situation is more complex. In a Fabry-Perot interferometer with high finesse, the optical field bounces back and forth between the mirrors for a number of times that essentially equals the Finesse to create very sharp peaks in the transmittance spectrum. The effect of the Gouy phase therefore accumulates over many bounces and becomes significant.

It should be noted that the Gouy phase can be ignored in the analysis of almost all MEMS interferometers that have been implemented to date. The reason is that it is difficult to create miniaturized, high-Finesse optical resonators using traditional optical MEMS technologies. This is so because both high-reflectivity mirrors and stable resonators are difficult to create in MEMS technology. MEMS interferometers have therefore either been low Finesse or relatively large. In either case the Gouy phase is of little importance.

12.3 Optical Lever

In the introduction to this chapter we described and compared the optical lever and the optical interferometer as displacement sensors in AFMs. Now we will make the comparison more quantitative by deriving an expression for the sensitivity of the optical lever and compare to the formulas we have found for interferometers. We start the derivation by considering a Gaussian beam on a detector of finite size as shown in Fig. 12.13.

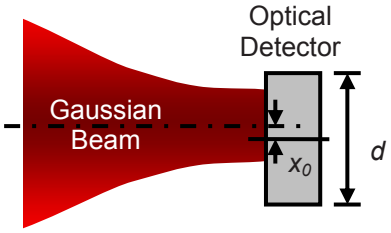


Figure 12.13 *Illustration of a Gaussian beam incident on a photodetector. The bell-shaped form of the Gaussian and the finite size of the detector make the set-up sensitive to the position of the beam. If the beam is centered on the detector (x_0=0), the detected signal is optimized. Any offset for the centered position is detected as a reduction in the received optical power.*

The fact that the Gaussian has a bell shaped distribution of optical power combined with the finite size of the detector makes this simple set-up an optical position sensor. The detector will measure the maximum optical power when the beam is centered, and any deviation from the center position will lead to a corresponding reduction in the measured optical power. In the figure, the Gaussian beam under fills the detector, but in the derivations we will not be making any assumption about the relative sizes of the beam and the detector.

The optical power received by the optical detector is given by:

$$P_D = \frac{P_{in}}{\omega\sqrt{\pi}} \int_{-d/2}^{d/2} e^{-\frac{(x-x_0)^2}{\omega^2}} dx \tag{12.29}$$

where the origin of the x-axis is at the center of the detector, and the Gaussian beam of beam radius, ω, is centered at x_0. We want to measure how the received power varies with the position of the beam, so we take the derivative of this expression with respect to x_0:

$$\frac{d\left(P_D/P_{in}\right)}{dx_0}=\frac{1}{\omega\cdot\sqrt{\pi}}\left(e^{-\frac{(-d/2-x_0)^2}{\omega^2}}-e^{-\frac{(d/2-x_0)^2}{\omega^2}}\right) \tag{12.30}$$

When the detector size is larger than the beam diameter ($d>2\omega$), we can ignore the first part of the sum in Eq. 12.30. The expression therefore has its maximum magnitude value of approximately $\left(\omega\cdot\sqrt{\pi}\right)^{-1}$ close to the points $x_0=\pm d/2$. In other words, the maximum sensitivity is obtained when the beam is centered at the edge of the detector.

The maximum sensitivity is inversely proportional to the beam radius. To achieve good accuracy in the measurements of the beam position, we therefore have to focus the beam tightly on the detector. This is what we would have guessed; the smaller the beam radius, the more precisely we can determine the location of the beam.

Due the fact that the sensitivity attains its maximum value when the beam is at the edge of the photodiode, the preferred set-up is to use two adjacent detectors with the nominal beam position at the dividing line between them as shown in Fig. 12.14. Here the beam is nominally centered at the dividing line between the two detectors so that each detector measures the position of the Gaussian beam with maximum sensitivity.

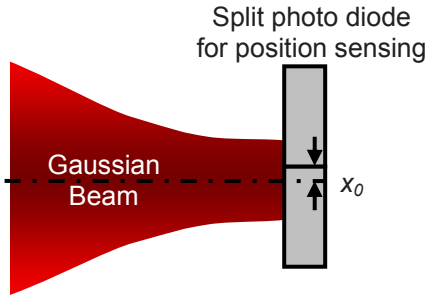


Figure 12.14 *Preferred set-up of optical position sensor for the optical lever with a split photodiode. The optical beam is nominally centered on the dividing line between to adjacent photodiodes that make up the position sensitive detector. Each photodiode is then measuring the beam position with optimum sensitivity and the total received power is very close to constant when the detectors are larger than the beam radius.*

By subtracting the signal from the two photo diodes, we obtain a measured signal that is zero for the nominal beam position and has positive values for positive displacement and negative values for negative displacement:

$$P_{PSD} = \frac{1}{\omega\sqrt{\pi}}\left(\int_0^{d/2} P_{in} \cdot e^{-\frac{(x-x_0)^2}{\omega^2}} dx - \int_{-d/2}^{0} P_{in} \cdot e^{-\frac{(x-x_0)^2}{\omega^2}} dx \right) \tag{12.31}$$

The derivative of this expression with respect to x_0 gives us the sensitivity

$$\begin{aligned} \frac{d\left(P_{PSD}/P_{in}\right)}{dx_0} &= \frac{1}{\sqrt{\pi}\cdot\omega}\left(2e^{-\frac{x_0^2}{\omega^2}} - e^{-\frac{(d/2-x_0)^2}{\omega^2}} - e^{-\frac{(-d/2-x_0)^2}{\omega^2}} \right) \\ &\approx \frac{2}{\sqrt{\pi}\cdot\omega}\cdot e^{-\frac{x_0^2}{\omega^2}} \end{aligned} \tag{12.32}$$

The approximation is valid when the detector size is larger than the beam diameter ($d>2\omega$).

The maximum magnitude of the expression is then $2/\left(\omega\cdot\sqrt{\pi}\right)$ for the nominal beam position ($x_0=0$). The maximum sensitivity is twice that of the single detector. This is of course expected because the maximum sensitivity is obtained with the beam is centered at the edge of each detector.

12.3.1 Displacement and Angle Sensitivity of the Optical Lever

In the geometry of Fig. 12.1, the vertical displacement (ΔL) of the cantilever is related to the beam displacement (Δx_0) on the PSD as

$$\Delta x_0 = \Delta L \cdot \frac{\sin 2\alpha}{\cos\alpha} = \Delta L \cdot 2\sin\alpha \tag{12.33}$$

where α is the incident angle on the cantilever. This expression is valid when the optical beam is at normal incidence on the photodetector as we assumed in the preceding section.

The maximum position sensitivity of the cantilever is then

$$\frac{d\left(P_{PSD}/P_{in}\right)}{dL} = \frac{d\left(P_{PSD}/P_{in}\right)}{dx_0}\frac{dx_0}{dL} = \frac{4\sin\alpha}{\sqrt{\pi}\cdot\omega} \tag{12.34}$$

The incident angle takes values between *0* and $\pi/2$, so the ratio $\Delta x/\Delta L$ has its maximum value of 2 at $\alpha=\pi/2$. Grazing incidence is, however, not very practical, so a more typical value is $\Delta x/\Delta L = \sqrt{2}$, which appears at $\alpha=\pi/4$.

Comparing the maximum fractional sensitivity of the lever $\left(\frac{4\sin\alpha}{\sqrt{\pi}\cdot\omega}\right)$ to that of the two-beam interferometer ($2\pi/\lambda$ – Eq. 12.5), we see that the fundamental limit on the sensitivity of the two measurement principles are not that different. We can make the beam radius on the order of half a wavelength, but not much smaller. The best sensitivity we can get from the optical lever is therefore $8\sin\alpha/\left(\lambda\sqrt{\pi}\right)$, which is only marginally smaller than the maximum sensitivity of the two-beam interferometer.

The optical lever is more often used to detect angle variation than displacement. The beam position on the PSD is related to angle variation as

$$\Delta x_0 = 2\Delta\alpha \cdot z \tag{12.35}$$

where z is the distance from the cantilever to the PSD. The maximum angular sensitivity of the cantilever is then

$$\frac{d\left(P_{PSD}/P_{in}\right)}{d\alpha} = \frac{d\left(P_{PSD}/P_{in}\right)}{dx_0}\frac{dx_0}{d\alpha} = \frac{2\cdot 2z}{\sqrt{\pi}\cdot\omega} \approx \frac{4\pi\cdot\omega_0}{\sqrt{\pi}\cdot\lambda} \tag{12.36}$$

where we have used the approximation $\omega = \frac{\lambda\cdot z}{\pi\cdot\omega_0}$, which is valid in the far field (see Chapter 4). The expression shows that once we are in the far field, further increase of the cantilever-PSD separation does not increase the angular sensitivity, because the beam size on the PSD increases linearly with distance in the far-field regime.

12.3.2 Grating Optical Lever

An optical lever with a very sharply focused beam is not useful in most situations. When focused to its smallest possible radius, the optical beam diffracts rapidly, which means that the working distance, i.e. the distance from the reflector that deflects the beam to the detector, must be very short. This puts restrictions on the design of systems with tightly focused beams, so most optical levers have sensitivity far less than the theoretical limit. In practice, the optical lever is therefore much less sensitive than optical interferometers.

As for the interferometer, which can be made more sensitive by using a high-finesse resonator, the optical lever can also be improved. One method for increas-

ing the sensitivity is to use a grating as shown in Fig. 12.15. Diffraction from the grating sets up an output beam at an angle Ψ_o in response to incident light at the angle Ψ_i, thereby enhancing the scan angle of the optical beam over that of the standard cantilever.

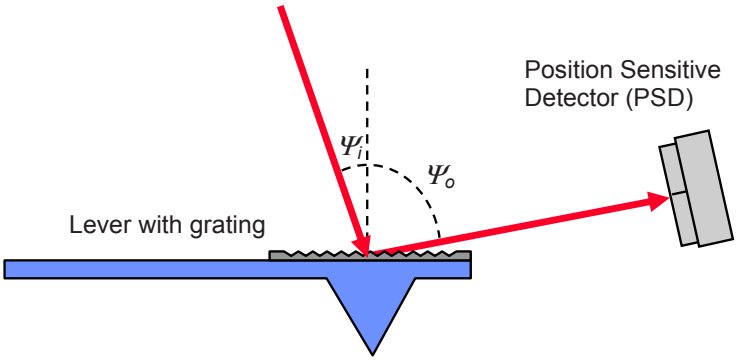


Figure 12.15 Optical lever with grating for sensitivity enhancement.

To analyze the scan angle enhancement, we start with the grating equation

$$p = \sin\Psi_i - \sin\Psi_o = \frac{\lambda}{\Lambda} \Rightarrow \Psi_i \approx \sin\Psi_o - \frac{\lambda}{\Lambda} \tag{12.37}$$

that shows that the sensitivity of the grating lever is increased by the factor

$$\frac{d\Psi_o}{d\Psi_i} = \frac{\cos\Psi_i}{\cos\Psi_o} \approx \frac{1}{\cos\Psi_o} \tag{12.38}$$

Unfortunately, the diffracted beam radius is decreased by the same factor, so the fundamental sensitivity of the grating lever is identical to that of the standard optical lever. Nevertheless, the grating lever is a very practical device, because it allows us to shrink the beam size of the outgoing beam without sacrificing sensitivity. This means that we can place the position sensitive detector closer to the lever, so that the size of the overall system is reduced.

12.4 Sources of Noise in Displacement Measurements

The ultimate limitation on any measurement system is noise, so we start our treatment of measurement accuracy by modeling the noise sources that are present in all optical systems. Figure 12.16 shows how shot noise, thermal noise, and Relative Intensity Noise (RIN) of the light source are added to the signal in a generic optical detection system. The shot noise and RIN are carried by the optical beam, while the thermal noise is a function of the dissipation in the photodetector circuit. In addition to these noise sources there is also added noise, often ex-

pressed in terms of a noise figure, of the amplifier stage that follows the detection, and *1/f* noise of various origins.

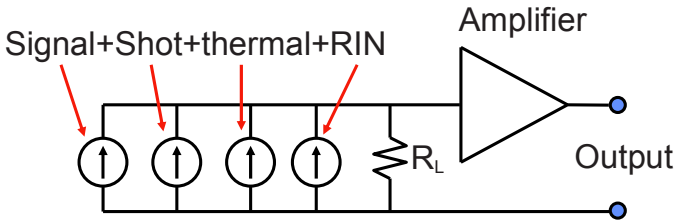


Figure 12.16 Signal and noise in optical detection systems. The optical signal carries shot noise and Relative Intensity Noise (RIN), and the input resistor (R_L) of the receiver adds thermal noise. The amplifier is also at a finite temperature and adds thermal noise, often characterized by noise figure or noise temperature.

12.4.1 Thermal Noise

Thermal Noise, also called Johnson noise or Nyquist noise, is present in all dissipative elements, including electrical resistors, where thermal noise is caused by the thermal motion of the charge carriers (electrons). Real resistors can be modeled as ideal (noise less) resistors in parallel with noise current sources as shown in Fig. 12.17.

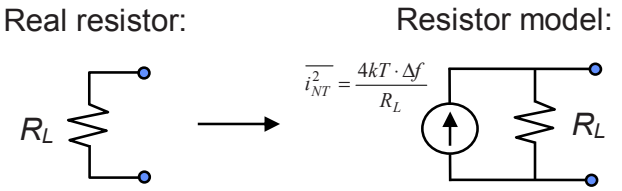


Figure 12.17 A real resistor can be modeled as an ideal (noiseless) resistor in parallel with a thermal-noise current source.

The noise current is given by

$$\overline{i_{NT}^2} = \frac{4kT \cdot \Delta f}{R_L} \tag{12.39}$$

where k is the Bolzmann constant, T is the absolute temperature, Δf is the electrical bandwidth of the detector, and R_L is the resistance value. The bandwidth is typically equal to the information bandwidth, but in some cases the detector bandwidth must be twice the information bandwidth. The thermal noise spectrum is uniform (white noise) up to about *10 GHz*.

12.4.2 Shot Noise

Shot Noise is caused by the quantization of the optical field and the finite value of the electrical charge carriers. All photodetectors (photo multipliers, photodiodes, photoresistors) exhibit signal degradation due to shot noise. As for thermal noise, we can model shot noise in detection circuits as an additive noise current source as shown in Fig. 12.18. The noise current is given by:

$$\overline{i_{NS}^2} = 2q\left(\overline{i_s} + I_D\right)\cdot \Delta f \tag{12.40}$$

where q is the electron charge, $\overline{i_s}$ is the average photocurrent, and I_D is the dark current

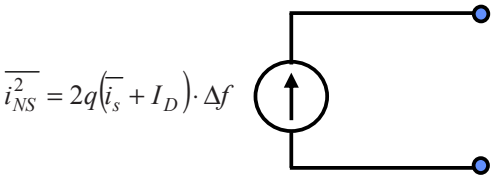


Figure 12.18 Shot noise model.

12.4.3 Relative Intensity Noise

Most lasers and other light sources generate excess noise, called Relative Intensity Noise (RIN), beyond the minimum quantization (or shot) noise. The noise is often higher close to the relaxation oscillation frequency of the laser, so RIN is more damaging at higher frequencies. For lasers, the absolute noise level typically peaks at threshold and stays constant as the output power is increased, so that RIN goes down as the laser is driven harder. As shown in Fig. 12.19, laser RIN can be modeled as a noise current source given by

$$\overline{i_{NL}^2} = RIN \cdot \overline{i_s^2} \cdot \Delta f \tag{12.41}$$

where *RIN* is the laser noise normalized to a one Hz bandwidth

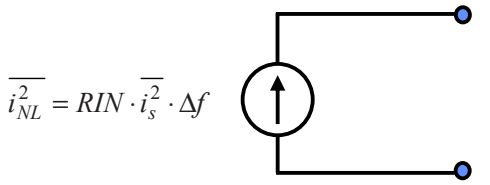


Figure 12.19 Relative Intensity Noise.

In addition to the fundamental noise sources of the optical detection process, noise is also added in the amplifiers. Amplification and the associated noise characterized in terms of the noise figure or noise temperature, is an integral part of any measurement system, and the challenges of designing good amplifiers are determined by the nature of the signal, so the relative ease or difficulty of amplification is important to consider when comparing measurement technologies.

12.5 Signal-to-Noise Ratio

The signal from an optical sensor can take many forms. It may be amplitude modulated, phase modulated, or have some more complex modulation format, e.g. pulse-width modulation. At some point the optical signal will be converted to an electrical signal in a square-law photodetector, i.e. an optical detector that is sensitive to the optical power, but not to the optical phase.

We will consider detectors that can be characterized by their *responsivity*, which is defined as the ratio of the output photo current produced by the detector to input optical power on the detector:

$$\rho = \frac{i}{P_{inc}} \tag{12.42}$$

The responsivity has the units of Ampere/Watt and it is the most important specification of most commonly used photodetectors, including semiconductor diodes (photo diodes) and Photon Multiplier Tubes (PMTs).

When a measurand affects an optical sensor, then it produces, directly or indirectly, a signal that can be expressed in terms of a change in optical power, ΔP, received by the photodetector. The change in optical power leads to a change in photocurrent according to Eq. 12.42. The optical power signal may take the form of a static offset or a harmonic power variation.

In the latter case, it is useful to introduce the modulation index, defined here as the ratio of the Root-Mean-Square (*RMS*) to the average of the optical power:

$$m \equiv \frac{P_{rms}}{\overline{P}} = \frac{\Delta P}{\overline{P}} \tag{12.43}$$

This definition is illustrated in Fig. 12.20.

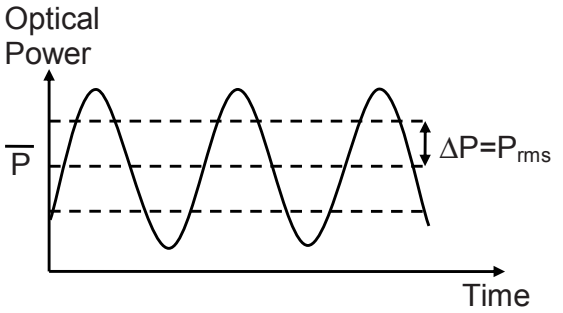


Figure 12.20 The modulation index of an optical signal is defined as ratio of the variation (RMS for harmonic signals) to the average of the optical power.

With this definition of modulation index combined with the expressions for noise (Eq. 12.39-12.41), we can write the signal-to-noise ratio of the electrical power of the detector as

$$\frac{S}{N}\cdot\Delta f=\frac{m^2\cdot\bar{i}^2R_L}{4kT+2q\left(\bar{i}+I_D\right)\cdot R_L+RIN\cdot\bar{i}^2R_L} \tag{12.44}$$

where $\bar{i}$ is the average photocurrent, I_D is the dark current produced by the photodetector, and Δf is the bandwidth of the detection circuit. In terms of the optical power this becomes

$$\frac{S}{N}\cdot\Delta f=\frac{\left(\rho\cdot\Delta P\right)^2R_L}{4kT+2q\left(\rho\cdot\overline{P}+I_D\right)\cdot R_L+RIN\cdot\left(\rho\cdot\overline{P}\right)^2R_L} \tag{12.45}$$

where $\overline{P}$ is the average power and ΔP is the power variation caused by the measurement. In these expressions, we have pulled the common factor Δf out of the denominators to emphasize the bandwidth dependence.

Figure 12.21 shows schematically how the Signal-to-Noise ratio depends on received electrical power. At low electrical powers, the noise is dominated by thermal noise, and the *S/N* increases linearly with electrical power. In this regime the *S/N* is proportional to the receiver input resistance.

As the electrical power is increased, the shot noise and RIN becomes more pronounced. If the receiver resistance is large enough, we enter a regime where shot noise is dominant. Here the *S/N* increases as the square-root of the electrical power.

As we increase the electrical power still further, the RIN finally becomes dominant. At this level the *S/N* no longer increases with electrical power. We have reached the highest *S/N* possible for the given light source.

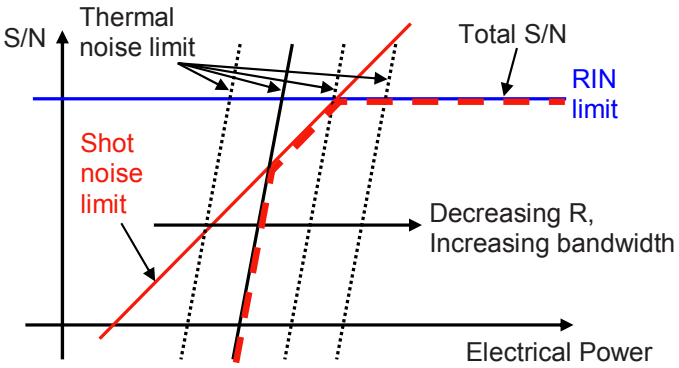


Figure 12.21 *Conceptual illustration of the dependence of Signal-to-Noise ratio (logarithmic scale) on received electrical power in optical detection circuits. As the electrical power is increased, the S/N goes from being limited by thermal noise, to shot noise, and finally to RIN.*

Figure 12.21 shows that to get the ultimate *S/N* in our measurements, we must operate at a power level where *RIN* is dominant, or in other words, we must make sure we have enough optical power at the receiver that shot noise and thermal noise can be neglected. This is of course not always practical. In many systems we wish to reduce the received power as much as possible. For example in long-haul fiber optic communication, it is economical to have the longest possible distance between detection circuits. In such systems the optical signals are severely attenuated by fiber loss by the time they arrive at the detectors, and the *S/N* is therefore limited by thermal noise of the detector[8].

In sensor systems we would like to be in the *RIN* limited regime to get the best possible data from our sensors. The high optical powers that are required are, however, not always practical. In some systems we might be limited by the high cost of high-power sources, and in other the measurement system itself may limit the optical power that can be applied before damage thresholds are reached.

12.5.1 Noise Equivalent Power

The sensitivity of an optical detector is often expressed in terms of its Noise-Equivalent-Power (NEP). The NEP is defined as the signal power required to provide a unity *S/N* ratio in a 1 Hz bandwidth. These choices of bandwidth and

[8] In the days before practical optical amplifiers this was the universal standard of operation, but now an increasing number of systems employ optical amplifiers on the receiving end.

S/N level are of course arbitrary, but they result in a well-defined figure of merit for comparison of different detector technologies, in addition to giving a useful intuitive sense of detector resolution.

Noise-Equivalent-Power is therefore a common specification used to describe the resolution of optical detectors. In such specifications, only the noise originating with the detector, i.e. the thermal noise[9], is considered. Here we extend the NEP concept to include all sources of noise in the system. Based on Eq. 12.45 we find the following expression for *NEP* in the presence of thermal noise, shot noise, and *RIN*

$$\frac{S}{N}=1\Rightarrow\frac{\Delta P_{\min}}{\sqrt{\Delta f}}=NEP=\frac{1}{\rho}\sqrt{\frac{4kT}{R_L}+2q\left(\rho\cdot\overline{P}+I_D\right)+RIN\cdot\left(\rho\cdot\overline{P}\right)^2} \tag{12.46}$$

Note that the units for NEP is $W/\sqrt{Hz}$. This may seem mysterious when NEP specifications are first encountered, but we see that it follows straightforwardly from the fact that NEP is defined in terms of optical power.

12.6 Detection Limits in Displacement Measurements

Optical displacement sensors face competition from a number of other displacement measurement technologies, including capacitive, piezoresistive, tunneling, and other sensor systems. In this section we will compare these technologies to understand their relative strengths and weaknesses, particularly for differential displacement measurements on the chip scale. We will use the concepts of Noise-Equivalent-Power and dynamic range to quantify the comparisons. The goal is to gain the perspective necessary to make good system-design choices.

12.6.1 Resolution of Optical Interferometers

Interferometric position sensors are fundamentally limited by the same noise sources that limit the capacity of optical communication and other optical systems, i.e. thermal noise, shot noise, and Laser Relative Intensity Noise (RIN). We combine this with our knowledge of the sensitivity of the interferometer to define a Noise-Equivalent-Displacement in units of $W/\sqrt{Hz}$.

By rewriting Eq. 12.5, we can relate the minimum resolvable deflection to the minimum resolvable reflected power

[9] Other noise sources, e.g. 1/f noise that we don't consider here, will of course also be included in the *measured* NEP specified by photodetector vendors.

$$\Delta L_{\min} = \frac{\lambda}{4\pi}\frac{\Delta P_{\min}}{\overline{P}} \tag{12.47}$$

where $\overline{P}$ is the average power, i.e. $\overline{P} = P_{in}/2$. Combining Eqs. 12.46 and 12.47, we find the following expression for the Noise-Equivalent-Displacement of a two-beam interferometer

$$\frac{\Delta L_{\min}}{\sqrt{\Delta f}} = \frac{\lambda}{4\pi}\cdot\sqrt{\frac{4kT}{R_L\left(\rho\cdot\overline{P}\right)^2} + 2q\left(\frac{1}{\rho\cdot\overline{P}} + \frac{I_D}{\left(\rho\cdot\overline{P}\right)^2}\right) + RIN} \tag{12.48}$$

where again ρ is the receiver responsivity, R_L is the photo detector load resistor, $\overline{P}$ is the average received optical power, I_D is the dark current, and *RIN* is the relative-intensity-noise of the optical source.

As expected, we find that the Noise-Equivalent-Displacement of the interferometer is inversely proportional to average power in the thermal-noise limit, inversely proportional to the square root of average power in the shot-noise limit, and independent of power in the *RIN* limit. For a *RIN* limited system with a *RIN* of 10^{-14}, Eq. 12.48 evaluates to approximately 10^{-14} m/Hz$^{0.5}$ at a wavelength of 1 um, in good agreement with experimental observations [4]. A low-noise laser with an RIN of 10^{-16} and a wavelength of 500 nm improves the Noise-Equivalent-Displacement to approximately $4\cdot10^{-16}$ m/Hz$^{0.5}$.

We define the dynamic range of a sensor as the ratio of the maximum signal that can be measured to the Noise-Equivalent signal. For a two-beam interferometer, the maximum displacement that can be determined without ambiguity is one quarter of the wavelength, so the dynamic range is

$$\frac{\Delta L_{\max}}{\Delta L_{\min}}\sqrt{\Delta f} = \frac{\pi}{\sqrt{\frac{4kT}{R_L\left(\rho\cdot\overline{P}\right)^2} + 2q\left(\frac{1}{\rho\cdot\overline{P}} + \frac{I_D}{\left(\rho\cdot\overline{P}\right)^2}\right) + RIN}} \tag{12.49}$$

We saw in section 12.2.4 that we can improve the sensitivity by using a high-finesse interferometer. Equation 12.24 shows that the improvement factor is F/π. The Noise-Equivalent-Displacement of an interferometer with a finesse F, is then given by

$$\frac{\Delta L_{\min}}{\sqrt{\Delta f}} = \frac{\lambda}{4\cdot F}\cdot\sqrt{\frac{4kT}{R_L\left(\rho\cdot\overline{P}\right)^2} + 2q\left(\frac{1}{\rho\cdot\overline{P}} + \frac{I_D}{\left(\rho\cdot\overline{P}\right)^2}\right) + RIN} \tag{12.50}$$

The maximum displacement is reduced by the same factor, so the dynamic range is the same as for a simple two-beam interferometer.

12.6.2 Resolution of Optical Levers

The sensitivity of the Optical Lever is expressed by Eq. 12.34. Rewritten, it gives the following relationship between displacement and power

$$\Delta L_{\min} = \frac{\omega \cdot \sqrt{\pi}}{4 \sin \alpha} \frac{\Delta P_{\min}}{\overline{P}} \tag{12.51}$$

where $\overline{P} = P_{in}$ is the average power. Combining Eqs. 12.46 and 12.50, we find the Noise-Equivalent-Displacement of the optical lever

$$\frac{\Delta L_{\min}}{\sqrt{\Delta f}} = \frac{\omega \cdot \sqrt{\pi}}{4 \sin \alpha} \cdot \sqrt{\frac{4kT}{R_L \left(\rho \cdot \overline{P}\right)^2} + 2q\left(\frac{1}{\rho \cdot \overline{P}} + \frac{I_D}{\left(\rho \cdot \overline{P}\right)^2}\right) + RIN} \tag{12.52}$$

The maximum displacement that can be measured is not so well defined. In principle we can make the PSD as wide as we want, and measure very large displacements. Off-center measurements have significantly worse resolution as expressed by the exponential in Eq. 12.32, however, so we set the maximum range equal to the beam radius.

With that definition, we see that the fundamental limits on minimum detectable displacement, Noise-Equivalent-Displacement, and dynamic range for the optical lever are very similar to those of optical interferometers. As pointed out in section 12.3.1, it is, however, practically impossible to achieve optimum performance in optical levers, because that requires that the beam is focused to its minimum beam radius, which again leads to impractically short working distances. In practice, the optical lever has worse Noise-Equivalent-Displacement than interferometers. The dynamic range is, however, the same, so the loss of resolution is accompanied by an increase in the maximum displacement that can be measured.

12.6.3 Resolution of Capacitive Sensors

Consider the capacitive sensor system shown in Fig. 12.22. The signal from the sensor is proportional to the bias voltage (V_S) and to the ratio of a sensing capacitor and a reference capacitor. This simple signal model covers most practical capacitive sensors. After demodulation, the capacitive signal is sent to a preamplifier. The electrical resistance in the circuit contributes thermal noise that degrades the signal-to-noise.

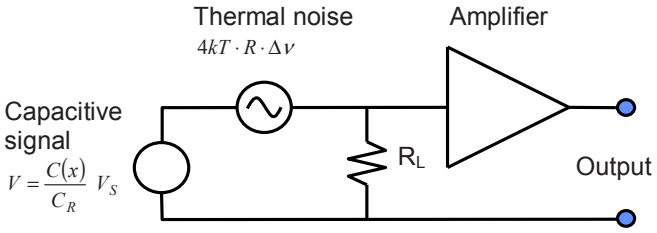


Figure 12.22. Model of signal and noise in capacitive sensor systems.

The thermal noise power is

$$P_{noise} = \frac{v_{rms}^2}{R_L} = i_{rms}^2 \cdot R_L = 4kT \cdot \Delta\nu \tag{12.53}$$

where $\Delta\nu$ is the bandwidth of the sensor system. If it is determined by the *RC(x)* product, the RMS noise voltage becomes:

$$v_{rms}^2 = 4kT \cdot \Delta\nu \cdot R_L = \frac{4kT}{2\pi \cdot C(x)} \tag{12.54}$$

In most practical cases, however, the bandwidth is not determined by the *RC* time constant of the sensor, but by the preamplifier, so we will treat the bandwidth as an independent variable. The signal-to-noise is then

$$\frac{S}{N} = \frac{\left(\frac{C(x)}{C_R} V_S\right)^2}{v_{rms}^2} = \frac{\left(\frac{C(x)}{C_R} V_S\right)^2}{4kT \cdot \Delta\nu \cdot R_L} \tag{12.55}$$

We now set the signal–to-noise ratio to unity to find the Noise-Equivalent Capacitance

$$\frac{S}{N} = 1 \Rightarrow \frac{\Delta C_{\min}(x)}{\sqrt{\Delta\nu}} = \frac{C_R \sqrt{4kT \cdot R_L}}{V_S} \tag{12.56}$$

To proceed we assume a gap-closing parallel-plate capacitor with area A, gap g_0, and a capacitor-plate spacing x. The capacitance value is then

$$C(x) = \frac{\varepsilon \cdot A}{x} \tag{12.57}$$

This gives us the displacement sensitivity of the capacitance

$$\frac{\partial C(x)}{\partial x} = -\frac{\varepsilon \cdot A}{x^2} = -\frac{C(x)}{x} \Rightarrow \frac{\Delta x}{\Delta C} = \frac{x}{C(x)} \tag{12.58}$$

The Noise-Equivalent-Displacement is

$$\frac{\Delta x_{\min}}{\sqrt{\Delta\nu}} = \frac{\Delta C_{\min}(x)}{\sqrt{\Delta\nu}} \frac{\Delta x_{\min}}{\Delta C_{\min}(x)} = \frac{C_S\sqrt{4kT\cdot R_L}}{V_S} \frac{x}{C(x)} \approx \frac{x\sqrt{4kT\cdot R_L}}{V_S} \tag{12.59}$$

where we assumed that the values of the sensing capacitor and the fixed capacitor are of the order of magnitude. To improve the capacitive sensor, we can increase the voltage and/or decrease the input impedance of the amplifier. The nominal gap (x) determines the maximum deflection, so it should be thought of as a parameter set by the application.

The Noise-Equivalent-Displacement is proportional to the capacitor-plate separation, x, which represent the upper limit on measurable displacement, so the dynamic range is

$$\frac{\Delta x_{\max}}{\Delta x_{\min}}\sqrt{\Delta\nu} \approx \frac{\sqrt{4kT\cdot R_L}}{V_S} \tag{12.60}$$

At room temperature with $x=1$ *um,* $V_S=10$ *V,* and $R_L=100$ *ohms*, the Noise-Equivalent-Displacement evaluates to about 10^{-16} m/Hz$^{0.5}$. In practice it is very hard to achieve the theoretical performance of a capacitive sensor. The ADXL150 from Analog Devices have a RMS position error of $130\cdot10^{-12}$ *m* in a *1,000 Hz* bandwidth. That corresponds to a Noise-Equivalent-Displacement of $4.3\cdot10^{-12}$ *mHz*$^{-0.5}$, with a voltage of *5 V* and a gap of *1.3 um*. [5] This is far from the theoretical limits, but still close to the thermo-mechanical noise floor. Capacitive sensors used for ultrasonic imaging are designed for detection of very small displacements, and get much closer to the theoretical limits.

12.6.4 Resolution of Piezoresistive Sensors

Piezoresistors were among the first MEMS sensor principles to be exploited in research and in commercial settings. The principle is illustrated in Fig. 12.23. A silicon resistor is subject to strain, typically by bending as shown in Fig, 12.23a. In response to the strain, the resistor changes its value, and the resistance change is converted to a signal voltage in a Wheatstone bridge as shown in Fig. 12.23b.

When the Wheatstone bridge is built with piezoresistors that are subject to opposite strains so that their resistance changes have opposite signs as in Fig. 12.23b, then the signal voltage is given by

$$\Delta V = \frac{\Delta R}{R} V_S \tag{12.61}$$

where V_S is the bias voltage on the Wheatstone bridge, R is the nominal resistivity of the bridge resistors, and ΔR is the resistance change.

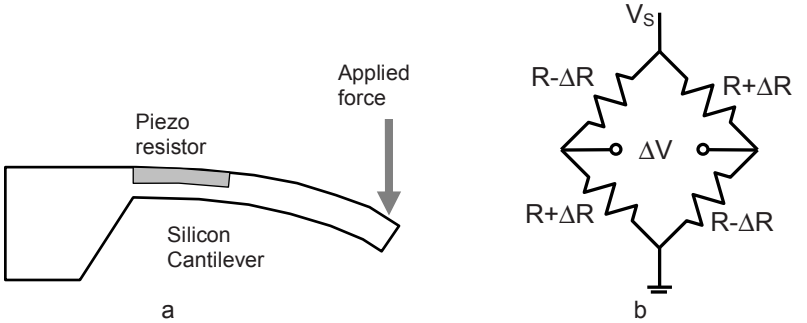


Figure 12.23 Piezoresistive sensor

The signal-to-noise ratio is then

$$\frac{S}{N} = \frac{\left(\frac{\Delta R}{R} V_S\right)^2}{v_{rms}^2} = \frac{\left(\frac{\Delta R}{R} V_S\right)^2}{4kT \cdot \Delta\nu \cdot R} \tag{12.62}$$

Setting the signal-to-noise ratio to unity, we find the Noise-Equivalent-Resistance of the piezoresistive sensor

$$\frac{S}{N} = 1 \Rightarrow \frac{\Delta R_{\min}}{\sqrt{\Delta\nu}} = \frac{R\sqrt{4kT \cdot R}}{V_S} \tag{12.63}$$

Piezoresistors and strain gages can be made in many different materials. In most materials the change in resistance is dominated by the geometrical effect, i.e. the resistance value changes as if the resistivity of the material is unchanged. Typically, the change in cross section can be neglected, so that the relative resistance change is equal to the relative length change, i.e.

$$R + \Delta R = \rho \frac{L + \Delta L}{A + \Delta A} \approx \rho \frac{L + \Delta L}{A} \Rightarrow \frac{\Delta R}{R} = \frac{\Delta L}{L} \tag{12.64}$$

where ρ is the material resistivity, L is the resistor length, and A is the uniform resistor cross section.

Silicon, as opposed to most other materials, has a very strong piezoresistive effect. A correctly oriented silicon piezoresistor achieves a resistance change that is 10 times larger than the value given by the geometrical effect alone. For silicon we can then write

$$\frac{\Delta R}{R} = 10 \frac{\Delta L}{L} \tag{12.65}$$

It follows that the Noise-Equivalent-Displacement can be expressed as

$$\frac{\Delta L_{\min}}{\sqrt{\Delta\nu}} = \frac{L}{10 \cdot R} \frac{\Delta R_{\min}}{\sqrt{\Delta\nu}} = \frac{L}{10 \cdot R} \frac{R \cdot \sqrt{4k_b T \cdot R}}{V_S} \approx \frac{L \cdot \sqrt{4k_b T \cdot R}}{10 \cdot V_S} \tag{12.66}$$

Comparing this equation directly to the Noise-Equivalent-Displacement of capacitive sensors (Eq. 12.59) it seems that the piezoresistor has about an order of magnitude better displacement sensitivity. This is not entirely correct, however, because the resistance values that are used in piezoresistive sensors typically are an order of magnitude higher or more, mitigating much of the difference. More importantly, the dynamic range of the piezoresistor is substantially less than that of capacitive and optical sensors. Silicon, like most other materials, cannot safely be strained to much more than 1%, so the dynamic range of piezoresistors is about one to two orders of magnitude lower than that of capacitive displacement sensors.

Piezoresistors are also difficult to utilize directly as displacement sensors. Typically we must use some kind of leverage to extend the travel. The bending arm of Fig. 12.23 is a good example. Here the motion of the end of the bending arm is much larger than the elongation of the piezoresistor. This extended travel leads to an equally large increase in the Noise-Equivalent-Displacement.

As noted above, however, the theoretical limits are very hard to achieve in any system, so practical considerations are often more important. The simplicity of the electronic interface (the Wheatstone bridge of Fig. 12.23b) makes the piezoresistor the sensor of choice in many applications.

12.6.5 Comparison of Displacement Sensors

The derivations and discussions of the preceding three sections make it clear that no one sensing concept is superior for all applications. Piezoresisitive sensors do not have the dynamic range of capacitive or optical sensors. In practices, this also means that they are not as sensitive. Still they are often preferred because of their simple fabrication and straight-forward integration with electronics.

Comparing capacitive and optical sensors, we have found that there are only small differences in terms of Noise-Equivalent-Displacement and dynamic range. The capacitive sensor can be made more sensitive by increasing the bias voltage (V_S in Eq. 12.59). In applications that can tolerate high bias voltages, i.e. ultrasonic pressure sensors with relatively small and stiff membranes, capacitive sensors are often the right choice. The counter point is low-frequency, highly-sensitive pressure sensors, e.g. microphones and Golay cells. These require very compliant pressure-sensing diaphragms that collapse under large bias voltages, so optical sensors have advantages over capacitive. In general, we can say that stiffer structures favor capacitive sensors, while optical sensors are preferred for compliant

constructions. Optical interferometers also have the advantage that their dynamic range can be vastly extended by lifting the ambiguity of the interferometric response (see Fig. 12.5). This adds complexity, however, and is therefore more common in macroscopic measurement systems than in chip-scale designs.

Optical, capasitive, and piezoelectric sensors account for the majority of MEMS and chip-scale measurement systems, but there are other technologies that find use in niche applications. Electron-tunneling sensors offer the ultimate in position sensing, having about 20 times better sensitivity than capacitive sensors under similar operations constraints (except maximum displacement). Unfortunately, their limited range restricts their use to sensor systems with feedback stabilization. Optical near-field, or photon tunneling, sensors do not have as good sensitivity as electron tunneling, but they have better dynamic range, and represent a very good choice for many miniaturized systems. A special class of photon-tunneling sensors based on Photonic Crystals is described in Chapter 15.

The conclusion is that comparisons between different types of displacement sensors based on their basic characteristics alone are difficult. When choosing between different measurement technologies, the systems designer should use the fundamental limitations outlined in this chapter as a guide, but as important are considerations of the practically achievable sensor properties. Those depend on the application, on environmental constraints, and on various aspects of the implementation including design, architecture, fabrication technology, tolerances, electronics, packaging, and, ultimately, cost.

12.7 Summary of Optical Displacement Sensors

Optical interferometers come in a bewildering array of architectures and implementations, but we show in Chapter 12.2 that their sensitivity to displacement depends only on the wavelength of light and the finesse of the interferometer. Further we show in 12.3 that the fundamental limits of the optical lever are similar to those of interferometers. This somewhat counterintuitive[10] finding allows us a compact description of the displacement sensitivity of all types of traveling-wave optical displacement sensors.

Sensitivity alone cannot predict resolution limits. For that we need the noise models that are developed in 12.4 and 12.5. Armed with these models, we investigate the limits on resolution and dynamic range of optical displacement sensors in 12.5, and compare them to the limits of capacitive and piezoresistive sensors. The sur-

[10] Or maybe not so counterintuitive when we consider the fact that optical-beam focusing, which is what determines the resolution of the optical lever, is limited by wave interference, i.e. the effect that gives interferometers their position sensitivity.

prising conclusion is that the fundamental limits are very similar, and that the best choice of measurement technology for a given system depends on the details of the application and the implementation.

Exercises

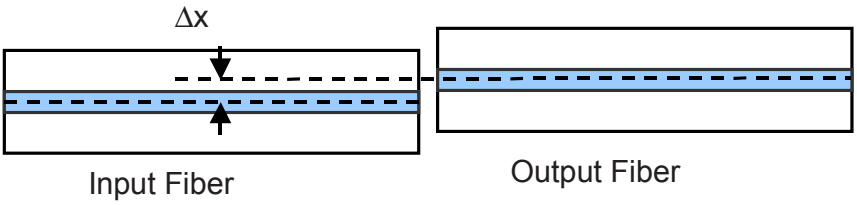
Problem 12.1 - Interferometric Displacement Sensing

You are designing an optical interferometer for distance measurements and you have the choice between two lasers: one at *500 nm* wavelength with a RIN of 10^{-14}, and one at *1,550 nm* wavelength with a RIN of 10^{-16}. Assume that the interferometer will be limited by thermal noise.

Which laser would you use? Explain your answer.

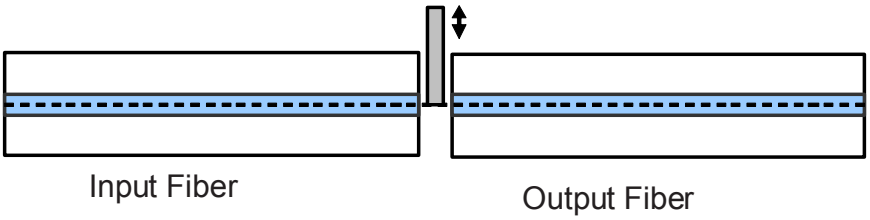
Problem 12.2 - Sensor Design

a. Calculate the maximum sensitivity of the fiber displacement sensor shown below. How should the fiber be designed to get maximum sensitivity?



Displacement sensor based on transmission between identical fibers that are offset laterally (ignore axial offset and reflections from the fiber ends.

The figure below shows another position sensor, in which the position of the beam block is measured by detecting the transmitted light between two perfectly-aligned fibers.

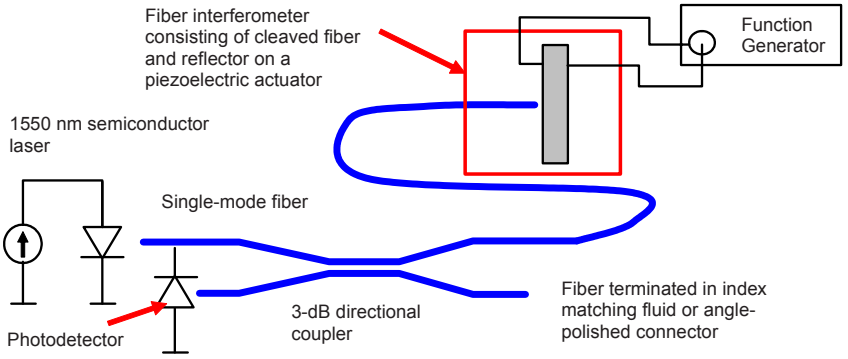


Displacement sensor based on partial blocking of transmission between identical, perfectly-aligned fibers.

b. Calculate the maximum sensitivity for this fiber displacement sensor.
c. How can you extend the range of the sensors? What figure of merit stays constant as you extend the range?
d. Compare the two sensors. What are the advantages and disadvantages of each of the two principles?

Problem 12.3 - Fiber Interferometer

Consider a fiber interferometer as shown below.



Interferometric fiber optic displacement sensor at 1.55 μm wavelength.

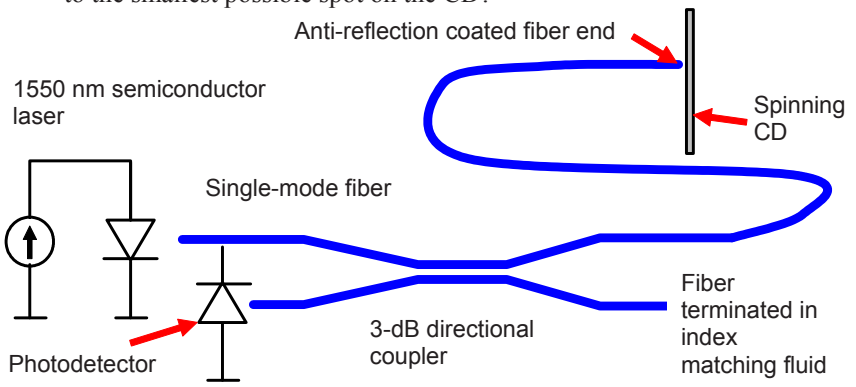
a. What is the optimum distance from the fiber to the mirror? Explain your reasoning.
b. Plot the back-coupled power as a function of distance over the range $L=L_{opt} \pm 5$ *micron*. Include the effects of both the phase shift and the displacement dependence of the reflection from the transducer back into the fiber.
c. Repeat the plot in b) for a wavelength of *1550nm+1nm*. What conclusion can you draw from this plot about the required line width of the laser?
d. In this experiment we may simply use an uncoated fiber, so the reflection from the fiber end is fixed at 4%. We can change the reflections by applying a (multi-layer) coating. What is the optimum reflectivity from the fiber end? Explain your reasoning.

Problem 12.4 - Reading a CD with a Single-Mode Fiber at 1.55 μm Wavelength

A fiber is brought to within 10 microns of a spinning CD, and the backcoupled light is detected as shown below.

a. What kind of signal do you expect to observe if you were to send the signal from the detector to an oscilloscope? Explain your reasoning.

b. How would the observed signal change if we were to focus the fiber output to the smallest possible spot on the CD?



Reading CDs with a standard single mode fiber at 1.55 μm wavelength.

Problem 12.5 - Optical Lever

In Section 12.3 we claimed to have optimized the optical lever when we maximized the differential with respect to displacement of the ratio of the detected optical power to the incident power.

a. Under what conditions is this figure of merit (differential with respect to displacement of the detected optical power relative to the incident power) the one that should be used for optimization?
b. If we are thermal noise-limited, what should be the figure of merit?
c. Is there a set of circumstances, and a corresponding figure of merit, that will change the basic conclusion that the maximum intensity point should be at the edge of the detector? If so, how should a split photo diode be designed for operation under these circumstances?

References:

1 See for example M. Born and E. Wolf, "Principles of Optics", 7th (expanded) edition, Cambridge University Press, 1999.
2 P. Hariharan, "Basics of Interferometry", Second edition, Academic Press, 2006.
3 http://en.wikipedia.org/wiki/LIGO
4 J.E. Bowers, "Fiber optic sensor for surface acoustic waves," Appl. Phys. Lett. **41**, 231 (1982).
5 S.D. Senturia, "Microsystem Design", Kluwer Academic Publishers, 2001, Chapter 19.

13: Micro-Optical Filters

13.1 Introduction to Micro-Optical Filters

Optical wavelength control is critical to the operation of many optical systems for communication, imaging, and measurement. Wavelength Division Multiplexed (WDM) communication systems require sources, (de)multiplexers, dispersion compensators, channel monitors, and receivers with accurate center wavelengths and bandwidths. Two-photon and other non-linear microscopy techniques use spectral filters to separate out the desired frequency components for imaging. Optical sensors depend on well-calibrated, wavelength-stabilized sources and filters for accurate and repeatable measurements. In the field of femto-second lasers, controlling the spectral phase is necessary for pulse shaping and diagnostics.

The Optical Microsystems we have described, i.e. scanners, mirror arrays, and micrograting, provide building blocks for wavelength *tunable* optical filters and spectrometers. This wavelength agility adds much-needed flexibility to wavelength control. Combined with the advantages of miniaturization, integration and parallel processing, it also enables systems with better performance and lower cost.

This chapter introduces the basic principles and architectures used in Optical MEMS filters and spectrometers. The field of optical filters is large and well developed. A comprehensive treatment would require a whole book, so in the interest of saving space, the chapter is focused on conceptual descriptions and on case studies of some of the most successful MEMS implementations. For the rich mathematical description of optical filtering, the reader is referred to the extensive bibliography of this chapter.

Optical filtering is traditionally done either by selective absorption or by interference as in Fabry-Perot and grating filters. Absorptive filters are bulky and difficult to tune, so our treatment is focused exclusively on filter concepts based on interference. We start by considering the filter properties of the Fabry-Perot interferometer and the waveguide ring resonator that were first introduced in Chapter 6. These are both examples of amplitude filters, although their operation and tuning are based on phase control.

We then turn our attention to filters that control spectral phase for dispersion compensation and pulse shaping. We find that these devices are simpler to implement, because, unlike amplitude filters, they do not require extreme out-of-band rejection. Optical spectrometers are treated next. The flexibility of optical MEMS enables a large variety of spectrometer architectures. We will concentrate on a few that illuminate the underlying principles. In particular, we will focus on the principles and implementations of dispersed-spectrum architectures. The chapter wraps up with a section on tunable lasers, which is an important application of MEMS tunable filters.

13.2 Amplitude Filters

13.2.1 Fabry-Perot Filters

Fabry-Perot resonators, or etalons, are among the simplest tunable optical MEMS filters. All that is required are two parallel mirrors with a distance between them that can be adjusted by a MEMS actuator. A typical implementation is shown in Fig. 13.1.

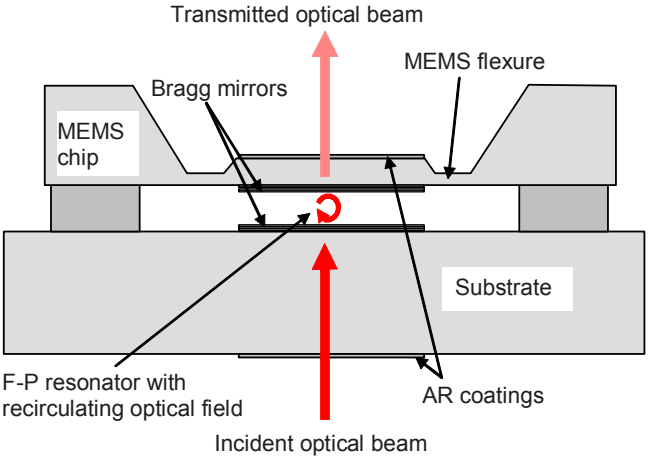


Figure 13.1. Typical implementation of a MEMS tunable Fabry-Perot filter. The F-P resonator is formed between two high-reflectivity Bragg mirrors. High reflectivity mirrors are needed to achieve good out-of-band rejection. The length of the resonator is adjusted by applying a voltage between the upper MEMS chip and the substrate. Anti-reflection coatings are used to suppress reflections from the other surfaces of the MEMS chip and substrate.

In this device a MEMS chip is bonded to a substrate such that their surfaces are parallel. A Fabry-Perot resonator is formed between the two facing surfaces that both have Bragg mirrors to enhance reflectivity. The other surfaces of the MEMS chip and substrate have Anti-Reflection (AR) coatings to avoid spurious reflections that would create additional interference effects. The MEMS chip is designed to have a stiff center region that supports the Bragg mirror. The stiff center region is suspended on compliant flexures such that the length of the Fabry-Perot cavity can be adjusted by applying a voltage between the MEMS chip and the substrate.

Fabry-Perot resonators are simple and versatile optical devices that are used in a wide range of applications. We looked at Fabry-Perot resonators as modulators in Chapter 6.7.6 and as displacement sensors in Chapter 12.2.4. In those applications the slopes of the transmittance or reflectance spectra that are most important, because they determine the modulation efficiency and measurement sensitivity. Filter applications are more demanding in that several other characteristic features, in addition to slope, must be precisely controlled. High-quality filters must meet stringent specifications on pass-band loss, out-of-band rejection, transition band width, and free spectral range (FSR).

To see how these qualities are related in F-P filters, we recall the F-P transmission spectrum derived in Chapter 12.2.4 (Eq. 12.16)

$$T = \frac{\dfrac{\left(1-r_1^2\right)\left(1-r_2^2\right)}{\left(1-r_2 r_1\right)^2}}{1+\dfrac{4r_2 r_1}{\left(1-r_2 r_1\right)^2}\cdot \sin^2\left(L\dfrac{2\pi}{\lambda}\right)} \tag{13.1}$$

where r_1 is the field reflectivity of the first F-P mirror, r_2 is the field reflectivity of the second F-P mirror, L is the F-P resonator length, and λ is the optical wavelength. In this description the Bragg mirrors are modeled as wavelength independent over the wavelength range of operation of the F-P. In other words, we assume that the filter is operated in the Bragg mirror's flat band where the reflectivity is uniform. That can only be valid if the length of the F-P cavity is much larger than the thickness of the Bragg stack, so we are making the implicit assumption that the F-P length is much longer than a wavelength. If this assumption is not fulfilled, then the dispersion of the Bragg mirrors much be considered in the modeling of the reflection and transmission spectra.

In the case of a symmetric F-P ($r_1=r_2=r$), Eq. 13.1 simplifies to (Eq. 12.19)

$$T = \frac{1}{1+\dfrac{4r^2}{\left(1-r^2\right)^2}\sin^2\left(L\dfrac{2\pi}{\lambda}\right)} = \frac{1}{1+C_F\sin^2\left(L\dfrac{2\pi}{\lambda}\right)} = \frac{1}{1+\dfrac{4F^2}{\pi^2}\sin^2\left(L\dfrac{2\pi}{\lambda}\right)} \tag{13.2}$$

where the parameter

$$C_F = \frac{4r^2}{\left(1-r^2\right)^2} \tag{13.3}$$

is the *coefficient of Finesse*, and

$$F = \frac{\pi}{2\cdot\sin^{-1}\left(\left(1-r^2\right)/2r\right)} \approx \frac{\pi\cdot r}{1-r^2} \tag{13.4}$$

is the *Finesse*, which is defined as the ratio of the period to the FWHM of the transmittance peak. The F-P transmission is plotted on a dB scale (i.e. $10\cdot Log_{10}[T]$) in Fig. 13.2. This is essentially the same plot as in Figs. 6.30 and 12.10, but the dB scale emphasizes and clarifies the characteristics of the F-P that are important for filter applications.

The first thing to note about the filter characteristics is that it is periodic and that the positions of the pass bands are dependent on the length of the resonator. The pass-band wavelength can therefore be tuned by controlling the resonator length with MEMS actuators, as described in the caption of Fig. 13.1. The periodic nature of the transmission limits the useful tuning range to one period, or one FSR.

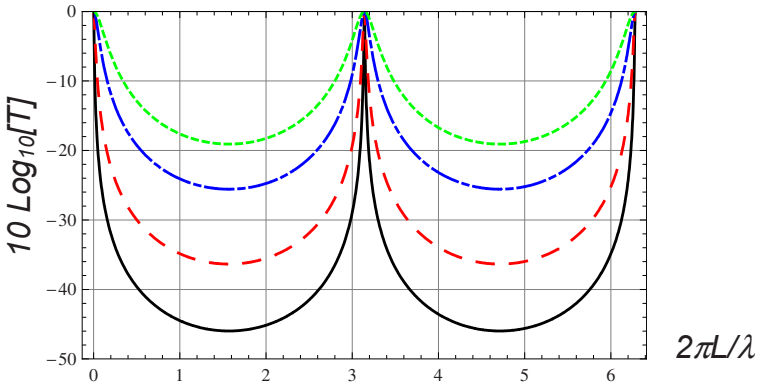


Figure 13.2. Transmission spectrum of symmetric a Fabry-Perot as a function of normalized length with mirror power reflectivities of $R=r^2=0.99$ (solid), $R=0.97$ (dashed), $R=0.90$ (dot-dashed), and $R=0.80$ (dotted).

Figure 13.2 demonstrates that high mirror reflectivities are required to make good F-P filters. The transmission spectrum of the symmetric F-P with power reflections of $R=0.8$ gives a maximum out-of-band rejection of less than *20dB*, which is not sufficient for many applications. A related problem is that the filter cannot separate more than a few channels across its FSR. Even if a *10dB* rejection is sufficient, the F-P with $R=0.8$ can only distinguish about 10 channels. As the mirror reflectivity is increased, the filter become progressively better and at $R\geq 0.97$, the

filter has excellent out-of-band rejection. We conclude that F-P filters need high-reflectivity mirrors, preferably with power reflectance above 0.95, to be useful for most applications[a].

High-reflectivity mirrors present difficulties in MEMS implementations. Metal mirrors that are the standard of optical MEMS, have too much transmission loss, or, if they are configured as thin films to lower the transmission loss, too low reflectivity. Until the arrival of Photonic Crystals, which is the subject of Chapter 14 and 15, high-reflectivity F-Ps were realized using multilayer dielectric mirrors. These multilayer dielectric mirrors require rigid substrates to avoid temperature dependent mirror curvature caused by the thermal stresses that build up in the mirrors stacks. This presents challenges for miniaturization that have been met through a variety of approaches.

Early work on silicon MEMS F-P filters used the full thickness of silicon wafer to provide a solid substrate [1]. These devices were fabricated by wafer bonding and were relatively bulky. Smaller devices have been created by using free-standing $Si\text{-}SiO_2$ mirror stacks, but these mirrors have some problems with curvature [2]. By careful compensation of the material stress in the dielectric stack, silicon-compatible, free-standing, dielectric mirrors with better than 99% power reflectivity have been demonstrated [3].

An approach that avoids the complications of bending due to thermal stress in free-standing dielectric stacks is to tune the filters thermally, rather than by mechanical motion. In such thermally-tuned devices the dielectric mirrors are deposited directly on a silicon substrate with an intermediate film of thermo-optical material. The temperature, and therefore the effective optical thickness, of the material between the dielectric mirrors, is controlled by thermal dissipation in integrated resistors. This approach has been used to create tunable channel-dropping WMD filters with narrow transition bands [4].

In contrast to silicon, the AlGaAs system enables growth of lattice-matched, thin film stacks with sufficiently large index variations to enable high reflectivity Bragg mirrors. Early work on this technology [5,6] has led to rapid development [7,8,9], and to the creation of MEMS tunable Vertical Cavity Surface Emitting Semiconductor Lasers (VCSELs) (For an excellent, in-depth description of MEMS tunable VCSELs see [10]). This type of fabrication process results in excellent mirrors, but the process is not compatible with silicon MEMS technology.

[a] We are talking exclusively about filter applications here. We have seen earlier that lower reflectivity mirrors can make excellent F-P sensors.

13.2.2 Bragg Filters

Fabry-Perot filters with high-reflectivity mirrors have good out-of-band rejection and narrow pass bands, as shown in Fig. 13.2. These F-P filters also have some undesirable features, however. The pass band, although it has low insertion loss at its center, is not flat. This means that signals that pass through the filter will see wavelength dependent loss. This is an undesirable trait both in communication and sensor applications.

One filter structure that has better pass-band flatness and narrower transition bands is the Bragg filter that we first discussed in Chapter 3.4.2 and again in 6.6 (where the focus is on waveguide implementations). This is clearly demonstrated in Fig. 3.11 that shows that we can get a flat passband and steep transitions bands with as little as three pairs of alternating layers of silicon and air.

The problem with this approach is that high-index contrast Bragg mirrors have large sidebands. This is shown in Fig. 3.11 where the transmission in the first set of sidebands reaches as high as 0.6. The passbands can be reduced by using more complex structures than the perfectly periodic mirror of Fig. 3.10, but high-index-contrast Bragg filters are nevertheless not well suited for applications where narrow transition bands and good out-of-band rejection are important figures of merit.

13.2.3 Microresonator Filters

The problems of MEMS implementations of high-Finesse Fabry-Perot filters are due to the bulky structures required to support high reflectivity mirrors. This difficulty is avoided in microresonator filters that are based on resonant waveguide coupling (see Chapter 6.7.7). Waveguide microrings, and other types of optical resonators, can be made in silicon waveguide technology. They can also be enhanced by MEMS actuators that provide tunability, but, as we will see shortly, the tuning functionality is different than that of tunable F-P filters. A generic implementation of a microresonator filter or switch is shown in Fig. 13.3. Here the microresonator is a disk, but the discussion is valid for any directional resonator, e.g. rings or spheres[b].

In microresonator filters the optical field on *Input 1* couples to and sets up a field in the resonator. If the circumference of the ring is an integer number of wavelengths, then the field in the ring builds up in phase. At wavelengths fulfilling this condition, we say that the ring is phase matched to the input field. The build up continues until the losses (intrinsic losses plus coupling to Output 2) matches the input coupling. At that point of steady-state operation, we say that we have im-

[b] Note that the characteristics of the filter are different if the modes of the resonator are non-directional, i.e. they couple equally to waves traveling in opposite directions on the waveguides.

pedance matching of the input coupling and resonator losses. Under simultaneous phase matching and impedance matching, we have maximum transfer of optical energy from Input 1 to Output 2. If the resonator is also intrinsically lossless (i.e. the only loss is though coupling to the two waveguides), then all the power on Input 1 is coupled into Output 2.

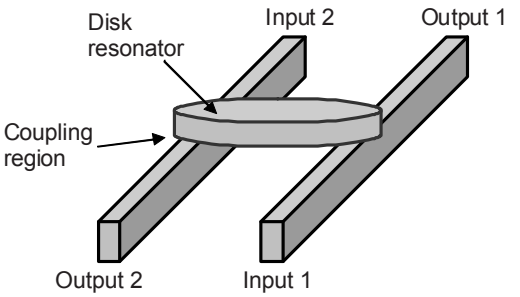


Figure 13.3. MEMS implementation of a waveguide microresonator filter. Part of Input 1 is transmitted to Output 1, and the rest is coupled into the disk resonator and from there to Output 2. The ratio of transmitted (Output 1) to reflected (Output 2) light is determined by the resonances of the disk and by the coupling from the waveguides to the disk. The effective optical circumference of the ring can be adjusted by thermal tuning or bychange-injection tuning, and the coupling can be adjusted by a MEMS actuator that controls the distance from the waveguides to the resonator.

The operation of the ring filter is conceptually similar to that of the Fabry-Perot: In the F-P, the input field couples into the cavity through the front mirror. If the round-trip length of the F-P is an integer number of wavelengths, then the input is phase matched and the recirculating field builds up until the total losses (i.e. internal losses plus output coupling through the mirrors) equal the input coupling. If the mirrors are equal and there is no internal loss in the cavity, then the transmission through the F-P is complete. Under these conditions, all the out-coupling (losses) from the cavity is through the back mirror. The part of the recirculating field that is coupled out through the front mirror is exactly equal in magnitude and exactly out of phase with the incoming light that is reflected from the front mirror, so these two fields interfere destructively and cancel each other exactly. This same explanation is correct for ring and disk resonators also as long as we substitute in and output couplers for front and back mirrors.

In Chapter 6.7.7 we show that the transmittance through a microring filter can be expressed as

$$T = \frac{\left(1 - r_1^2\right) \cdot \left(1 - r_2^2\right)}{1 + r_1^2 r_2^2 e^{-4\alpha \cdot \pi \cdot R} - 2 r_1 r_2 e^{-2\alpha \cdot \pi \cdot R} \cos\left(\beta_R \cdot 2\pi \cdot R\right)} \tag{13.5}$$

where $r_{1,2}$ are the coefficients of forward coupling[c] in each of the two coupling regions, α represents the internal losses of the ring mode, R is the radius of the ring mode, $\beta_R = \frac{2\pi}{\lambda_R}$ is the propagation constant of the ring mode, and λ_R is the wavelength of the ring mode. The waveguide-to-ring field coupling coefficients ($t_{1,2}$) are related to the forward coupling coefficients as

$$t_{1,2}^2 = 1 - r_{1,2}^2 \tag{13.6}$$

When the ring losses are negligible ($\alpha=0$), the transmission expression becomes

$$T = \frac{\left(1-r_1^2\right)\cdot\left(1-r_2^2\right)}{1+r_1^2 r_2^2 - 2r_1 r_2\cos\left(\beta_R\cdot 2\pi\cdot R\right)} = \frac{\dfrac{\left(1-r_1^2\right)\cdot\left(1-r_2^2\right)}{\left(1-r_1 r_2\right)^2}}{1+\dfrac{4r_1 r_2}{\left(1-r_1 r_2\right)^2}\sin^2\left(\beta_R\cdot\pi\cdot R\right)} \tag{13.7}$$

which in a symmetric ring ($r_1=r_2=r$) filter further simplifies to

$$T = \frac{1}{1+\dfrac{4r^2}{\left(1-r^2\right)^2}\sin^2\left(\beta_R\cdot\pi\cdot R\right)} \tag{13.8}$$

Comparing Eq. 13.8 to the transmission through a Fabry-Perot filter given by Eq. 13.2, we see that the expressions are identical assuming the following substitution

$$\beta_R\pi\cdot R = \frac{2\pi}{\lambda_R}\pi\cdot R \rightarrow \frac{2\pi}{\lambda}L \tag{13.9}$$

which simply says that the optical circumference of the ring corresponds to twice the optical thickness of the etalon. This is what we would expect, given the conceptual similarities between the operation of the F-P and the microresonator filter.

Comparison of Figs. 13.1 and 13.3 makes it clear that using MEMS technology, ring filters are simpler to fabricate than Fabry-Perot filters. The ring filter of Fig. 13.3 can be made with a few etch steps in a SOI wafer and large numbers of different designs and large arrays can easily be integrated in a relatively small area. The F-P of Fig. 13.1, on the other hand, require wafer bonding and include bulky support structures that make it difficult to integrate large number of devices on a common substrate.

[c] In principle there will also be coupling to the backward going wave and to radiation modes, but these can be neglected in well-designed waveguide couplers.

Practical ring filters do not have the same tunability as F-P filters. Comparing Eqs. 13.2 and 13.6, we see that to change the position of the pass band of a ring or disk filter as we would a F-P filter, we need to control the optical mode radius of the microresonator mode. This can be done by changing the effective refractive index of the mode by thermal tuning or by charge injection, but it is difficult to get the large tuning ranges that can be achieve in F-P filters tuned by MEMS actuators.

It is straight forward, however, to change the coupling from the waveguides to the ring by using MEMS actuators to control the separation of the resonator from the waveguides. This type of tuning does not lead to a change in the position of the pass bands, but enables switching [11]: Referring to Fig. 13.3, we see that the optical power on Input 1 can be directed to either Output 1 (no coupling from the waveguide to the disk) or Output 2 (complete coupling from the waveguide to the disk), depending on the waveguide-disk separation.

Just as for Fabry-Perot interferometers, we usually want the losses of ring filters to be as low as possible. In practical situation, the losses are often dominated by scattering from roughness on the side walls of the ring waveguide. Good patterning and etching technologies are therefore very important for successful implementations of ring filters.

13.3 Dispersion Compensators

In almost all applications of amplitude filter it is important to be able to substantially suppress out-of-band signals and to have an abrupt transition between the pass bands and rejection bands. These required abrupt transitions between high transmission and very low transmission makes amplitude filters hard to realize. Filters that are primarily applied to phase corrections are less challenging, because they do not require high out-of-band suppression.

Dispersion compensation in Wavelength Division Multiplexed (WDM) fiber optic systems are good examples. For such systems, low-Finesse F-Ps provide a convenient means of dispersion compensation. To avoid unwanted amplitude variations, dispersion compensation is preferably carried out with Gires-Tournois (G-T) interferometers [12], which is a variation of the F-P filter. In the G-T interferometer the back mirror is highly reflective. Ideally it has 100% reflection, so the ideal G-T is an all-pass filter with a strong phase variation around resonance.

Figure 13.4 shows a MEMS implementation of a G-T interferometer. It is based on the mechanical antireflection switch (MARS) device [13]. In spite of the fact that the back mirror has close to 100% reflectance, the MARS device is a low-Finesse resonator, because the front mirror is a single, free-standing, $\lambda/4$ silicon-nitride film as shown in Fig. 13.4. The MARS phase filter performs very well as a

dispersion-slope compensator. A linear dispersion tunable from -100 ps/nm to 100 ps/nm over 50 GHz in C-band has been experimentally demonstrated [14], verifying that high Finesse is not required for this function.

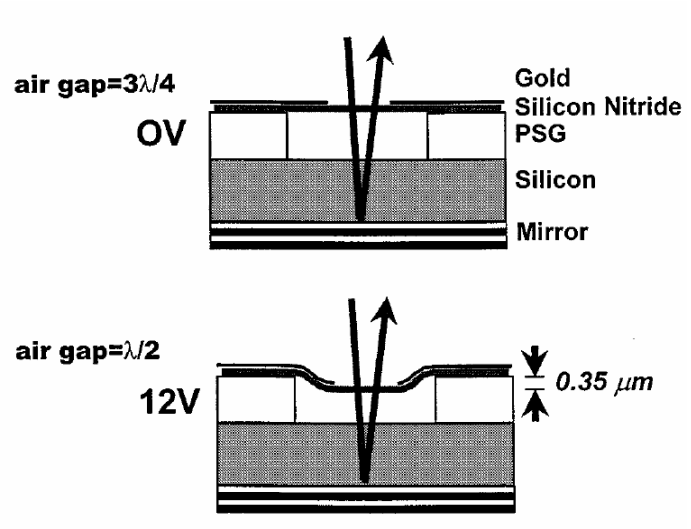


Figure 13.4. MEMS implementation of a Gires-Tournois all-pass filter designed for dispersion compensation. The back mirror is a high-reflectivity, multi-layer Bragg mirror, while the front mirror is a simple l/4 film. The length of the cavity is controlled by pulling the nitride mirror towards the substrate using electrostatic actuation. Reprinted from [13] with permission.

The G-T interferometer can also be operated on oblique incidence so that the optical beam follows a zigzag pattern as shown in Fig. 13.5. In this set-up, the reflections from the back mirror are spatially separated, and the output is the interference pattern of the beams coming off the front mirror. This geometry allows individual phase modulation of the beam that creates the output interference pattern. The G-T interferometer at oblique incidence therefore operates more like a grating than a standard F-P or G-T at normal incidence. This grating-like operation, combined with phase modulation of the individual the reflections, enables a variety of filter and switching applications, including tunable (de)interleavers [15], amplitude filters [16], and dispersion compensators [17].

The G-T interferometer at oblique incidence is, however, not an all-pass filter, even with an ideal, 100% reflectivity mirror. This is due to the fact that the interference of the reflected beams creates a number of higher-order diffraction modes on the output. Careful attention must therefore be paid to avoiding parasitic amplitude modulation as the phase is tuned.

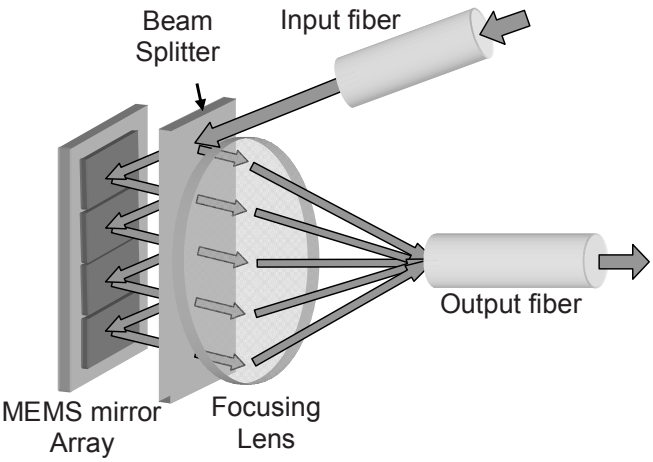


Figure 13.5. Schematic diagram of the MEMS Gires-Tournois interferometer. The back mirror is replaced with a MEMS micromirror array, in which the individual mirrors are electrostatically actuated in piston motion so that the reflections are phase modulated. The output is formed by the interference of multiple spatially separated beams coming off the beam splitter.

13.4 MEMS Spectrometers

In the preceding sections we emphasized the use of filters to *manipulate* the optical spectrum. In section 13.2 the focus was on separating different parts of the spectrum into pass bands and rejection bands and in section 13.3 we briefly looked at controlling and adjusting the spectral phase. Now we will turn our attention to spectroscopy, where the objective is to *measure* the optical spectrum.

Figure 13.6 show two commonly used spectrometer architectures. The first uses a tunable filter with a pass band that is swept across the spectrum of interest and a single detector that measures the transmitted optical power for each setting of the filter. Its simple and compact design makes it well suited for miniaturization, particularly when the tunable filter is implemented as a compact MEMS device.

The diffractive spectrometer of Fig. 13.6b uses a diffractive element to disperse the spectrum onto a detector array. The dispersive element is shown as a grating here, but could take other forms. This spectrometer architecture has the advantages that it measures the whole spectrum all the time, i.e. it is not wasting photons. It is, however, more complex and requires more components and more space than the swept-filter spectrometer, so it is more difficult to miniaturize.

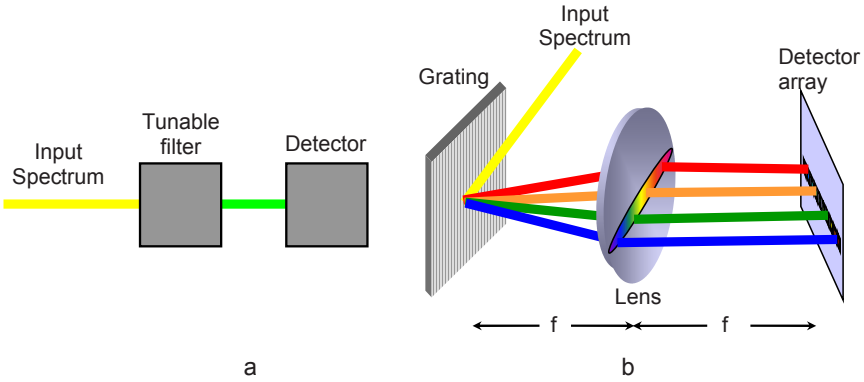


Figure 13.6. Commonly used spectrometer designs. The tunable-filter spectrometer (a) uses only a single detector combined with a filter that must be tuned over a set of transfer functions that allows the reconstruction of the input spectrum. In the diffractive spectrometer (b) the incident light is dispersed by a diffractive element (e.g. a grating) onto a detector array that measures the spectrum without the need for tuning of any sort

The basic geometry shown here does not require tunable elements, but there are a number of variations that utilize MEMS actuators to enhance the functionality and/or simplify the overall system. Examples include the use of MEMS gratings that tune the diffraction of the incident light, and the replacement of the detector array with a reflective MEMS modulator array that enables individual modulation of the spectral components of the incident light. Both these two types of systems are discussed below.

13.4.1 Swept Pass Band Spectrometers

The conceptually simplest way to create a MEMS spectrometer is to sweep a narrow pass-band, tunable filter across the spectral range of interest and thereby obtain the spectral distribution of optical power of the incident light. The Fabry-Perot filter described in section 13.2.1 works well in this capacity. Unambiguous data can only be obtained over one Free Spectral Range (FSR) of the F-P filter, and the spectral resolution is determined by the width of the pass band. The number of independent spectral bands that can be recorded is therefore equal to the Finesse. (Remember that the Finesse is defined as the ratio of the period to the FWHM of the transmittance peak).

We therefore want high-Finesse F-P to get good resolution, but for that we pay the price of reduced optical efficiency. At any given time the tunable F-P filter only lets through a fraction of the total spectrum equal to the inverse of the Finesse. This low optical efficiency is a serious limitation of the swept-passband, F-P spectrometer, and the major motivation for finding alternatives with higher throughput.

13.4.2 Generalized Transform Spectrometers

To make the spectrometer of Fig. 13.6a more efficient, i.e. less wasteful of photons, we have to give up on the idea of a narrow pass band surrounded by rejection bands with very low transmission. Good rejection is important in WDM channel filters and other signal-processing applications, but there is no need for strong rejection bands in spectroscopy. What we need is a set of sampling functions, or basis functions, that each is used to capture part of the spectrum such that the complete[d] spectrum can be reconstructed from the measurement data.

To see how a spectrum can be captured with a set of non-orthogonal sampling functions, consider a digitized[e] spectrum of N spectral measurements. To capture these N values we need N sampling functions. The measured optical power is related to the power spectrum by the following expression

$$\begin{pmatrix} b_1 \\ b_2 \\ \cdot \\ b_N \end{pmatrix} = \begin{pmatrix} s_{11} & s_{12} & \cdot & s_{1N} \\ s_{21} & s_{22} & \cdot & s_{2N} \\ \cdot & \cdot & \cdot & \cdot \\ s_{N1} & s_{N2} & \cdot & s_{NN} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \cdot \\ a_N \end{pmatrix} \Leftrightarrow \overline{b} = S \cdot \overline{a} \tag{13.10}$$

where a_n is the optical power in the spectrum at wavelength n, b_n is the total measured power using sampling function n, and s_{nm} is the transfer function at wavelength m of the n^{th} sampling vector. The sampled spectrum is then found from the interferogram by inverting the matrix

$$\overline{a} = S^{-1} \cdot \overline{b} \tag{13.11}$$

The measurement matrix is invertible, so that the power-spectrum vector can be found if the sampling vectors are linearly independent. This is of course a much less stringent requirement than that they be orthogonal. We will use the term Generalized Transform Spectrometer for devices that use the approach described by Eq. 13.11.

The measurement matrix is particularly simple if we use a swept narrow band filter. In the ideal case it is simply diagonal with ones on the diagonal and zeros everywhere else. That matrix is its own inverse, so using it simplifies the post processing of the spectral data, but post processing by matrix multiplication (note that the inversion of the matrix only have to be done once as a part of the programming of the spectrometer) is an inexpensive operation so in practical MEMS

[d] No finite set of basis functions allows perfect reconstruction of the spectrum. Any realizable set will impose restrictions on spectral resolution and range.

[e] The fact that the spectrum is digitized does not represent any kind of practical limitation. Any modern MEMS spectrometer has a digitized representation of the measured spectrum as its output.

implementations it is more important to simplify the hardware and to design more efficient (less wasteful) sampling functions.

Up to now, the conventional wisdom in spectrometer design has been to create hardware that makes the digital signal processing simple. There is an increasing trend towards a design philosophy that says that the hardware should be designed to gather the most complete data about the spectrum, irrespective of how complicated the post processing has to be. Examples of such designs are described in section 13.4.4.

13.4.3 Fourier Transform Spectrometers

One highly-efficient spectrometer that is well suited to MEMS implementations, *and* requires only relatively simple post processing, is the Fourier Transform Spectrometer. It is based on the Michelson interferometer (described in Chapter 12.2.1) as shown in Fig. 13.7. The moving mirror of the Michelson interferometer is oscillated back and forth in a periodic manner. Harmonic motion is common, because it simplifies the mechanical design and control, but other time waveforms are sometimes used to simplify the signal processing (although this is typically not of great concern in modern systems).

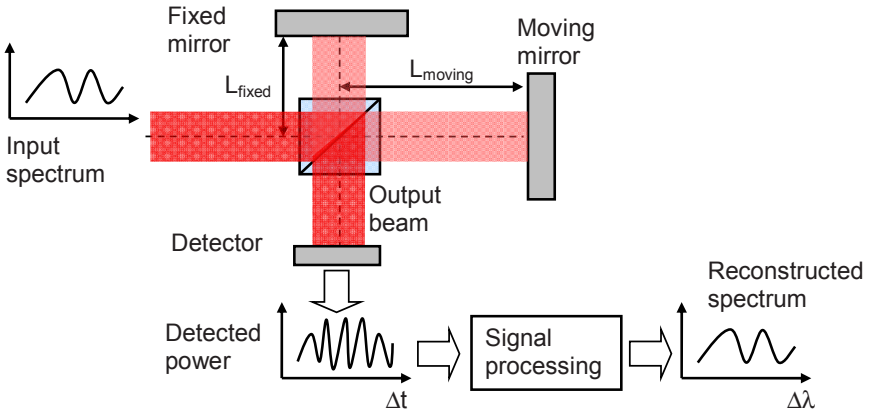


Figure 13.7 *Fourier transform spectrometer based on a Michelson interferometer. The output of the Michelson interferometer varies as a function of time as the moving mirror goes through its periodic motion. This time sequence is the inverse Fourier transformed to recreate the input spectrum.*

The motion of the mirror creates a time dependent output of the optical detector. The time sequence is first converted to a power-vs.-position function, which is then inverse-Fourier-Transformed to yield an estimate of the input spectrum. Due

to the efficiency of Fast-Fourier-Transform algorithms, this is less computationally intensive than matrix multiplication.

From Eq. 12.1 we know that the power transfer function for monochromatic light through the Michelson interferometer is

$$\frac{P_{out}}{P_{in}} = 0.5 \cdot \left[1 + \cos\left(\frac{2\pi \cdot 2\Delta L}{\lambda} \right) \right] = 0.5 \cdot \left[1 + \cos\left(\frac{2\pi \cdot \delta}{\lambda} \right) \right] \tag{13.12}$$

where P_{out} and P_{in} are the output and input optical powers, λ is the wavelength, and $\delta = 2\Delta L = 2\left(L_{fixed} - L_{moving}\right)$ is the total path length difference for the light reflected from the fixed and moving mirrors. Equation 13.12 shows that the Fourier-Transform spectrometer measures spectra by sampling with harmonic functions.

The part of the measured power that varies with the path length is called the interferogram, $B(\delta)$. For reasons that will become obvious, we prefer to use the wave number, $\nu = \frac{1}{\lambda}$, in these calculations. The expression for the monochromatic interferogram is then

$$B(\delta) = \cos(2\pi \cdot \nu \cdot \delta) \tag{13.13}$$

For a spectrally extended source with a power spectral density $I(\nu)$, we must integrate over the wave numbers to find the interferogram

$$B(\delta) = \int_{-\infty}^{+\infty} I(\nu) \cdot e^{-2\pi \cdot \nu \cdot \delta} \cdot d\nu \tag{13.14}$$

We recognize this as the Fourier Transform, so it follows that

$$I(\nu) = \int_{-\infty}^{+\infty} B(\delta) \cdot e^{2\pi \cdot \nu \cdot \delta} \cdot d\delta \tag{13.15}$$

We conclude that we can find the spectrum by inverse Fourier transform of the interferogram.

Perfect reconstruction of the spectrum requires that we scan the path difference from minus to plus infinity. In practical systems we scan from zero up to a maximum path length difference δ_{max}. (Note that since $B(\delta)$ is symmetric in δ, we need not include negative path-length differences. It is important to include $\delta=0$, however.) The imperfectly reconstructed spectrum is then

$$I'(\nu) = \int_{-\infty}^{+\infty} \Pi(\delta/\delta_{\max}) \cdot B(\delta) \cos(2\pi \cdot \nu \cdot \delta) \cdot d\delta \tag{13.16}$$

where $\Pi(\delta/\delta_{\max})$ is the *Rectangular* function[f], defined as

$$\Pi(\delta/\delta_{\max}) = \begin{cases} 1 & \text{if } -1 < \delta/\delta_{\max} < 1 \\ 0 & \text{if } |\delta/\delta_{\max}| > 1 \\ 0.5 & \text{if } \delta/\delta_{\max} = \pm 1 \end{cases} \tag{13.17}$$

Note that the values at the discontinuities are sometimes defined differently than in Eq. 13.17.

The inverse Fourier transform of the Rectangular function is [18]

$$h(\nu) = 2\delta_{\max} \frac{\sin(2\pi\nu \cdot \delta_{\max})}{2\pi\nu \cdot \delta_{\max}} \equiv 2\delta_{\max} \cdot \operatorname{sinc}(2\pi\nu \cdot \delta_{\max}) \tag{13.18}$$

The reconstructed spectrum is the convolution of the true spectrum with this *sinc* function

$$P'(\nu) = P(\nu) \otimes h(\nu) = \int_{-\infty}^{+\infty} P'(u) \cdot h(\nu - u) \cdot du \tag{13.19}$$

This is exactly what we would expect: The measured, or reconstructed, spectrum is the true spectrum convolved with, or broadened by, the response function of the spectrometer.

Equation 13.19 illustrates the main problem with Fourier-Transform spectroscopy in its most basic form: The *sinc* function that is convolved with the spectrum has large side lobes, and these side lobes create errors in the reconstructed spectrum. The solution is to *apodize* the inteferogram. Apodization refers to the technique of differentially weighting the interferogram such that the multiplying function in Eq. 13.16 is not the *Rectangular* function, but some other function with less abrupt transitions and therefore less prominent side lobes in its Fourier Transform.

The design of apodization functions is complex and the optimum choice depends on the application. Reference 19 gives a good introduction to apodization in Fourier-Transform spectroscopy. Here we will just quote the results most relevant to our discussion.

Popular apodizations include triangular functions and truncated harmonics and Gaussians. The truncated harmonic apodization function of the form

[f] The Rectangular function is also known as the rectangle function, rect function, unit pulse, boxcar function, or the top-hat function.

$$\Pi(\delta/\delta_{\max})\cdot\frac{1}{2}\left[1+\cos\left(\frac{\pi\cdot x}{\delta_{\max}}\right)\right] \tag{13.20}$$

has a Fourier Transform with side lobes that have maxima that are *-2.7%* of the central lobe maximum. The Gaussian apodization function

$$\Pi(\delta/\delta_{\max})\cdot e^{-\frac{1}{\ln 2}\left(\frac{\pi\cdot x}{2\delta_{\max}}\right)^2} \tag{13.21}$$

does even better. The maximum side lobes of its Fourier Transform are only *-0.45%* of the central-lobe value. These ratios are substantially smaller than the side-lobe-to-central-lobe ratio of *-21%* of the sinc function.

This reduction in side-lobe strength, and the associated reduction in errors caused by their presence, come at the cost of reduced spectral sensitivity. The *sinc* function has its first zeros for $\nu=\pm 1/2\delta_{\max}$, so the full width of the central lobe is $1/\delta_{\max}$, and its Full Width at Half Maximum (FWHM) is approximately $0.6/\delta_{\max}$. The Fourier Transform of the truncated harmonic function has a FWHM of exactly $1/\delta_{\max}$ and the value for the truncated Gaussian is only insignificantly larger ($\sim 1.02/\delta_{\max}$).

In practical applications, the reduction in errors caused by side lobes is well worth the increase in resolution, so apodization is commonly used. We will therefore say, somewhat arbitrarily, that the resolution[g] of the transform spectrometer is given by

$$\Delta\nu=\frac{1}{\delta_{\max}} \tag{13.22}$$

Expressed in terms of wavelength, this becomes

$$\Delta\lambda=\frac{\Delta\nu}{\nu^2}\Rightarrow\Delta\lambda=\frac{\lambda^2}{\delta_{\max}} \tag{13.23}$$

These equations point out a hard limit on scaling Fourier Transform spectrometers; the application-specific resolution determines a minimum acceptable mirror translation.

[g] What is the best measure of resolution in spectroscopy is a much debated question. It is definitively application dependent. Our choice of FWHM is somewhat conservative and it simplifies the mathematical expressions.

13.4.4 MEMS Implementations of Transform Spectrometers

There are many considerations other than resolution that go into the design of Fourier Transform spectrometers. One is the minimum measurable wave number that can be measured. It is determined by the minimum-resolvable path-length difference, which is typically limited by the sampling speed. Another important issue for traditional designs is the variable losses incurred as the path length is changed. Typically this is not a difficulty for MEMS implementations, due to the relatively short path-length variations that can be sustained by MEMS actuators. The resolution as expressed in Eq. 13.23 therefore highlights the main challenge of MEMS implementations of transform spectrometers.

The spectral resolution is inversely proportional to the maximum actuation distance, so long-travel actuators are required. As an example, let's say we want to construct a transform spectrometer for channel monitoring in a WDM system with 100 GHz channel spacing at 1.55um wavelength. A channel spacing of 100 GHz corresponds to 0.8 nm at 1.55um wavelength, so we find

$$\Delta L_{\max} = \frac{\delta_{\max}}{2} = \frac{\lambda^2}{2 \cdot \Delta\lambda_{FWHM}} \approx \frac{\left(1.55 \cdot 10^{-6}\right)^2}{2 \cdot 0.8 \cdot 10^{-9}} m \approx 1.5 \cdot 10^{-3} m \tag{13.24}$$

Motion in excess of millimeters is difficult, if not impossible, to achieve in MEMS. The calculation demonstrates that the micron-scale displacements that are sufficient for many MEMS applications are not useful in transform spectrometers, and that even long-range MEMS actuators, e.g. electrostatic combdrives with several tens of microns of motion, achieve only modest resolution.

The Michelson interferometer is the building block of the traditional Fourier Transform spectrometer, but many other configurations using other types of interferometers can be used. In MEMS implementations it is often beneficial to use non-traditional architectures that are designed to fit into the MEMS-fabrication environment and to facilitate miniaturization. It is important to find architectures that minimize both the size and number of components. The optimum design depends on the available manufacturing methods, so many different structures have been demonstrated as MEMS have been extended and refined. Here we will give a few typical examples to illustrate how different MEMS technology leads to different solutions.

Figure 13.8 [20] illustrates a classic design based on a Michelson interferometer in its most traditional implementation. The optical beam to be analyzed is split in two by a beam splitter and the two parts are reflected from two different mirrors. One of the mirrors is actuated by an electrostatic combdrive to create a variable path length. After reflection, the two parts of the optical beam recombine on the beam splitter, and interfere on the detector to give a detected power that is harmonically dependent on the path-length difference as described above.

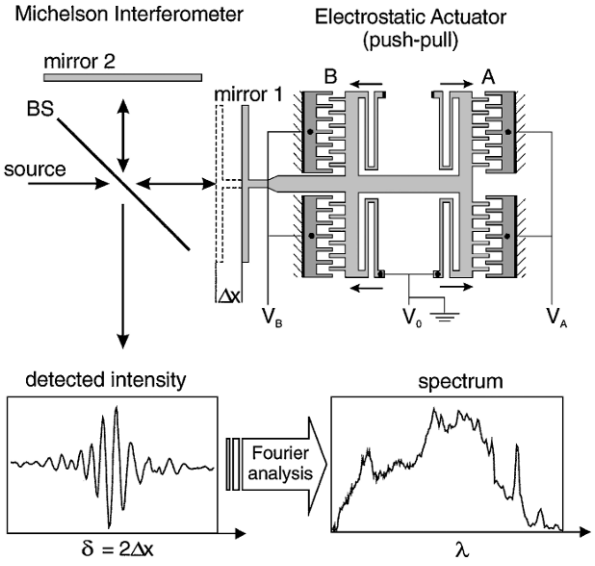


Figure 13.8. Schematic of a Michelson interferometer used as a Fourier Transform spectrometer. The movable mirror of the F-T spectrometer is controlled by a MEMS electrostatic actuator. Reprinted from [20] with permission.

Figure 13.9 shows a variation of the traditional transform spectrometer design [21]. Here the geometry is changed to allow more of the components of the spectrometer to be fabricated in the MEMS technology directly, as opposed to being manufactured separately and positioned on the chip using some type of hybrid integration approach. The structure shown in Fig. 13.9 is implemented on a single chip through a combination of anisotropic etching and Deep Reactive Ion Etching (DRIE).

The starting point for the MEMS fabrication is a SOI wafer with a <110> device layer. The device layer is shaped by a combination of anisotropic etching and DRIE (Deep Reactive Ion Etching). Anisotropic etchants, e.g. KOH, etches the <111> planes of silicon at much lower speeds than other crystalline planes. These etchants can therefore be used to create very smooth, vertical surfaces that can be used as mirrors and beam splitters for optical beams that propagate in the plane of the device layer. The <111> planes that are perpendicular to the chosen <110> surface intersect at angles of *70.6* and *109.4* degrees as shown in Fig. 13.10. These angles dictate the geometry of the spectrometer shown in Fig. 13.9. Note in Fig. 13.10 that there are other <111> planes that are not perpendicular to the chosen <110> surface that will interfere with the formation of vertical surfaces during anisotropic etching.

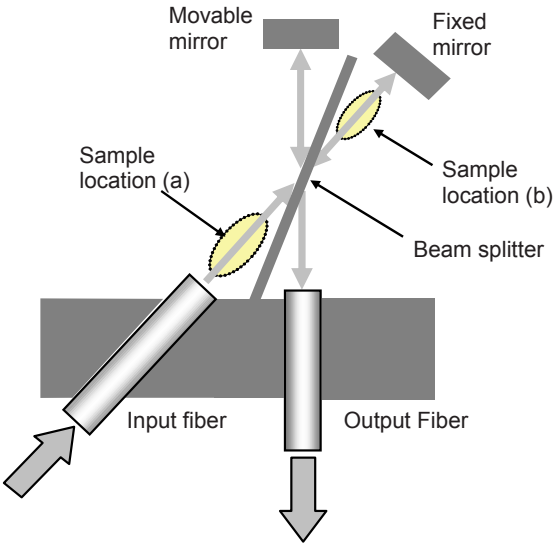


Figure 13.9. Schematic of single-chip integrated transform spectrometer based on vertical micromirrors with integrated MEMS actuators. The non-normal incidence on the beam splitter is due to the restrictions of the surfaces that can be defined by anisotropic etching of Si.

The specific orientations of the <111> planes limit the type of geometries that can be created by anisotropic etching alone, so DRIE is used to add flexibility to the design. Electrostatic actuators and fiber grooves that do not need atomically smooth surfaces are therefore created by DRIE. In the spectrometer of Fig. 13.9, the beam splitter and the movable mirror are defined using anisotropic etching, while the fixed mirror is defined by DRIE.

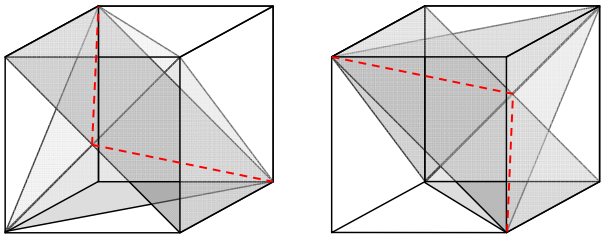


Figure 13.10. The intersections between a specific <110> plane (going from the upper to the lower corner on the front surface) and the <111> that are perpendicular to it are shown as dashed lines. These lines intersect at 70.6 degrees in the corners and at 109.4 degrees at the mid points of the front and back surfaces.

The two transform spectrometers shown in Fig. 13.8 and 13.9 are both of the traditional design based on the Michelson interferometer. The characteristic advantages and challenges of MEMS technology have inspired non-traditional solutions of different kinds. One such architecture uses the variable-amplitude grating shown in Fig. 13.11 [22].

As described in Chapter 10, the reflected optical power from the grating has a harmonic dependence on the grating amplitude. This matches the harmonic dependence on the optical path-length difference of the output of traditional Fourier transform spectrometers. With variation in grating amplitude playing the role of path length difference, the grating-transform spectrometer maps readily onto the traditional design. The main advantage of this solution from a MEMS point of view is that the grating acts both as a beam splitter and as a two-beam interferometer with a variable path-length difference.

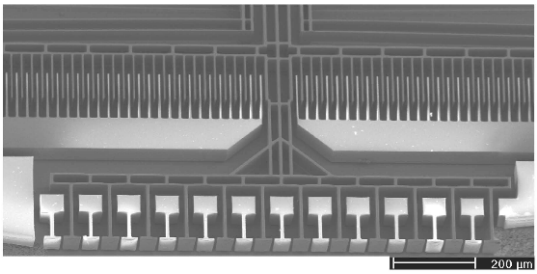


Figure 13.11. Transform spectrometer based on a diffraction phase grating with tunable grating amplitude. The grating consists of alternating fixed (light) and movable (dark) mirror elements. The movable mirrors are displaced by an electrostatic actuator to create a variable path length difference. Reprinted from [22] with permission.

The spectrometer shown in Fig. 13.12 [23] represents a more drastic departure form traditional transform-spectrometer design. Here the incident optical field forms a standing wave between the semitransparent front mirror and the highly reflecting back mirror. The standing wave is sampled by a semi-transparent detector in a fixed location, and the period of the standing wave pattern is varied by moving the rear mirror of the standing-wave cavity. As in other transform spectrometers, the response is a harmonic function of the mirror position, so the resolution has the same dependence on maximum mirror displacement given by Eq. 13.22.

The non-traditional implementations of transform spectrometers of Figs. 13.11 and 13.12 show how the flexibility of MEMS technology enables non-traditional solutions. Both spectrometer designs are compact and require few components, so they are well-suited for miniaturization and integration with electronics. The drawback of these geometries is that they only achieve modest resolution due to the limited maximum displacement of their actuators.

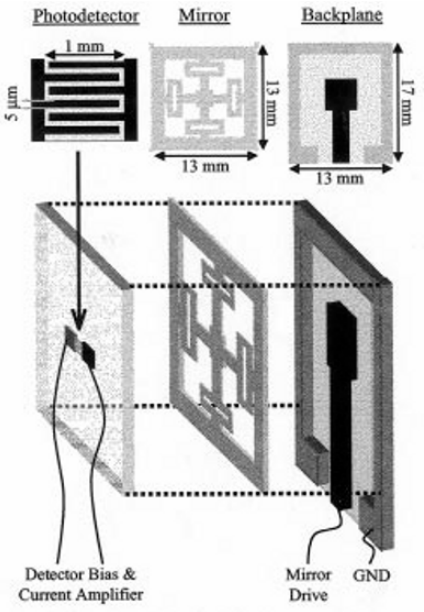


Figure 13.12. Transform spectrometer using a semitransparent detector in a standing wave cavity. The back mirror is moved to change the period of the standing wave and create an output that is a harmonic function of mirror position. Reprinted from [23] with permission.

13.5 Diffractive Spectrometers

13.5.1 Spectral Synthesis

The filters we have studied so far are simple devices with only one or at most a few (in the case of the oblique-incidence Gires-Tournois interferometer) degrees of freedom of mechanical motion. This simplicity enables compact and low cost MEMS implementations, but it also limits the range of spectral manipulations that are supported by the filters. Much more flexible and functional optical filters and optical synthesizers can be created if the spectrum is first spread by a dispersive element and then the spectral components are individually modulated before recombination. This device, called the Heritage-Weiner optical synthesizer after its inventors [24], was first described in Chapter 5.6.3 and is shown conceptually in Fig. 13.13. This optical synthesizer is similar to the grating spectrometer of Fig. 13.5b, but instead of a detector array in the spectral plane, we use a MEMS Spatial Light Modulator (SLM).

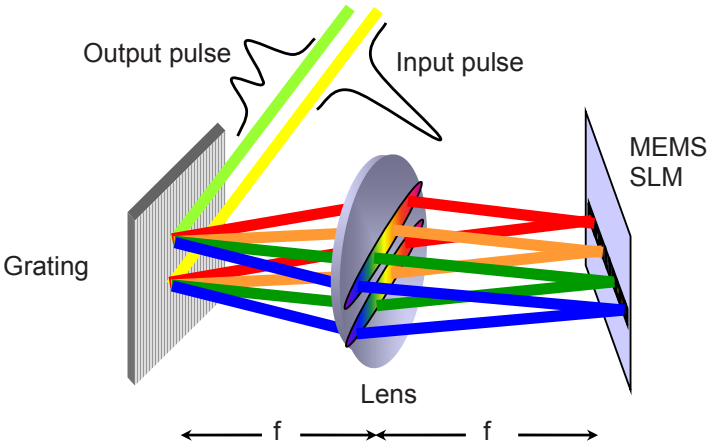


Figure 13.13. Basic optical system for manipulation and synthesis of optical spectra. The incident light is collimated onto a diffraction grating and dispersed on a Spatial Light Modulator (SLM) through a lens placed one focal length away from the grating. The spectral components are modulated by the reflective SLM and then recombined on the grating, before the output is focused onto an optical fiber, detector, detector array, or other output device. This flexible system has many applications, including pulse shaping as shown here.

In the optical synthesizer of Fig. 13.13, the spectral components of the incident light are dispersed by the grating and modulated by the MEMS SLM. Depending on the application, the SLM is constructed to modulate the amplitude or phase, or both, of the dispersed light. The synthesizer of Fig. 13.13 employs a reflective SLM and a folded geometry, in which the same grating-lens combination is used both to disperse the incoming light and to recombine the output light. The reflective SLM is slightly tilted with respect to the optical axis so that the input and output beams are spatially separated. Transmissive modulators work just as well, but require a separate lens and grating for the recombination of the output. The reflective geometry is preferred in microsystems, because this configuration requires fewer components and less space, and because reflective SLMs are easier to implement using MEMS technology.

The versatility of the optical synthesizer of Fig. 13.13 makes it a favored design. The flexibility in size, form, and function of Optical MEMS has made it the technology of choice for a large number of applications. Channel-extraction filters for Wavelength-Division-Multiplexed optical fiber networks are good examples. This application require a flat pass band to avoid signal distortion, combined with strong (> 40 dB) side band rejection and narrow transitions between the pass band and rejection bands to suppress cross talk between channels.

Such a filter can be realized using the optical synthesizer configuration with a simple MEMS SLM as shown in Fig. 13.14. The SLM used here is an adjustable aperture created by two beam blocks that are each positioned by MEMS actuators. In this configuration the center wavelength of the filter is tuned by moving the beam blocks in common mode and the pass band width is adjusted by moving the beam blocks differentially. The width of the transition bands between the pass bands and the rejection bands are determined by the spot size on the SLM. Filters with better than 50 dB sideband rejection and 1 GHz 1dB to 40 dB transition bands in a 1 by 1mm footprint have been demonstrated using this design [25].

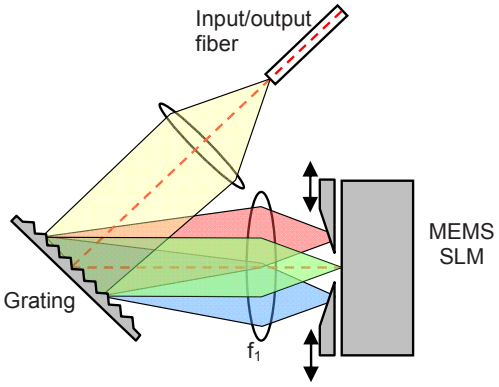


Figure 13.14. MEMS implementation of an optical synthesizer set up with an adjustable aperture in the spectral plane. The two beam blocks that form the adjustable aperture can be independently positioned to tune the pass band width and center frequency.

The simple SLM used in the filter of Fig. 13.14 controls the spectral amplitude of the optical output field. Changing the relative phase of the spectral components has the effect of changing the temporal shape of the input. This effect is much used in femto-second optical pulse shaping [26] and spectral phase measurements [27]. The same effect is also used to create adjustable time delays [28,29,30] in Optical Coherence Tomography systems. In the latter application, the SLM is simply a scanning mirror that rotates around an axis perpendicular to the direction of spectral dispersion. Tilt of the mirror around this axis results in a variable, linear gradient of the spectral phase, which translates into a variable time delay in the temporal regime.

Variations of the geometry of Fig. 13.13 are also used to interchange wavelength channels between fibers in Wavelength-Division-Multiplexed optical fiber systems. The SLM then consists of arrays of rotating micromirrors that are configured to switch signals between different fibers as shown in Fig. 8.6 in Chapter 8.3. Both wavelength selective optical WDM switches [31] and WDM add-drop filters [32] based on this principle have been demonstrated.

The optical synthesizer is also very well suited to spectroscopy. The SLM can be used to set up a large variety of sampling functions that can be used to optimize spectral measurements for specific applications. This has lead to the development of a rich field of mathematics and practice, generally referred to as Hadamard spectroscopy [33].

13.5.2 Diffractive MEMS Spectrometers

For as powerful as the general optical synthesizer of Fig. 13.13 is, it is also complex and contains many optical components that must be accurately aligned. It is therefore not easy to miniaturize and integrate, so it is practical to find simpler alternatives wherever possible. A variation of the grating spectrometer that uses optical MEMS, not as a SLM to modulate dispersed light, but as a tunable dispersing element, is shown in Fig. 13.15. The point of this design is to reduce the number of components in the system by combining the grating and the SLM. This architecture is neither as powerful, nor as efficient as the optical synthesizer, but it requires fewer components and less space, so it preferable for miniaturized systems and has been demonstrated as a viable geometry for implementation of a variety of applications, including spectral synthesis [34], optical filters [35] and pulse shapers [36].

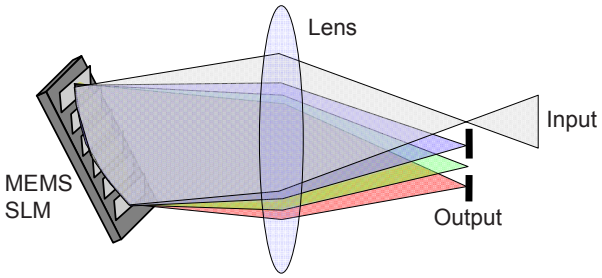


Figure 13.15. Optical synthesizer that combines the functions of the dispersive element and SLM into one component. The MEMS SLM is set up to diffract the desired spectral components of the incident light onto the output, which in this case is a simple aperture.

The objective of the simplified synthesizer of Fig. 13.15 is to create a MEMS SLM that can be set up such that any desired part of the input spectrum is diffracted onto the output. Ideally this should be done with no loss and perfect recreation of the desired spectrum, but in practice there will be limitations on both throughput and spectral control.

To understand these limitations, consider the impulse response of the filter synthesizer. The output is an impulse train that contains impulses that each are delayed and attenuated replicas of the input pulse as shown in Fig. 13.16. The delay for

each impulse on the output is given by the total path length from the input to the diffractive surface and back for that particular pulse. The delays therefore correspond to the height distribution of the individual reflectors of the diffractive surface. Likewise will the attenuation of each pulse correspond to the area of the reflective surface that creates it. The impulse response is therefore determined by the height distribution of the diffractive surface. Neglecting weak wavelength dependencies in diffraction efficiency and output coupling, the transmission of the filter is then given by the Fourier transform of the height distribution of the reflectors [37].

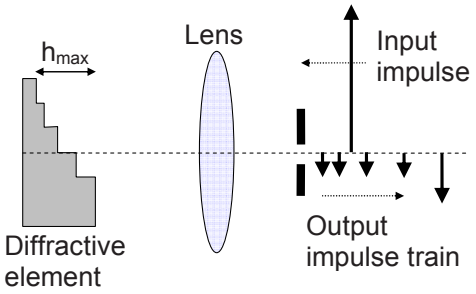


Figure 13.16. The impulse response of a diffractive MEMS element is a train of impulses with amplitudes determined by the areas of the reflectors and spacing determined by the height difference between the reflectors.

The impulse-response picture of Fig. 13.16 gives us a powerful tool for analysis and design of diffractive filters. For details of the underlying mathematical and computational details the reader is referred to [38]. Here we will focus on the conceptual insight necessary for MEMS design and implementation.

The first observation we make is that the transfer function of the diffractive filter is the Fourier transform of a non-negative sequence. The diffractive filter can therefore in principle synthesize *any* transfer function to within a constant. In terms of spectral synthesis, the diffractive filter is just as powerful as the synthesizer of Fig. 13.13.

In perfect analogy with the Fourier Transform spectrometer, the resolution of the diffractive filter is the inverse of the total path length difference between the leading and trailing impulses in the output pulse train. This maximum path difference is equal to the distance, h_{max}, along the optical axis between the first and last reflector. The resolution of the diffractive filter is then given by

$$\Delta\nu_{df} \approx \frac{1}{2h_{\max}} \tag{13.25}$$

which in terms of wavelength becomes

$$\Delta\lambda \approx \frac{\lambda^2}{2h_{\max}} \tag{13.26}$$

We have expressed these relations as approximations to reflect the fact that the exact values depend on the apodization.

Equations 13.25 and 13.26 demonstrate that the spectral resolution of the diffractive filter of Fig. 13.16 is inversely proportional to the height difference of the diffractive element *along the optical axis*. This is also true for the grating used in the spectrum synthesizer of Fig. 13.13. The conclusion is that the synthesizer and the diffractive filter have similar resolution provided that their diffractive elements have the same size and incident angle.

Diffractive filters and synthesizers have similar resolution and synthesizing capabilities, but there are large differences in optical efficiency. In the synthesizer the losses at different wavelengths are independent, so the loss in each spectral band can be optimized to give the desired synthetic spectrum with the minimum required loss. In the diffractive filter, on the other hand, the losses at one wavelength will in general depend on losses at other wavelengths in a complex manner. The consequence is that, even though any spectrum can be synthesized to within a constant, the losses are larger than the minimum required to replicate the desired spectral distribution.

To understand the origin of the extra loss, we model the diffractive surface as a beam splitter that separates the incident light into N spatially separate channels, and then recombines them into a single output channel after having imparted a phase shift on each channel. To simplify the argument we assume that the channels are of equal strength. We learned in Chapter 2 that such recombination onto a single output leads to $1/N$ loss on average over all frequencies (or over one FSR of a periodic spectrum). Therefore only wavelengths that are reflected in phase from all N reflectors of the diffractive element are completely transmitted by the filter. In other words, we can synthesize a spectrum with one perfectly-transmitted peak within the FSR of the filter. (Note that this leads to the prescribed $1/N$ loss on average across the FSR.)

Now consider what happens if we try to synthesize a spectrum with M peaks within the FSR. The input is first split into N channels. The best we can do is to group the N channels into M subsections that each creates one spectral peak. Each of these M subsections only handles $1/M$ of the input power and each has a $1/M$ coupling loss due to mode mismatch. The maximum value of each peak is then $1/M^2$. If we interleave the sub sections so that the total path length difference, and therefore the resolution, is close to the same as for the whole diffractive element, then the total loss averaged over the FSR is $1/(M^2N)$.

It is the extra loss of $1/M^2$ caused by power splitting and mode-mismatch loss that separates the diffractive filter from the spectral synthesizer. This extra loss becomes prohibitive when the number of peaks in the desired spectrum is large. The synthesizer geometry should therefore be used for applications that require complex spectra, while the diffractive filter is a simpler, and therefore better, alternative in applications where the synthesized spectrum has only one , or at most a few, peaks.

The resolution, given by Eqs. 13.25 and 13.26, and the extra loss have important implications for the MEMS implementation of diffractive spectrometers/filters. Spatial light modulators designed for normal incidence are only useful for low-resolution applications [39]. Most spectroscopy applications require better resolution than what we get from the height of practical MEMS structures by themselves, so gracing incidence and large diffraction angles are necessary. One way to achieve high-diffraction efficiency at gracing incidence is shown conceptually in Fig. 13.15. In this implementation the MEMS diffractive element behave much like a tunable blazed grating. It consists of a series of tilted reflectors so that the diffraction angle is high, and it is operated such that all reflections are in-phase in the optical pass band. Such MEMS structures can be fabricated by a combination of anisotropic etching and DRIE [40], and have been demonstrated as amplitude filters and tunable WDM interleavers [41]. Alternatively, each reflector can be replaced by a diffractive structure with high diffraction angle [42].

13.6 Tunable Lasers

We have seen in this chapter that miniaturization and microfabrication enable compact and functional optical filters that cannot practically be created using traditional technologies. In addition, MEMS technology provides not only the means for fabricating optical components, but also a substrate for system integration and packaging. A compelling approach is therefore to create complete systems-on-a-chip solutions that include both MEMS optical filters and MEMS alignment and integration structures.

One application that benefit from such an approach is the tunable semiconductor laser. In this section we will consider the two most common MEMS tunable lasers geometries: (1) Vertical Cavity Surface Emitting Lasers with a built-in MEMS Fabry-Perot filter for direct tuning of the cavity modes, and (2) External Cavity Semiconductor Diode Laser (ECSDL) with MEMS optical filters.

In addition to these common designs, MEMS enable a number of simpler solutions with less general tuning characteristics. One such design with intermediate complexity and proven market potential is to use an array of fixed-wavelength semiconductor lasers and select the output of the one with the most appropriate wavelength, using a MEMS mirror like the ones described in Chapter 7 for the se-

lection. If the laser selection is combined with temperature tuning of the individual lasers, this approach achieves continuous tuning [43].

13.6.1 MEMS Vertical Cavity Surface Emitting Lasers

Perhaps the simplest and most elegant way to make a tunable laser is to place the laser gain medium in a very short F-P cavity that can be tuned by moving one mirror with respect to the other. This is the approach that is taken in the MEMS tunable Vertical Cavity Surface Emitting Laser (VCSEL) shown schematically in Fig. 13.17.

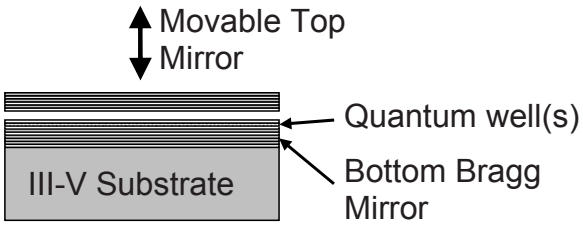


Figure 13.17. Cross section of MEMS tunable VCSEL. The laser cavity is formed by two high-reflectivity Bragg mirrors. Between the mirrors is a thin gain region with one or more quantum wells, and an air-gap that allows the upper Bragg mirror to be moved to tune the wavelength of the cavity mode(s). The structures supporting current injection and MEMS actuation are not included in the schematic.

The fabrication of the tunable VCSEL starts with the growth of a Bragg mirror on a III-V semiconductor substrate. Following the formation of the bottom mirror, the gain medium, consisting of one or more quantum wells, a spacer layer, and finally the top Bragg mirror are grown. The spacer layer is then sacrificially etched to allow the upper mirror to move vertically under the control of a MEMS actuator.

The Bragg reflectors are lattice-matched semiconductor mirrors, most typically fabricated using Molecular Beam Epitaxy (MBE). Tunable F-P filters based on MBE-grown reflectors have been demonstrated in AlGaAs for the 850 nm wavelength [44,45], and in InP-based materials for the 1550 nm wavelength band [46,47,48]. Excellent mirrors can be made in both bands, but it is challenging to integrate the air gap and MEMS actuators with the current-injection and mode-profile-control structures required in VCSELs, and the laser cavity has to be designed with care to avoid detrimental effects from internal reflections from the semiconductor-airgap interfaces. A good historical account and in-depth description of MEMS tunable VCSELs are given in ref. [49].

13.6.2 MEMS External Cavity Semiconductor Diode Lasers

A F-P cavity laser with a movable mirror can only be tuned over a wavelength range that is less or equal to the spacing of the cavity modes, given by the Free Spectral Range (FSR) of the cavity. These modes coincide with the transmission peaks of the cavity. (We leave it to the reader to prove that the internal field in the cavity is maximized at the wavelengths of the transmission maxima, thus making these the cavity-mode wavelengths.) From Eq. 13.2 we know that the wavelengths of the transmission maxima are given by

$$\lambda_C = \frac{2L}{m} \tag{13.27}$$

where m is an integer. The cavity mode spacing is then

$$\frac{d}{dm}\lambda_C = -\frac{2L}{m^2} \Rightarrow \frac{\Delta\lambda_C}{\lambda_C} = \frac{\lambda_C}{2L} \tag{13.28}$$

Good semiconductor lasers have broad-band gain of as much as 2% relative bandwidth in the 1550 nm wavelength band. To achieve mode-hop free tuning over such a broad band, we must restrict the laser cavity length to values fulfilling the inequality

$$L < \frac{\lambda_C}{2\dfrac{\Delta\lambda_C}{\lambda_C}} \approx 25 \cdot \lambda_C \tag{13.29}$$

Practical VCSELs have cavity lengths that are substantially shorter than this, so they can indeed be tuned over their full gain bandwidth by simply moving one cavity mirror. Edge emitting semiconductor lasers, on the other hand, are almost always much longer than the values given by Eq. 13.29, so tunable edge emitters must be designed to not only have tunable cavity modes, but must also include optical filters that suppress all but one of these cavity modes, so that single-mode operation over a broad optical wavelength band can be achieved. Tunable edge emitters are therefore substantially more complex than the simple tunable VCSEL, but they also achieve higher power and better wavelength stability due to their longer gain regions and cavities.

Two different ECSDL geometries capable of broad-band tuning are shown in Figs. 13.18 and 13.19. The traditional Littrow configuration of Fig. 13.18 consists of a semiconductor gain medium with a single-mode wave guide, a collimating lens, and a grating that acts as a wavelength filter. The front facet of the gain medium is Anti-Reflection (AR) coated to establish an optical cavity between the back mirror of the semiconductor diode and the diffraction grating and to prevent spurious lasing of the semiconductor diode by itself. The output of the single-mode

waveguide is collimated onto a diffraction grating that retro-reflects the mode back into the waveguide.

The combination of the grating and the back mirror of the semiconductor diode creates two interacting filters. The grating itself is a wavelength filter that only diffracts light in a narrow wavelength band back into the semiconductor waveguide. The second filter is the cavity formed by the grating and the diode back mirror. Both these filters will be tuned when the grating is moved. The center passband wavelength of the grating filter depends on the angle of the grating, so it is tuned by rotation. The cavity modes are determined by the length of the cavity, so their center wavelengths are tuned by translation of the grating along the optical axis[h].

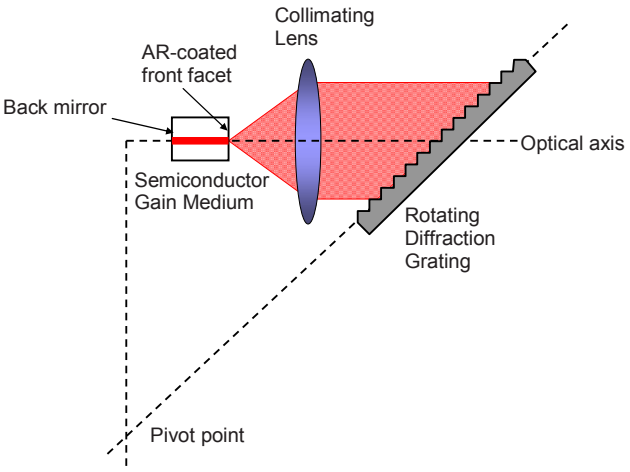


Figure 13.18. Schematic diagram of the traditional Littrow configuration for tunable external cavity lasers. The optical field of the single-mode waveguide in the gain medium is collimated onto a grating that acts as a filter and diffracts a narrow band of wavelengths back into the gain medium. All but one cavity mode is then suppressed and prevented from lasing. The wavelength of the lasing mode is controlled by controlling the angle of the grating with respect to the optical axis. If the grating is rotated around the correct pivot point, then the cavity wavelength and the grating-filter wavelength are tuned equally, so that continuous wavelength tuning is achieved.

[h] It seems counterintuitive that translation of the grating perpendicularly to the optical axis does not change the lasing wavelength, but it is nevertheless true to first order. To understand how, consider light at the center wavelength of the grating. If the grating is moved perpendicularly to the axis, the individual reflectors of the grating are still in phase and the wavelength is not changed.

The length of the cavity is many times larger than the spacing of the reflecting surfaces of the grating, and the FSR of the cavity filter is correspondingly smaller. The function of the grating filter is then to pick out one specific cavity mode and suppress all others, so that only the desired mode achieves lasing without an excessively high pumping threshold. In other words, the two filters must be aligned in wavelength, which means that the cavity length has to be controlled with sub-wavelength accuracy. Accurate alignment and cavity length control are therefore necessary and motivates the use of MEMS.

The key to continuous, wide-range, mode-hop-free wavelength tuning, is to simultaneously rotate and translate the grating so that the cavity modes and the grating filter are tuned the same amount, i.e. so that the desired cavity mode stays aligned with the grating filter. It is well known that if the grating is rotated around a pivot point located at the intersection of the plane through the grating surface and the normal to the optical axis at a point that is a distance $n \cdot \lambda_{vac}$ from the rotating element along the optical axis, where n is the number of wavelengths in the cavity and λ_{vac} is the vacuum wavelength, then the cavity mode and grating filter stays aligned during rotation [50,51].

An alternative design, dubbed the Littman configuration, is shown in Fig. 13.19. This laser architecture also uses a grating filter, but instead of sending the light directly back into the semiconductor diode, the grating diffracts the light onto a mirror. Only the narrow wavelength band that hits the mirror at normal incidence is reflected back into the semiconductor diode via the grating. An optical cavity is now established between the external mirror and the back facet of the diode.

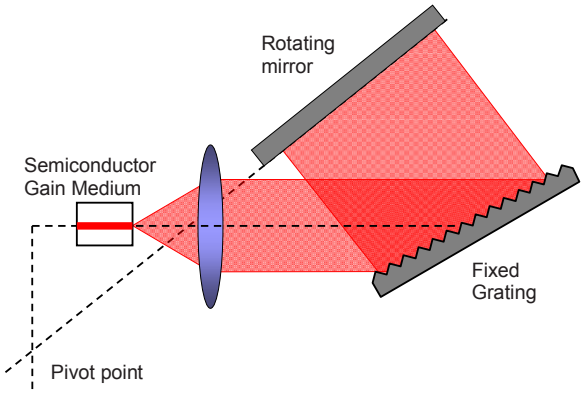


Figure 13.19. The Littman tunable laser is similar to the Littrow configuration, except that the optical mode is diffracted onto a mirror. The cavity modes are then established between the rotating mirror and the back facet of the diode, while the grating filter picks out the narrow wavelength band that is diffracted perpendicularly onto the rotating mirror.

Again we have two interacting filters; the grating filter that is tuned by rotating the mirror, and the cavity filter that is tuned by translating the mirror. Continuous tuning requires that the mirror is rotated around a pivot point at the intersection of the plane through the mirror surface and the normal to the optical axis at a point that is a distance $n \cdot \lambda_{vac}$ from the rotating element along the optical axis, where again n is the number of wavelengths in the cavity and λ_{vac} is the vacuum wavelength.

The advantage of the Littman configuration is that the light is diffracted from the grating twice per round trip of the cavity. The two passes through the grating lead to a shaper transmission function with better side-mode suppression. Lasers in the Littman configuration therefore have a cleaner lasing mode and achieve better coherence, and it is therefore the architecture of choice in industrial developments.

Both the Littman and the Littrow architectures have a fixed pivot point that in principle allows the grating filter and the cavity modes to be tuned together with a single degree of freedom of motion. This compellingly simple solution does not work in practical MEMS implementations, where we need at least one extra degree of freedom for alignment of the cavity mode and grating filter and for compensation of the dispersion in the optical components.

Still it is practical to have one major degree of motion that does most of the tuning, while a second control is used for adjustments. One school of thought in MEMS implementations of Littrow [52,53] and Littman [54] ECSDLs has therefore been to focus on actuator design, with the goal to develop accurate, one-degree-of freedom actuators that can provide stable, mode-hop-free tuning after initial alignment. In practice the one-degree-of-motion actuator has been augmented by a second degree of freedom of motion for initial alignment and dispersion compensation [55,56]. These types of lasers have excellent stability and optical characteristics.

13.6.3 Tunable External Cavity Semiconductor Diode Lasers with Diffractive Filters

The drawback of using a rotating grating or mirror to tune an ECSDL as in Fig. 13.18 and 13.19 is that the required total range of motion is large. Typically the widths of the rotating elements are in the mm range, leading to motion of several hundred micron. Motion on this scale is difficult to support in MEMS actuators, and for that reason, the tunable laser has been a driver in the development of long-range MEMS actuators.

MEMS diffractive filters represent a compelling alternative to traditional designs. As shown in Chapter 13.5.2, MEMS diffractive filters achieve separate control of

the phase and amplitude response of the filter[i]. Therefore when a diffractive filter is used instead of the grating in the ECSDLs of Fig. 13.18, no macroscopic motion of the diffractive element is required for tuning. Tunable lasers based on this principle have been demonstrated both with commercial MEMS SLMs [57] and with diffractive elements design especially for the purpose [58].

The diffractive filter requires only sub-wavelength motion, as opposed to the macroscopic motion of rotating gratings and mirrors. This advantage is somewhat offset by the added complexity of control. Typically a diffractive filter designed for ECSDL tuning requires several dozen reflectors to cover the gain band width of semiconductor diode. Each of these reflectors must be individually controlled to fully utilize the tuning range of the diffractive element. The simple designs with one (in practice two) degree of freedom of Fig. 13.18 and 13.19 are therefore replaced by systems with several dozen degrees of freedom. This means that complexity is moved from the mechanical domain in traditional designs to the digital-electronics domain for the diffractive filters. This is usually a worthwhile trade-off, and it is only getting better as digital electronics improves.

13.7 Summary of Microoptical Filters

This Chapter describes MEMS implementations of traditional filter and spectrometer architectures, as well as several designs that rely for their operation on the characteristics of MEMS technology. The field of optical MEMS filters is too large for a comprehensive coverage, so the emphasis is on the basic principles, as well as the unique advantages and challenges of optical filters implemented in MEMS technology.

Traditional resonant filters like Bragg gratings, Fabry-Perots, and other optical resonators are not conceptually different in MEMS implementations. These structures have one, or at most a few, degrees of freedom of motion, so they do not make use of the potential for complexity that MEMS technology provides. However, both amplitude filters and phase filters (dispersion compensators) benefit from MEMS actuation that provides tunability and from the miniaturization and integration inherent to MEMS fabrication technology.

MEMS spectrometers and spectral synthesizers, on the other hand, take more full advantages of the unique characteristics of MEMS technology. Section 13.4 describes several transform-spectrometer architectures that are designed specifically

[i] This fact follows from the math, but is also simple to understand in physical terms: If the individual microreflectors are all moved in common mode, then only the phase of the reflected light is affected. If the reflectors are moved differentially, then both the amplitude and the phase are changed, but the phase can be corrected by an associated common-mode actuation.

for MEMS implementations. A common trend of these spectrometers is non-traditional signal processing, where the emphasis is on adapting the data handling to the hardware, rather than the other way around.

Optical spectral synthesizers must by definition have large numbers of degrees of freedom to support the generation of general spectra. In section 13.5 we describe two different general filters/synthesizers: The Heritage-Weiner frequency synthesizer and the MEMS diffractive filter. Each of these can make use of MEMS Spatial Light Modulators with many degrees of freedom to perform near-arbitrary spectral phase and amplitude filtering.

The last section of the Chapter is devoted to tunable lasers. The VCSEL with one MEMS actuated mirror is the simplest conceivable, broad-band tunable laser. It is a one-degree-of-freedom device with a F-P tuning filter. MEMS implementations of the traditional Litthrows and Littman external-cavity designs are also in principle simple to control, needing only one degree of freedom in theory (two in practice). The problem with these architectures is that the long range of travel required by the rotating tuning elements is difficult to achieve with standard MEMS devices. The MEMS diffractive filter represents an alternative that requires only sub-wavelength motion of the mechanical elements, at the cost of a sharply increased number of degrees of motion that must be accurately controlled.

Exercises

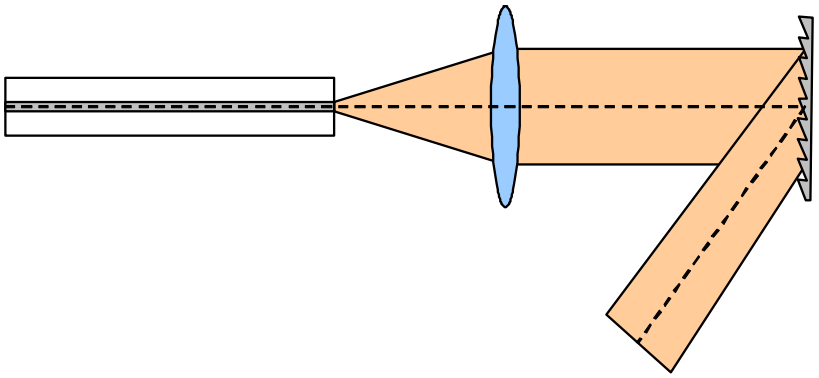
Problem 13.1 - Asymmetric Fabry-Perot

Loss less Fabry-Perot filters have symmetric pass bands, but the symmetry can be broken if we use lossy mirrors.

a. Use the formalism of Chapter 3 together with the metal data of Chapter 7 (Table 7.1) to design a F-P with an asymmetric pass band.
b. How can an asymmetric passband be used? Optimize the asymmetry for a specific application.

Problem 13.2 - Grating Demultiplexer

You are designing a WDM demultiplexer at a center wavelength of *1.55 um*. The dispersive element of the demultiplexer is a reflective blazed grating, on which you collimate the output of the WDM fiber as shown in the figure. Assume that the incident angle is zero, and the diffraction angle of the center wavelength is *60* degrees.



Grating-based WDM multiplexer.

a. What is the periodicity of the grating? (Hint: There might be more than one correct answer. Make sure to list all possibilities.)
b. How large do you have to make the spot on the grating to be able to separate 160 wavelengths with –40 dB cross talk between channels?
c. How can you design a smaller demultiplexer, i.e. what parameters would you change to be able to shrink the size of the beam on the grating?

Problem 13.3 - Time Delay

If we use a scanning mirror as the SLM in Fig. 13.13, we create a variable time delay. This can be understood by noting the fact that the Fourier Transform of a linear function (tilted mirror) in the frequency domain is a temporal shift in the time domain.

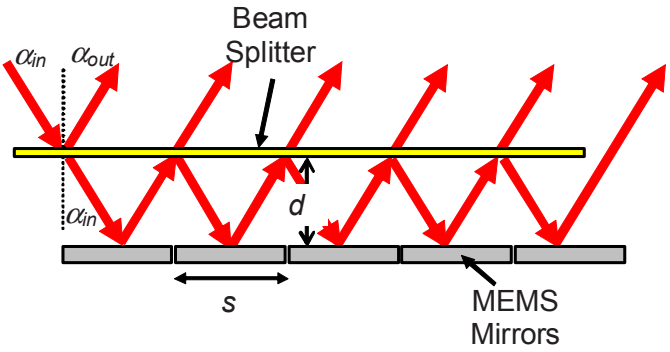
a. Explore this effect by drawing the path of a pulse through the synthesizer with a tilted mirror as the SLM. Assume that the beam is retro reflected after it has passed the grating for a second time, and that it therefore makes a third and fourth pass before it propagates back in the direction of the incident beam.
b. Derive an expression for the time delay in this double pass configuration. Express your answer in terms of the design parameters of the synthesizer (size and relative position of the components, period and order of operation of the grating), as well as the characteristics of the input beam.

Problem 13.4 - Hybrid Spectrometer

a. Design a Fourier-Transform spectrometer based on the time delay of Problem 13.3.
b. The grating in this F-T spectrometer determines the time delay and therefore the resolution of the spectrometer. Compare the F-T resolution to what you would get by using the same grating in a grating spectrometer.

Problem 13.5 - MEMS GT Interferometer

The figure below shows a cross section of a MEMS Gires-Tournois interferometer. The output of the GTI is created by interference of the partially-transmitted beams coming off the beam splitter. The phases of these beams are controlled by the MEMS mirror that can move vertically.



MEMS GT interferometer. We assume that the micromirrorarray has unity fill factor, i.e. that the mirror size is equal to the mirror spacing. This is only approximately true in practical devices.

We would like to design the GTI to have as many MEMS mirrors as possible to have the best possible control of the optical output field. The maximum number of mirrors will be limited by diffraction of the input beam. Assume that the input beam is Gaussian and focused at the halfway point through the interferometer and that all the mirrors of the arrays are identical. The minimum mirror size is then

$$s = k \cdot \frac{\omega}{\cos \alpha_{in}}$$

where ω is the Gaussian beam radius at the first mirror, and k is a constant that has to be on the order of 3 to avoid excessive cross talk between consecutive reflections.

a. Using standard formulas for Gaussian beam propagation, express the mirror size in terms of the total length of propagation through the GTI.
b. Solve for the total length and maximize.
c. Find the maximum number of mirrors in terms of the input angle, the mirror size, the wavelength, and the factor k.
d. What input angle maximizes the number of mirrors? (You will find that the maximizing angle is one that is also important in a very different area of MEMS.)

Problem 13.6 - ECL Pivot Point

a. Prove the assertion made in section 13.6.2 that an External Cavity Laser can exhibit mode-hop-free tuning if the pivot point of the grating is chosen correctly.
b. Design a compact MEMS actuator that mimics the rotation around the ideal pivot without having to extend all the way to that pivot point.
c. Show that the grating, and therefore also the pivot point, may slide perpendicularly to the optical axis without causing mode hops.
d. Can you design a MEMS actuator that takes advantage of this extra degree of freedom (translation perpendicularly to the optical axis) to simplify the implementation of the ECL?

References

1 J.H. Jerman, S.R. Mallinson, “A miniature Fabry-Perot Interferometer fabricated using silicon micromachining techniques”, 1988 Solid State Sensor and Actuator Workshop. Technical Digest, 6-9 June 1988, Hilton Head Island, SC, USA; p.16-18.

2 A. T. T. D. Tran, Y. H. Lo, Z. H. Zhu, D. Haronian, E. Mozdy,” Surface Micromachined Fabry-Perot Tunable Filter“, IEEE Photonics Technology Letters, Vol. 8, No. 3, March 1996, pp. 393-395.

3 K. Cao, W. Liu, J.J. Talghader, “Curvature Compensation in Micromirrors with High-Reflectivity Optical Coatings Flat-free-standing dielectric mirrors”, Journal of Microelectromechanical Systems, Vol. 10, No. 3, September 2001, pp. 409-417.

4 D. Hohlfeld, H. Zappe, "All-dielectric tunable optical filter based on the thermo-optic effect", Journal of Optics A: Pure and Applied Optics, vol. 6, no. 6, 2004, pp. 504-511.

5 E.C. Vail, M.S. Wu, G.S. Li, L. Eng, C.J. Chang-Hasnain ,” GaAs micromachined widely tunable Fabry-Perot filters“, Electronics Letters, 2nd February, 1995 vol. 31, no. 2, pp. 228-229.

6 M. C. Larson, B. Pezeshki, Member, IEEE, J. S. Harris, Jr., “Vertical Coupled-Cavity Microinterferometer on GaAs with Deformable-Membrane Top Mirror”, IEEE Photonics Technology Letters, Vol. 7, No. 4, April 1995, pp. 382-384.

7 A. Spisser, R. Ledantec, C. Seassal, J. L. Leclercq, T. Benyattou, D. Rondi, R. Blondeau, G. Guillot, P. Viktorovitch, “Highly Selective and Widely Tunable 1.55-um InP/Air-Gap Micromachined Fabry–Perot Filter for Optical Communications”, IEEE Photonics Technology Letters, vol. 10, no. 9, September 1998, pp. 1259-1261.

8 P. Tayebati, P. Wang, M. Azimi, L. Maflah, D. Vakhshoori, "Microelectromechanical tunable filter with stable half symmetric cavity", Electronics Letters, 1st October 1998 vol. 34, no. 20, pp. 1967-1968.
9 M. Garrigues, J. Danglot, J.-L. Leclercq, O. Parillaud, "Tunable High-Finesse InP/Air MOEMS Filter", IEEE Photonics Technology Letters, Vol. 17, No. 7, July 2005, pp. 1471-1473.
10 C.J. Chang-Hasnain, "Tunable VCSEL", IEEE Journal of Selected Topics in Quantum Electronics; Nov.-Dec. 2000; vol.6, no.6, p.978-87
11 M.-C.M. Lee, M.C. Wu, "Variable bandwidth of dynamic add-drop filters based on coupling-controlled microdisk resonators," Optics Letters, vol. 31, no. 16, pp. 2444–2446, 2006.
12 F. Gires and P. Tournois, "Interferometer utilizable pour la compression d'impulsions lumineuses modules en frequence," C.R. Academie Sci., vol. 258, pp. 6112-6115, 1964.
13 K. Goossen, J. Walker, and S. Arney, "Silicon modulator based on mechanically-active antireflection layer with 1Mbit/sec capability for fiber-in-the-loop applications," IEEE Photon. Technol. Lett., vol. 6, pp. 1119-1121, Sept. 1994.
14 C. K. Madsen, J. A. Walker, J. E. Ford, K. W. Goossen, T. N. Nielsen, and G. Lenz, "A tunable dispersion compensating MEMS all-pass filter," IEEE Photon. Technol. Lett., vol. 12, pp. 651-653, Jun. 2000.
15 K. Yu, O. Solgaard, "MEMS optical wavelength deinterleaver with continuously variable channel spacing and center wavelength," IEEE Photon. Technol. Lett., vol. 15, pp. 425-427, Mar. 2003.
16 K. Yu, O. Solgaard, "Tunable optical transversal filters based on a Gires-Tournois interferometer with MEMS phase shifters," IEEE J. of Sel. Topics in Quantum Electronics, vol. 10, pp. 588-597, May 2004.
17 K. Yu and O. Solgaard, "Tunable chromatic dispersion compensators using MEMS Gires-Tournois interferometers," IEEE/LEOS International Conference on Optical MEMS, Lugano, Switzerland; pp. 181-182, Aug. 2002.
18 Wikipedia at http://en.wikipedia.org/wiki/Rectangular_function
19 P.R. Griffiths, J.A. de Haseth, "Foiurier Transform Infrared Spectroscopy", Wiley, 1986.
20 O. Manzardo, H.P. Herzig, C.R. Marxer, N.F. de Rooij, "Miniaturized time-scanning Fourier transform spectrometer based on silicon technology", Optics Letters, vol. 24, no. 23, December 1, 1999, pp. 1705-1707.
21 Yu, K., Lee, D., Krishnamoorthy, U., Park, N., Solgaard, O., *Transducers*, Seoul, June 2005.
22 Hans Peter Herzig, GLV transform spectrometer: Manzardo, O., Michaley, R., Schadelin, F., Noell, W., Overstolz, T., De Rooij, N., Herzig, H. P., *Opt. Lett.* **29,** 1437-1439 (2004).
23 Kung, H. L., Bhalotra, S. R., Mansell, J. D., Miller, D. A. B., Harris, J. S., *IEEE J. Sel. Topics Quantum Electron.*, **8**, 98-105 (2002).

24 J.P. Heritage, A.M. Weiner, R.N. Thurston, "Picosecond Pulse Shaping by Spectral Phase and Amplitude Manipulation," Optics Letters **10**, 609-611 (1985).

25 K. Yu, D. Lee, N. Park, O. Solgaard, "Tunable Optical Bandpass Filter with Variable-Aperture MEMS Reflector", Journal of Lightwave Technology, vol. 24, no. 12, December 2006, pp. 5095-5102.

26 A. M. Weiner, J. P. Heritage, J. A. Salehi, "Encoding and Decoding of Femtosecond Pulses," *Optics Letters,* vol. 13, no. 4, April 1988.

27 J.P. Heritage, A.M. Weiner, R.N. Thurston, Rn, "Picosecond pulse shaping by spectral phase and amplitude manipulation", Optics Letters; Dec. 1985; vol.10, no.12, pp.609-11.

28 G.J. Tearney, B.E. Bouma, J.G. Fujimoto, "High-Speed Phase- and Group-Delay Scanning with a Grating-Based Phase Control Delay Line," *Optics Letters* **22**, 1811-1813 (1997).

29 Y. Pan, H. Xie, G.K. Fedder, "Endoscopic optical coherence tomography based on a microelectromechanical mirror", Optics Letters; 15 Dec. 2001; vol.26, no.24, p.1966-8.

30 K.T. Cornett, P.M. Hagelin, J.P. Heritage, O. Solgaard, M. Everett, "Miniature Variable Optical Delay using Silicon Micromachined Scanning Mirrors," *CLEO 2000*, San Francisco, CA, 383-384, (2000).

31 P.M. Hagelin, U. Krishnamoorthy, J.P. Heritage, O. Solgaard, "Scalable Optical Cross-Connect Switch Using Micromachined Mirrors", IEEE Photonics Technology Letters, vol. 12, no. 7, pp. 882-885, July 2000.

32 J.E. Ford, V.A. Aksyuk, D.J. Bishop, J.A. Walker, "Wavelength add-drop switching using tilting micromirrors", Journal of Lightwave Technology; May 1999; vol.17, no.5, p.904-11.

33 R.A. Deverse, R.M. Hammaker, W.G. Fateley, "Realization of the Hadamard Multiplex Advantage Using a Programmable Optical Mask in a Dispersive Flat-Field Near-Infrared Spectrometer", Applied Spectroscopy, Vol. 54, No. 12, 2000, pp. 1751-1758.

34 M.B. Sinclair, M.A. Butler, A.J. Ricco, and S.D. Senturia, "Synthetic spectra: a tool for correlation spectroscopy", *Applied Optics*, vol. 36, no. 15, pp. 3342-48, May 20, 1997.

35 R. Belikov, X. Li, O. Solgaard, "Programmable Optical Wavelength Filter Based on Diffraction From a 2-D MEMS Micromirror Array," Conference on Lasers and Electro-Optics (CLEO), Technical Digest, Baltimore, MD, June 1-6, 2003.

36 R. Belikov, C. Antoine-Snowden, O. Solgaard, "Femtosecond Direct Space-to-Time Pulse Shaping with MEMS Micromirror Arrays," Proc. 2003 IEEE/LEOS International Conf. on Optical MEMS, Waikoloa, Hawaii, 18-21 August 2003, pp. 24-25.

37 R. Belikov, O. Solgaard, "Optical Wavelength Filtering by Diffraction from a Surface Relief", Optics Letters vol. 28, No. 6, March 15, 2003, pp.447-449.

38 R. Belikov, "Diffraction-Based Optical Filtering: Theory and Implementation with MEMS", Doctoral Dissertation, Stanford University, UMI Dissertation Publishing, 2005.
39 G. B. Hocker *et al.*, "The Polychromator: A Programmable MEMS Diffraction Grating for Synthetic Spectra," *Tech. Dig., Solid-State Sensor and Actuator Wksp.*, Hilton Head, SC, 4–8 June 2000, pp. 89–92.
40 X. Li, C. Antoine, D. Lee, J.-S. Wang, O. Solgaard, "Tunable Blazed Gratings", Journal of Microelectromechanical Systems, vol. 15, no. 3, June 2006, pp. 597-604.
41 C. Antoine, X. Li, J.-S. Wang, O. Solgaard, "Reconfigurable Optical Wavelength Multiplexer Using a MEMS Tunable Blazed Grating", Journal of Lightwave Technology, vol. 25, no. 10, October 2007, pp. 3100-3107.
42 H. Sagberg, M. Lacolle, I-R. Johansen, O. Løvhaugen, R. Belikov, O. Solgaard, A. Sudbø, "Micromechincal Gratings for Visible and Near-Infrared Spectroscopy", IEEE Journal of Selected Topics in Quantum Electronics, vol. 10, no. 3, May/June, 2004, pp. 604-613.
43 B. Pezeshki, E. Vail, J. Kubicky, G. Yoffe, S. Zou, J. Heanue, P. Epp, S. Rishton, D. Ton, B. Faraji, M. Emanuel, X. Hong, M. Sherback, V. Agrawal, C. Chipman, T. Razazan, "20mW widely tunable laser module using DFB array and MEMS selection", Technical Digest of the 2002 Optical Fiber Communication Conference (OFC 02).
44 E. C. Vail, M. S. Wu, G. S. Li, L. Eng, and C. J. Chang-Hasnain, "GaAs micromachined widely tunable Fabry–Pérot filters," Electron Lett., vol. 31, no. 3, pp. 228–229, Feb. 1995.
45 M. C. Larson, B. Pezeshki, and J. S. Harris, "Vertical coupled-cavity microinterferometer on GaAs with deformable-membrane top mirror," IEEE Photon. Technol. Lett., vol. 7, no. 4, pp. 382–384, Apr. 1995.
46 A. Spisser, R. Ledantec, C. Seassal, J. L. Leclercq, T. Benyattou, D. Rondi, R. Blondeau, G. Guillot, and P. Viktorovitch, "Highly selective and widely tunable 1.55-μm InP/air-gap micromachined Fabry–Pérot filter for optical communications," IEEE Photon. Technol. Lett., vol. 10, no. 9, pp. 1259–1261, Sep. 1998.
47 P. Tayebati, P. Wang, M. Azimi, L. Maflah, and D. Vakhshoori, "Microelectromechanical tunable filter with stable half symmetric cavity," Electron. Lett., vol. 34, no. 20, pp. 1967–1968, Oct. 1998.
48 M. Garrigues, J. Danglot, J. L. Leclercq, and O. Parillaud, "Tunable high-finesse InP/Air MOEMS filter," IEEE Photon. Technol. Lett., vol. 17, no. 7, pp. 1471–1473, Jul. 2005.
49 C. J. Chang-Hasnain, "Tunable VCSEL," IEEE J. Sel. Topics Quantum Electron., vol. 6, no. 6, pp. 978–987, Nov./Dec. 2000.
50 M.G. Littman, H.J. Metcalf, "Spectrally narrow pulsed dye laser without a beam expander," *Appl. Opt.*, vol. 17, pp. 2224–2227, 1978.
51 W.R. Trutna, L.F. Stokes, "Continuously tuned external cavity semiconductor laser," *J. Lightwave Technol.*, vol. 11, pp. 1279–1286, 1993.

52 R.R.A. Syms, A. Lohmann, ”MOEMS Tuning Element For A Littrow External Cavity Laser”, Journal of Microelectromechanical Systems, vol. 12, no. 6, December 2003, pp. 921-928.
53 X. M. Zhang, A.Q. Liu, C. Lu, D. Y. Tang, ”Continuous Wavelength Tuning in Micromachined Littrow External-Cavity Lasers”, IEEE Journal of Quantum Electronics, vol. 41, no. 2, February 2005, pp. 187-197.
54 W. Huang, R.R.A. Syms, J. Stagg and A. Lohmann, “Precision MEMS flexure mount for a Littman tunable external cavity laser”, IEE Proc.-Sci. Meas. Technol. vol. 151, no. 2, March 2004, pp.67-75.
55 H. Jerman, J.D. Grade, “A Mechanically-Balanced, DRIE Rotary Actuator For A High-Power Tunable Laser”, Proceedings of the 2002 Solid-State Sensors and Actuators Workshop, Hilton Head, South Carolina, June, 2002.
56 D. Anthon, D. King, J.D. Berger, S. Dutta, A. Tselikov, ”Mode-hop free sweep tuning of a MEMS tuned external cavity semiconductor laser”, Conference on Lasers and Electro-Optics (CLEO), Technical Digest, San Francisco, CA, May 16-21, 2004, paper CWL2.
57 R. Belikov, C. Antoine-Snowden, O. Solgaard, “Tunable external cavity laser with a stationary deformable MEMS grating.,” Conference on Lasers and Electro-Optics (CLEO), Technical Digest, San Francisco, CA, May 16-21, 2004, paper CWL3.ï
58 C. Antoine, X. Li, D. Sesko, O. Solgaard, “An External Cavity Tunable Laser with a Low-Loss Narrowband MEMS Tunable Blazed Grating”, 2007 IEEE/LEOS Annual Meeting Conference Proceedings, Lake Buena Vista, Florida, 21-25 October 2007, pp. 834-835.

14: Photonic Crystal Fundamentals

14.1 Introduction to Photonic Crystals

Photons are very hard to manipulate and control, because most materials either absorb[a] the photon or interact with it only weakly. In earlier chapters of this book we have seen how transparent, weakly-interacting materials can be used to build lenses and waveguides, and we have learned how to use absorptive, strongly-interacting materials (metals) as mirrors and gratings. A third method for controlling optical fields is to use many weak interactions that combine coherently to a strong interaction. This approach to control of electromagnetic radiation is the basis for Photonic Crystals (PCs).

We have already looked at one simple type of Photonic Crystal. The Bragg reflectors covered in Chapters 3.4.2 and 6.6, and briefly in Chapter 13.2.2, are one-dimensional PCs. These Bragg gratings consist of periodic layers of alternating high and low refractive index. The periodicity is on the order of the wavelength of the light (half the wave length to be exact). This concept can be generalized to two and three dimensions, so we may loosely define Photonic Crystals as periodic structures in one, two, or three dimensions with periodicities on the order of the wavelength of electromagnetic radiation. Photonic Crystals can therefore take any size depending on the wavelength.

In current practice, Photonic-Crystal technology has been developed for Radio Frequencies with wavelengths up to tens of centimeter, and for visible and near-infrared light with wavelengths from 0.4 μm to 2 μm. It is a fortuitous fact that the size range of PCs for visible and near-IR radiation is compatible with standard MEMS fabrication and packaging. This creates opportunities for integration of MEMS and Photonic Crystals, enabling new and improved photonic devices, some of which are described in Chapter 15.

In this chapter we introduce the fundamentals of Photonic-Crystal theory and practice. The field is very large and well documented [1,2,3,4,5], so our goal is

[a] Electronic buffs will maintain that the best way to control a photon is to absorb it and convert it to an electronic excitation!

not a comprehensive overview, but rather a focused introduction with emphasis on concepts and technologies that are most important and exciting to designers of microoptical systems and optical MEMS.

We start by describing at the band structure of PCs and how their band-gaps can be used to create very compact waveguides and resonators for integrated photonics. Then we switch our attention to the use of one and two-dimensional PCs that are of special interest because of their relatively simple fabrication. We cover the concept of Guided Resonance, and describe a simple, analytical theory for how it changes the flow of light through 2-D PCs. With the help of the theory, we show how PCs can be used to create a variety of optical components for miniaturized, free-space optics. We finish the Chapter with a brief section that compares and contrast Photonic Crystals and their role in optics, to natural crystals and their role in electronics.

14.2 Photonic Crystal Basics

The examples in Fig. 14.1 illustrate the variety of Photonic Crystals. Structures of one, two, or three dimensions have very different characteristics, and also present very different design and fabrication challenges. All the crystal structures of Fig. 14.1 are important in microphotonics, some directly and some in supporting roles.

One-dimensional Bragg reflectors are used as high-quality mirrors in sophisticated Optical MEMS, but represent a fabrication challenge due to the thermal stresses caused by different thermal expansion coefficients of the layers of the stack. These fabrication challenges can be overcome, but it complicates and increases the cost of manufacture, so Bragg Mirrors are used sparingly in microphotonics. The exceptions are VCSELs (Chapter 13.6.1) and other III-V devices. Bragg mirrors are important components in the supporting structures that interface the microphotonics to the external world, however. The same is true for another technologically important 1-D Photonic Crystal, the Fiber Bragg Filter or Fiber Bragg Sensor (not shown in Fig. 14.1 – see Chapter 6.6).

One-dimensional, planar gratings and two-dimensional PC slabs are straightforward to integrate with ICs and MEMS, because of their planar geometry and size compatibility. These structures therefore form the basis of most experimental demonstrations and product developments of PC microphotonics. One of their compelling properties is that they can be designed to have many of the desirable characteristics of Bragg reflectors, and may therefore be used as replacements for multilayer-stacks in integrated optics and microphotonics.

The Holey Fiber [6,7] is a low-dispersion medium that allows practical delivery of femto-second laser pulses to microphotonic components and subsystems. It is fabricated by stacking quartz tubes in a hexagonal, or other configuration, and draw-

ing the stack in a fiber-drawing tower to achieve the correct dimensions. In the low-dispersion Holey Fiber shown in Fig. 14.1, the center tube is removed from the stack to produce a center void that supports one or more guided modes due to the surrounding photonic-bandgap material. Photonic Crystal fibers with solid centers can also be made for various applications, but they don't have the low dispersion of the Holey Fiber.

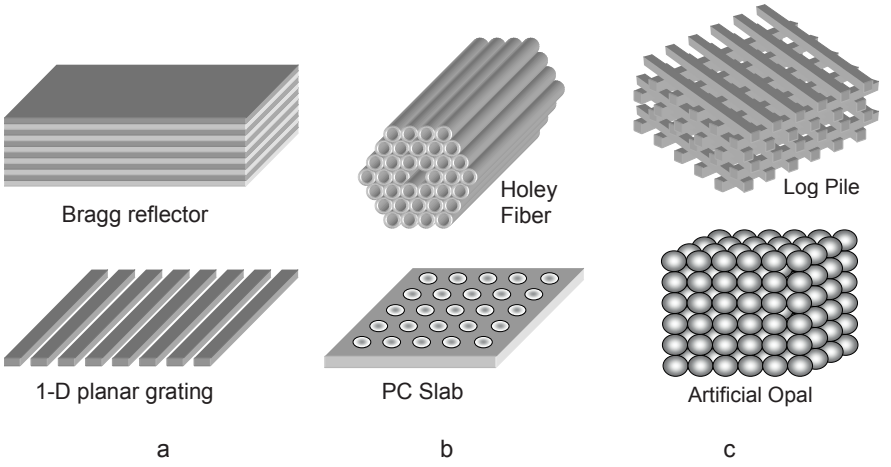


Figure 14.1 *Photonic Crystals of one (a), two (b), and three (c) dimensions. The Bragg reflector and Holey fiber play important supporting roles in many microphotonics systems. The 1-D planar grating and the 2-D PC slab are simple to fabricate and integrate with electronics and MEMS. They are the building blocks of most of the PC MEMS devices described in Chapter 15. The log pile, artificial opal and other 3-D Photonic Crystals hold high promise, but are difficult and expensive to fabricate with the accuracy required by typical optics applications.*

Three-dimensional Photonic Crystals represent the ultimate in photon control. Their structure enable the formation of complete (i.e. omni-directional) band gaps, for complete photon confinement. The difficulty is to fabricate 3-D PCs with sufficient uniformity and dimensional control. The artificial opal of Fig. 14.1 is one promising approach. It is made by self-assembly [8,9] of monodisperse spheres. After assembly the opal may be "inverted" by filling the voids between the spheres by a high-index material. Another approach is to build the photonic crystal layer by layer, either by sequential deposition and patterning [10,11] or by self-aligned etching [12,13]. Alignment and uniformity are difficult to control, however, so these techniques typically cannot make large enough number of layers to take full advantage of the omni-directional bandgap. As fabrication technology is improved by conceptual breakthrough and improved practice, this is likely to change and we'll see an increased use of optical devices based on 3-D PCs.

14.2.1 1-D Photonic Crystals

Photonic Crystals of one, two, and three dimensions have very different characteristics and uses, but many of the underlying concepts are the same, so it is useful to consider the simplest case; the 1-D photonic crystal. We will modify the results of Chapter 6.6, where we used coupled-mode theory to model a periodic wave guide, to find the propagating electromagnetic waves at normal incidence on the Bragg Grating of Fig. 14.2.

This simple 1-D Photonic Crystal is fully described by its average refractive index (n), its index difference (Δn), and its period (Λ). We are only considering normal incidence, so there will be no dependence on the directions of the fields as required by the scalar theory of Chapter 6.6. (In 6.6 we considered only TE wave guide modes, but the treatment is readily modified to cover TM.)

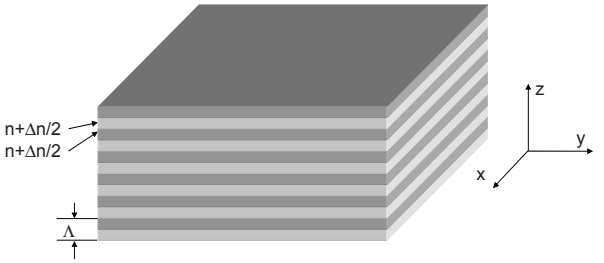


Figure 14.2 1-D Photonic Crystal of period Λ and index variation Δn.

We assume that the corrugation has a square-wave shape that can be expressed as a series

$$\Delta n^2(z) = \Delta n^2 \frac{-j}{m\pi} \sum_m e^{j\frac{2m\pi \cdot z}{\Lambda}} \tag{14.1}$$

where the summation is over all odd-integer values of m. Only modes that are close to phase matched will experience significant coupling, so we may ignore all terms of the series other than the one that has a period close to half the wave length.

As in Chapter 6.6, we will explore solutions of the form

$$E_{x,y} = A(z)u(x,y)\exp\left[j(\omega \cdot t - \beta \cdot z)\right] + B(z)u(x,y)\exp\left[j(\omega \cdot t + \beta \cdot z)\right] \tag{14.2}$$

where A and B are the amplitudes of the forward and backward propagating waves, and $u(x)$ is the mode profile. The direction of the e-field is in the plane of the layers of the Bragg grating. Because of symmetry, it is immaterial if we consider the x or the y component. Using coupled-mode theory we can write the expressions

$$\frac{dA}{dz} = B \cdot \frac{\omega \cdot \varepsilon_0}{4m\pi} e^{j2\beta \cdot z} e^{-j\frac{2m\pi \cdot z}{\Lambda}} \cdot \Delta n^2 \int_{-\infty}^{\infty} u^2(x) dx \tag{14.3}$$

and

$$\frac{dB}{dz} = A \cdot \frac{\omega \cdot \varepsilon_0}{4m\pi} e^{-j2\beta \cdot z} \cdot e^{j\frac{2m\pi \cdot z}{\Lambda}} \cdot \Delta n^2 \int_{-\infty}^{\infty} u^2(x) dx \tag{14.4}$$

These equations are of the form

$$\frac{dA}{dz} = K_m^* B e^{j2\Delta\beta \cdot z} \tag{14.5}$$

$$\frac{dB}{dz} = K_m A e^{-j2\Delta\beta \cdot z} \tag{14.6}$$

where

$$K_m = \frac{\omega \cdot \varepsilon_0}{4m\pi} \Delta n^2 \int_{-\infty}^{\infty} \cdot u^2(x) \cdot dx \tag{14.7}$$

$$\Delta\beta = \beta - \frac{m\pi}{\Lambda} \tag{14.8}$$

Note that the parameter K_m, which we will call the coupling coefficient, depends on the harmonic number m, such that higher-order harmonics are more weakly coupled.

The solutions to Eqs. 14.5 and 14.6 are the amplitudes of the forward and backward propagating waves of the general Bragg grating at normal incidence. The propagation constants of these waves are

$$\beta_{bragg} = \beta - \Delta\beta \pm j\sqrt{K_m^2 - \Delta\beta^2} = \frac{m\pi}{\Lambda} \pm j\sqrt{K_m^2 - \left(\beta - \frac{m\pi}{\Lambda}\right)^2} \tag{14.9}$$

The real and imaginary parts of this expression are plotted in Fig. 6.20. Here we want to emphasize the band gap in photon energy, so we solve the expression for β, which is the propagation constant of the unperturbed wave and therefore proportional to the photon energy

$$\beta = \frac{m\pi}{\Lambda} \pm \sqrt{\left(\beta_{bragg} - \frac{m\pi}{\Lambda}\right)^2 + K_m^2} \tag{14.10}$$

This expression is plotted in Fig. 14.3 for *m=1*. We see that at wavelengths far from resonance $\left(\left(\beta_{bragg} - \frac{\pi}{\Lambda}\right) >> K_1\right)$, we have $\beta \approx \beta_{bragg}$, i.e. waves propagate through the Bragg reflector as if it was a homogeneous medium with an average refractive index. In this range of wavelengths the layered structure does not lead to coherent interference of multiple reflections.

On or close to resonance $\left(\left(\beta_{bragg} - \frac{m\pi}{\Lambda}\right) << K_1\right)$, the Bragg propagation constant is independent of the unperturbed propagation $\left(\beta = \frac{m\pi}{\Lambda} \pm K_1\right)$. There is a forbidden range of propagation constants from $\left(\frac{\pi}{\Lambda} - K_1\right)$ to $\left(\frac{\pi}{\Lambda} + K_1\right)$. In photon energy this gap corresponds to $\left(\hbar c \cdot \left[\frac{\pi}{\Lambda} - K_1, \frac{\pi}{\Lambda} + K_1\right]\right)$. Photons in this energy band cannot propagate in the Photonic Crystal and are therefore totally reflected.

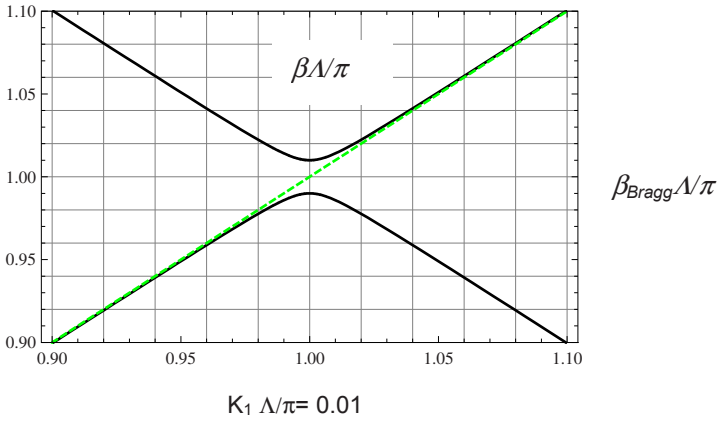


Figure 14.3. Normalized propagation constant of modes at normal incidence in a Bragg reflector (solid lines). Far from resonance, the propagation constant approaches that of the unperturbed modes (dashed line), but close to resonance there is a forbidden range of values, a band gap.

Equation 14.9 is strictly-speaking only valid in a range of wave vectors around $m\frac{2\pi}{\Lambda}$, but, as pointed out above, at wavelengths far from each resonance, we have $\beta \approx \beta_{bragg}$. In other words, away from the resonances, all the equations give the same propagation constant. That means that we can piece together the com-

plete solution by simply using the appropriate harmonic number m in Eq. 14.9 in the vicinity of the resonances and set $\beta \approx \beta_{bragg}$ in between.

14.2.2 Bloch States

The solutions to the wave equation in Bragg gratings can be expressed in terms of forward and backward propagating waves with amplitudes A and B as in Eq. 14.2. The solutions can of course equally well be presented as any other linear and independent combination of the forward and backward propagating waves. It is particularly useful to write the solutions in terms of combinations that are eigen modes, i.e. modes with well-defined wave vectors just as plane waves. We used the eigen-mode picture in Chapter 6.4.2 and 6.4.3 to describe propagation in side-coupled waveguides.

We will adopt the same approach here and represent the waves in the 1-D Photonic Crystal in terms of eigen modes. The famous and very useful Bloch theorem [4,14,15] states that the eigen-modes, or Bloch states ($\psi(z)$), can be expressed as the product of a plane wave and a periodic function

$$\psi(z) = e^{j\beta z} \cdot \Phi(z) \tag{14.11}$$

These eigen states are called Bloch waves or Bloch states, and the periodic function, $\Phi(z)$, is called the Bloch function. It is periodic with the periodicity of the grating, i.e.

$$\Phi(z + n \cdot \Lambda) = \Phi(z) \tag{14.12}$$

where n is any integer.

Adding an integer multiple of $2\pi/\Lambda$ to the propagation constant of the Bloch states only changes the solution by an insignificant phase factor. It is therefore customary to add and subtract multiples of $2\pi/\Lambda$ so that the full set of solutions can be graphed within the first Brillouin zone, which is the range of propagation-constant values from $-2\pi/\Lambda$ to $2\pi/\Lambda$. We have used this convention to plot the allowed values of the propagation constant of the Bragg-reflector modes in Fig. 14.4, where we have used a higher coupling constant ($K_1 \Lambda/\pi = 0.2$) to make the higher-order bandgaps easier to observe.

The graph shows bandgaps at odd integer values of Λ/π. Note that the bandgaps are progressively smaller for increasing values of β as prescribed by Eq. 14.7. The lack of a band gap at $\beta\Lambda/\pi=2$ is due to the square-wave shape of the index variations as given by Eq. 14.1. A more general index corrugation with a non-zero second-harmonic spatial frequency term will give bandgaps at even integer values of $\beta\Lambda/\pi$.

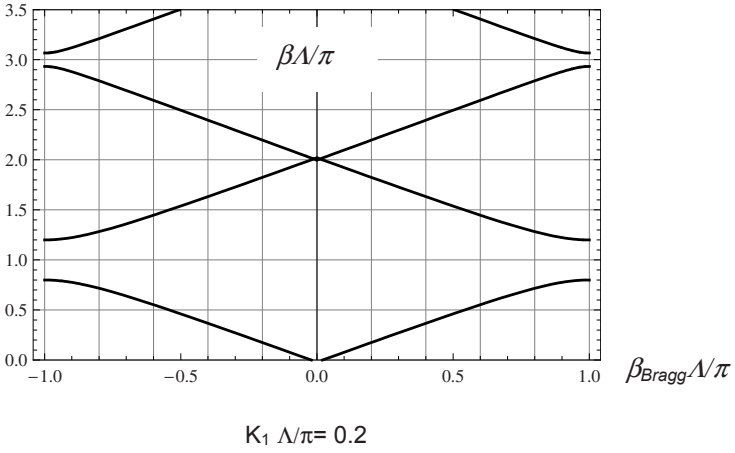


Figure 14.4. Normalized propagation constant of modes at normal incidence in a Bragg reflector over an extended range that includes two band gaps. Integer values of $2\pi/\Lambda$ have been added to β_{Bragg} to move the solutions inside the first Brillouin zone. Note the lack of a band gap at $\beta\Lambda/\pi=2$.

Scaling of Photonic Crystals

Figure 14.4 reveals a very useful fact about Photonic Crystals; everything scales with the lattice constant. In other words, PCs do not have any absolute length scales, which means that a PC device that operates at a given wavelength can in principle be scaled to any other wavelength by simply scaling the lattice constant of the PC by the same ratio as the wavelength. This is very convenient, because normalized calculations of band structure can be used over a wide range of frequencies, and successful device designs can be "reused" in other parts of the spectrum. Of course, we have to be careful when applying this principle. We must take into consideration the facts that material constants change as a function of wavelength, and that diffraction characteristics that are determined by device apertures will not scale the same way.

14.2.3 Band Structure of 2-D and 3-D Photonic Crystals

The conclusion of our treatment of the 1-D PC in the preceding section is that the periodic nature of Photonic Crystals modifies the propagation of electromagnetic waves, and may create a forbidden energy bands for photons. The situation we have analyzed is the simplest possible; propagation in a fixed direction in a transversally homogeneous medium so that polarization effects can be ignored. The general concept, however, also holds true for 2-D and 3-D Photonic Crystals, although the structure of the energy bands and energy gaps are considerably more complex than those of Fig. 14.4. Just as for the 1-D PC we have analyzed, there

will in general be allowed and forbidden photon energies for propagation in all directions within 2-D and 3-D PCs. If a particular range of energies is forbidden for all propagation directions (in the case of 2-D, all means all directions in the plane of the crystal) and both polarizations, then we say that the PC has a complete band gap.

A typical 2-D PC is shown Fig. 14.5. It is a plate made of a high-dielectric constant material with a 2-D periodic array of cylindrical holes on a square lattice. The practical implementations of such 2-D PC are described in Chapter 15. This simple structure is fully described by four parameters; the dielectric constant, the lattice constant, the plate thickness, and the hole radius. To calculate the modes of the 2-D PC we must also specify the dielectric constant of the surrounding medium. In many cases the surrounding medium is air (or vacuum) with unity dielectric constant. As described in the previous section, all length scales can be normalized to the lattice constant, so the calculations have a total of four degrees of freedom.

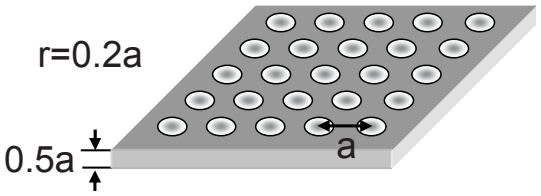


Figure 14.5 *A two-dimensional Photonic Crystal of lattice constant a defined in a high-index plate with a dielectric constant of 12 and a thickness of 0.5a. The plate has a square lattice of holes of radius 0.2a.*

The energy bands of a 2-D, square-lattice PC with a dielectric constant of $\varepsilon=12$, hole radii of $r=0.2a$, and thickness of $t=0.5a$, in air is shown in Fig. 14.6 [16]. The even and odd modes of the PC are shown separately in Fig. 14.6 a and b for clarity. The calculations of these bands rely on a period boundary condition for the unit cell of the crystal, which means that the results are for a crystal of infinite extent, and that all unit cells of the crystal are the same, i.e. the crystal is perfect. The vertical axes in these plots give frequency in units of the ratio of the speed of light to the lattice constant. The horizontal axes give the direction of the wave vector, with Γ indicating the direction normal to the plane of the crystal, X the in-plane direction along the holes, and M the in-plane diagonal direction as shown in the inset in Fig. 14.6b[b].

[b] These directions are really referring to directions in the reciprocal lattice, which for a square unit cell (or cubic unit cell for the 3-D case) has the same shape as the real lattice. For a lucid introduction to the concept of the reciprocal lattice see reference [13].

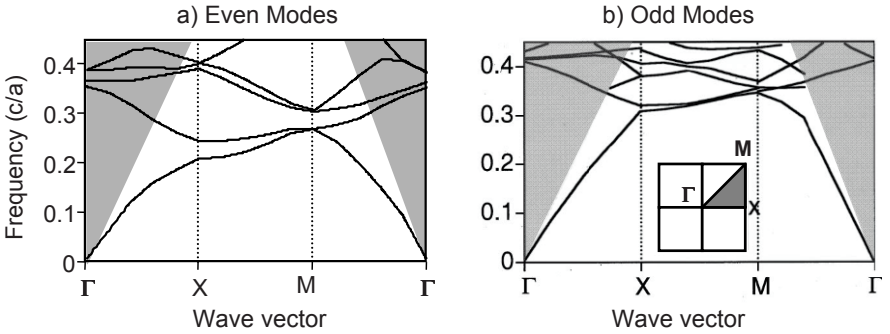


Figure 14.6 *Band structure for the PC of Fig. 14.5. The diagrams show the bands of modes of even (a) and odd (b) symmetry with respect to the center plane of the plate. The shaded regions extending from the Γ point (normal incidence) are above the light line and support a continuum of radiation modes. The solid lines in the unshaded regions represent guided modes of the crystal, while the solid lines in the shaded regions represent guided resonances. (Courtesy Professor Shanhui Fan, Stanford University)*

The band diagrams of Fig. 14.6a and b are divided into shaded and unshaded regions. The dividing line between the regions is called the light line and is given by

$$f = \frac{n \cdot c \cdot k}{2\pi} \tag{14.13}$$

where c is the speed of light in vacuum, k is the wave vector, and n is the effective refractive index of the PC. In the shaded regions above this light line, the crystal supports a continuum of modes, and these modes are phase matched to plane waves outside the crystal. In other words, for wave vectors in the shaded regions, the PC permits propagation of plane waves that couple to plane waves outside the crystal.

It is tempting to conclude that for photon energies and wave vectors that correspond to the shaded regions of the band diagrams, the PC behaves just like a uniform dielectric plate of the same effective refractive index. This is, however, not correct. The presence of the PC modes above the light line, shown as solid lines in the shaded regions of Fig. 14.6a and b, substantially changes the transmission and reflection of the 2-D PC, as explained and explored in the next section. These PC modes that exist above the light line and that are unique to 2-D PCs, are called guided resonances, due to their similarity to guided modes of uniform dielectric plates.

The band diagrams show that the crystal of Fig. 14.5 does not have a complete band gap, i.e. there is no band of energies for which there are no modes in any di-

rection of the crystal. Due to the modes above the light line, this is true for all 2-D PCs, irrespective of their specific crystalline structure. However, 2-D PC, including the one shown in Figs. 14. 4 and 14.5, have partial bandgaps for waves propagating in, or close to, the plane. These incomplete band gaps can be used to create waveguides that rely on PC confinement of the light in the plane and Total-Internal-Reflection (TIR) confinement perpendicularly to the plane. The important consideration in the design of such wave guides is that the surrounding material must have a bad gap that extends to all wave vectors that are present in the guided mode.

The lack of a complete band gap in 2-D PCs limits their use in integrated optics, so since the early days of Photonic Crystal research, the search was on for a 3-D crystal that supports a complete band gap. The first such structure to be discovered was a PC constructed from dielectric spheres surrounded by air (vacuum) in a diamond lattice [17]. The calculated band diagram of such a crystal with dielectric spheres of index 3.6 filling 34% of the crystal volume, is shown in Fig. 14.7. This crystal has a complete band gap at the frequency *c/a=0.5*.

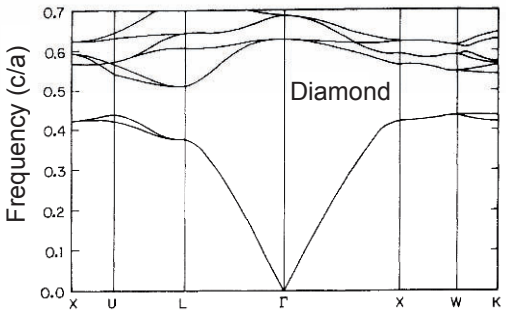


Figure 14.7 *Calculated photonic band structure of a 3-dimensional Photonic Crystal made of dielectric spheres in a diamond lattice. The spheres have a refractive index of 3.6 and fills 34% of the volume of the crystal. This crystal supports a large band gap around a frequency of c/a=0.5, where a is the cubic lattice constant of the diamond structure. Reprinted with permission from [18].*

The PC of Fig. 14.7 has had little direct practical impact, because it is difficult to fabricate for operation at visible and near IR wavelengths of the optical spectrum. The fill factor of 34% means that the spheres are overlapping, and not just touching as in an artificial opal (see Fig. 141c). Beside, spheres tend to aggregate into a Face-Centered-Cubic lattice, and not a diamond structure. However, the realization that the dielectric-sphere diamond has a complete bandgap, has led to the discovery of several related crystals, also with complete band gaps, but with structures that are simpler to fabricate. The most promising for practical applications include Yablonovite [19], the wood pile [20,21], and diamond-like structures made of patterned, offset dielectric layers [11].

Numerical Calculations

The complexity of 2-D and 3-D PCs makes calculations analytically intractable. The band structure, transmission, and other optical properties of 2-D and 3-D Photonic Crystals are therefore found numerically. A number of software systems have been developed by academic and commercial institutions for different types of PC simulations. A very useful package is the Photonic Crystal t-Software developed at MIT. It is based on a Finite Difference Time Domain [22,23] approach and enables calculations of transmitted and reflected signals in the temporal and spectral domains, as well as band diagrams. Another system, the MIT Photonic-Bands (MPB) software [24], is designed for calculations of field distributions of PC modes. It computes eigen modes of Maxwell's equations in 3-D periodic dielectric structures.

14.3 Guided Resonances

It is intuitively obvious how materials with complete band gaps, as the 3-D PC of Fig. 14.7, can be used to advantage in the construction optical waveguides and optical resonators. By simply surrounding a core (in the case of a wave guide) or small volume (in the case of a resonator) by a photonic band-gap material, we concentrate the electromagnetic energy to the regions where we want it. One of the main advantages of PCs is therefore precisely that we can that confine the radiation to small volumes and miniaturize optics beyond the limits of other technologies.

Not quite so obvious is how to take advantage of the states above the light line, the guided resonances, but the fact is that they can be used to create useful effects. It has been shown theoretically and experimentally that broad band mirrors with high reflectivity can be made from free-standing 2-D PCs [25,26,27,28,29,30] and from PCs placed on a thin dielectric film on a silicon substrate[31,32,33,34,35]. This effect can be exploited to design a number of very compact optical devices, including mirrors, filters, lasers, and optical sensors.

To understand the effect of guided resonances on free-space optical beam propagation, consider what happens when a plane wave at near-normal incidence (Γ point in Fig. 14.6) interacts with the PC at a photon energy (frequency) that corresponds to a guided resonance. When the optical field is incident on the crystal plate, some of the light is reflected from the plate surfaces and some is propagating through the plate as a plane wave. We will call this the direct path way through the PC.

So far the PC behaves just like a homogeneous dielectric plate, but in addition, the incident optical field can also excite the guided resonances. An excited guided resonance will then couple to a plane wave on the far side of the PC (this will al-

ways be the case when the PC is surrounded by the same material on both sides) and thereby provide a second path way through the PC. If there are multiple guided resonances at the frequency of the incident plane wave, then there will be a third (and possibly a fourth and a fifth so on) pathway through the plate. We will call these the indirect path ways through the PC.

The direct pathway, created by plane waves inside the PC, and the indirect pathway(s), created by guided resonances, will interfere and thereby fundamentally change the transmission and reflection of the PC. This is illustrated in Fig. 14.8 that shows a 2-D Photonic Crystal designed to operate as the most basic of optical MEMS devices; a high-reflectivity mirror. In this 2-D PC, the direct and indirect pathways are set up to interfere destructively on transmission and constructively on reflection. Consequently, the crystal becomes highly reflective. This type of mirror opens up for more compact devices with better temperature characteristics and more robust surfaces than can be achieved by devices with the metal mirrors used in most Optical MEMS applications.

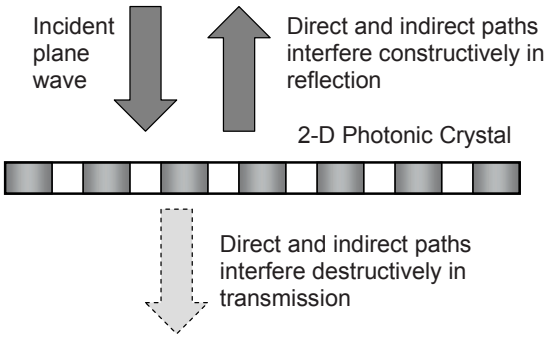


Figure 14.8 *High-reflectivity 2-D PC slab. The incident optical plane wave excites two different types of modes in the crystal; plane waves and guided resonances. These two types of modes set up two (or more if there are more than one guided resonance) pathways through the plate. In a crystal that is designed for high reflectivity, these two pathways interfere destructively in transmission over the wavelength band of interest. These modes then interfere constructively in reflection and establish high reflection from the single-layer crystal.*

14.3.1 Reflection and Transmission through 2-D Photonic Crystals

Our phenomenological explanation of interference effects in 2-D PCs can be captured in a coupled-mode theory that allows us to calculate transmission and reflection spectra. Following references [36,37], we consider a loss less optical resonator with *m* ports as illustrated in Fig. 14.9. Each port has an incoming and an out

going wave, and each wave is coupled to the resonator. The model also allow for direct coupling between each port.

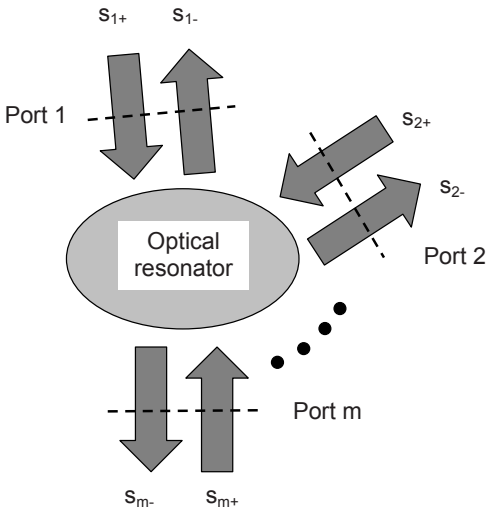


Figure 14.9 *Optical resonator coupled to m ports. Each port has an incoming and an outgoing wave, all of which are coupled to the resonator. The incoming waves are also directly coupled to all outgoing waves and vice versa. The phases of the incoming and outgoing waves are determined with respect to reference planes on each port.*

We will described the incoming and outgoing waves on each port as vectors in the following forms

$$\left|s_{+}\right\rangle = \begin{pmatrix} s_{1+} \\ s_{2+} \\ \cdots \\ s_{m+} \end{pmatrix} \tag{14.14}$$

$$\left|s_{-}\right\rangle = \begin{pmatrix} s_{1-} \\ s_{2-} \\ \cdots \\ s_{m-} \end{pmatrix} \tag{14.15}$$

The outgoing waves are coupled to the incoming waves and to the amplitude, a, of the resonator

$$|s_-\rangle = S|s_+\rangle = C|s_+\rangle + a|d\rangle = C|s_+\rangle + a\begin{pmatrix} d_1 \\ d_2 \\ \cdots \\ d_m \end{pmatrix} \tag{14.16}$$

where S is the scattering matrix for the overall system, C is the direct coupling matrix and $|d\rangle$ is the out-coupling vector.

In this formalism, the resonator amplitude is normalized such that $|a|^2$ represents the energy stored in the resonator. The rate of change of the amplitude of the resonator mode at the resonance frequency, ω_0, can be written as

$$\frac{da}{dt} = \left(j\omega_0 - \frac{1}{\tau} \right) a + \left(\langle \kappa | * \right) s_+ \rangle \tag{14.17}$$

where $\langle \kappa |* = (\kappa_1, \kappa_2, \cdots_1, \kappa_m)$ is the in-coupling vector. The combined out coupling is here represented by the life time, τ, of the resonator mode.

The coupling coefficients κ, d, and C, and the life time τ are not independent. Using energy conservation (remember that we assumed a loss less resonator!) and time reversal, it can be shown that [36]

$$\langle d | d \rangle = \frac{2}{\tau} \tag{14.18}$$

$$|\kappa\rangle = |d\rangle \tag{14.19}$$

$$C|d\rangle^* = -|d\rangle \tag{14.20}$$

14.3.2 Reflection and Transmission for a Mirror-Symmetric 2-port with One Guided Resonance

The equations can be further simplified when we restrict ourselves to a two port with mirror symmetry, e.g. a 2-D PC as shown in Fig. 14.8. If we define the ports such that the reference planes are placed symmetrically about the mirror plane, the overall scattering matrix takes the form

$$S = \begin{bmatrix} s_{11} & s_{21} \\ s_{21} & s_{11} \end{bmatrix} = \begin{bmatrix} r_d & t_d \\ t_d & r_d \end{bmatrix} + \frac{\frac{1}{\tau}}{j(\omega - \omega_0) + \frac{1}{\tau}} \cdot \begin{bmatrix} -(r_d \pm t_d) & \mp(r_d \pm t_d) \\ \mp(r_d \pm t_d) & -(r_d \pm t_d) \end{bmatrix} \tag{14.21}$$

where r_d and t_d are the direct reflection and transmissions coefficients. The sign convention here is to use the upper signs for modes that are even and the lower signs for modes that are odd with respect to the mirror plane. The total reflection and transmission become

$$r = s_{11} = s_{22} = r_d - \frac{\frac{1}{\tau}(r_d \pm t_d)}{j(\omega - \omega_0) + \frac{1}{\tau}} \tag{14.22}$$

$$t = s_{12} = s_{21} = t_d - \frac{\mp\frac{1}{\tau}(r_d \pm t_d)}{j(\omega - \omega_0) + \frac{1}{\tau}} \tag{14.23}$$

The direct reflection and transmission coefficients for normal incidence on a PC slab are given by the formulae for the Fabry-Perot etalon (Eqs. 12.12 and 12.13). For convenience, they are repeated here in slightly modified form

$$r_d = \frac{r - r \cdot e^{-j2kL}}{1 - r^2 \cdot e^{-j2kL}} \tag{14.24}$$

$$t_d = \frac{\left(1 - r^2\right) \cdot e^{-jkL}}{1 - r^2 \cdot e^{-j2kL}} \tag{14.25}$$

where r is the field reflection from the PC-air interface, L is the etalon thickness, and $k = \frac{2\pi}{\lambda} = n\frac{2\pi}{\lambda_0}$ is the propagation constant for plane waves at normal incidence in the PC, λ is the vacuum wavelength, and n is the effective index of the PC slab.

Equations 14.22-24, combined with Eq. 3.28 that gives the normal-incidence reflection from a dielectric interface, allow us to plot the reflection from a 2-D slab with one guided resonance. That is done in Fig. 14.10a and b for PC slabs with a square lattice of holes of lattice constant a and thickness $0.44\ a$. The effective dielectric constant of the slab is 12. The PC supports one even guided resonance

with an inverse life time of $\frac{1}{\tau} = 0.002 \cdot 2\pi \frac{c}{a}$. In Fig 14.10a the resonance frequency is $\omega_a = 0.33 \cdot 2\pi \frac{c}{a}$ and in b it is $\omega_b = 0.26 \cdot 2\pi \frac{c}{a}$.

Figure 14.10 shows that the reflection from the PC slab is very close to the background reflection, which is that of a uniform slab, except close to the resonances. At resonance the reflection undergoes a rapid change as a function of frequency. When the resonance is placed at, or close to a background reflection minimum as in 14.10a, then the reflection has a Lorentian-like line shape, while a resonance placed away from reflection minima results in a Fano-like line shape as shown in 14.10b. One of the very useful features of the PC is that its reflection goes to unity at resonance (in the Lorentian case) or close to resonance (in the Fano case).

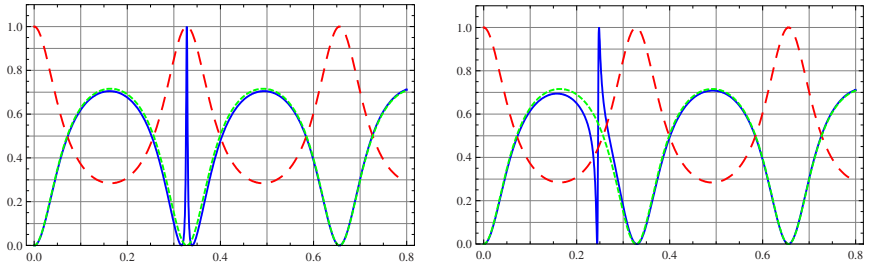


Figure 14.10 *Reflection (solid), background reflection (dotted), and background transmission (dashed) of square-lattice PC slabs with lattice constant a and thickness 0.44 a. The effective dielectric constant of the slab is 12, and the PCs supports one even-symmetry, guided resonance with* $1/\tau = 0.002 \cdot 2\pi \cdot c/a$. *In a the resonance frequency is* $0.328 \cdot 2\pi \cdot c/a$ *and in b it is* $0.246 \cdot 2\pi \cdot c/a$. *The horizontal frequency axis is in units of* $2\pi \cdot c/a$.

The line width of the resonance in Fig. 14.10 is quite narrow compared to the modulation of the back ground reflection caused by the Fabry-Perot. Figure 14.11 shows reflections from PCs with substantially wider resonances with inverse life times of $0.07 \cdot 2\pi \cdot c/a$. All other parameters are as in Fig. 14.10. We see that when the resonance is centered on a transmission maximum as in 14.11a, then the PC has a broad and very strong reflection. Such high-reflection bands clearly have many applications. When the resonance is shifted away from the transmission maximum, then we still get broad band reflection, but the band has a double-humped shape that is less useful for most applications.

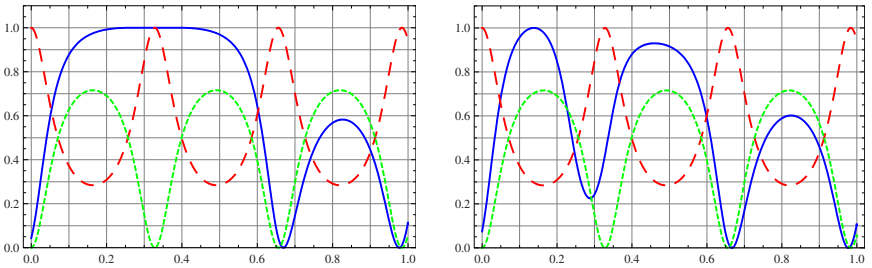


Figure 14.11 *Reflection (solid), background reflection (dotted), and background transmission (dashed) of square-lattice PC slabs with lattice constant a and thickness 0.44 a. The effective dielectric constant of the slab is 12, and the PCs supports one even-symmetry, guided resonance with* $1/\tau = 0.07 \cdot 2\pi \cdot c/a$. *In (a) the resonance frequency is* $0.328 \cdot 2\pi \cdot c/a$ *and in (b) it is* $0.246 \cdot 2\pi \cdot c/a$. *The horizontal frequency axis has units of* $2\pi \cdot c/a$.

14.3.3 Reflection and Transmission for a Mirror-Symmetric 2-port with Two Guided Resonances

The theory of coupled resonances can be extended to systems with multiple resonances. It can be shown [37] that with two orthogonal guided resonances the transmission is

$$t_{\perp} = t_d \mp \frac{\frac{1}{\tau_1}(r_d \pm t_d)}{j(\omega-\omega_1)+\frac{1}{\tau_1}} \pm \frac{\frac{1}{\tau_2}(r_d \mp t_d)}{j(\omega-\omega_2)+\frac{1}{\tau_2}} \tag{14.26}$$

where the top sign should be used if mode 1 is even, and the bottom sign should be used if mode 1 is odd. If the two guided resonances have the same symmetry, then the transmission is given by

$$t_{\parallel} = t_d \mp \frac{(r_d \pm t_d)\left[\frac{1}{\tau_1} j(\omega-\omega_2)+\frac{1}{\tau_2} j(\omega-\omega_1)\right]}{\left[j(\omega-\omega_1)+\frac{1}{\tau_1}\right]\cdot\left[j(\omega-\omega_2)+\frac{1}{\tau_2}\right]} \tag{14.27}$$

An example of the use of formula 14.26 is demonstrated in Fig. 14.12 that shows the reflection from a PC slab with two resonances. Again the PC has a square lattice of lattice constant *a*, a thickness *0.44 a*, and an effective dielectric constant of

12. The PC supports one odd guided resonance at $\omega_1 = 0.62 \cdot 2\pi \cdot c/a$ and one even guided resonance at $\omega_2 = 0.76 \cdot 2\pi \cdot c/a$. The inverse life time is $1/\tau = 0.16 \cdot 2\pi \cdot c/a$ for both resonances. In this case we see that we can create a broad, high-reflectivity band away from the transmission maxima.

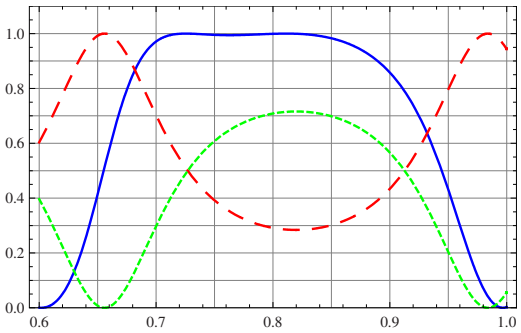


Figure 14.12 *Reflection (solid), background reflection (dotted), and background transmission (dashed) of square-lattice PC slabs with lattice constant a and thickness 0.44 a. The effective dielectric constant of the slab is 12, and the PCs supports one odd guided resonance at* $\omega_1 = 0.62 \cdot 2\pi \cdot c/a$ *and one even guided resonance at* $\omega_2 = 0.76 \cdot 2\pi \cdot c/a$. *The inverse life times for both resonances are* $1/\tau = 0.16 \cdot 2\pi \cdot c/a$, *and the horizontal frequency axis is in units of* $2\pi \cdot c/a$.

The analytical coupled-mode theory we have described in this section is not a replacement for detailed simulations. The results we get using the formulas for transmission and reflection are only as good as the data we have on the frequency and line width of the guided resonances, and those data comes from simulations or measurements.

We can give some simple rules of thumb to guide design. In general, the life time decreases with increased hole radius, because larger holes lead to stronger scattering, which again leads to stronger coupling and lower life times. Increasing the film thickness has the opposite effect: It leads to reduced coupling and longer life times. The life time of certain modes can also be understood by considering the symmetry of the PC as described in the next section.

In spite of the difficulty of relating crystal structure to modal lifetimes, the theory is nevertheless a great tool for understanding Photonic Crystal devices. The theory is intuitive so it is easy to suggest a set of parameters for a given purpose, and the usefulness of that set can quickly be assessed. It also clarifies how crystal structures translate into transmission characteristics.

Equally important is the fact that the theory helps us explain and systemize experimental observations. Recreating observed spectra using the theory allow us to determine the mode frequencies and life times and provide a short cut for investigating the effect of changes of these parameters. In short, the empirical theory we have presented is a great tool that simplifies and speeds up the design of PC devices.

14.3.4 Coupling to Guided Resonances – Symmetry

The coupled-mode theory for transmission through 2-D PCs requires the knowledge of resonance frequencies and life times, which are determined by coupling of the guided resonances to incoming and out going modes. The resonance frequencies mostly depend relatively straightforwardly on crystal structure and dimensions, while knowledge of life times typically requires detailed simulations. The symmetry of the PC does, however, give us important information about coupling and life times.

To see how, consider the unit cell of a 2-D PC slab with a square lattice of cylindrical holes as shown in Fig. 14.13. The unit cell has eight distinct symmetries; four mirror planes and four rotation symmetries (including the trivial 360 degree-rotation symmetry). It can be shown [5] that all the modes of a crystal of such symmetry must fall in one of the six symmetry classes of Fig. 14.14.

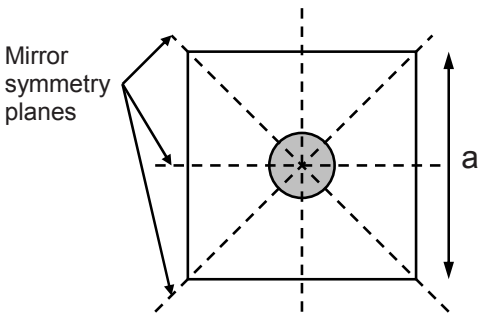


Figure 14.13 Square unit cell with four mirror-symmetry planes (the three shown plus a mirror plane parallel to the PC surface at the half thickness point of the plate) and four-fold rotation symmetry.

The convention that is used in Fig. 14.14 is that if equal signs overlap after a particular symmetry operation has been performed, then the mode is even under that operation, and if not, it is odd. This follows the usage of [5], as does the designations of the mode types. If we use the same convention for describing the symmetry of plane waves at normal incidence, we get the diagram shown in Fig. 14.15.

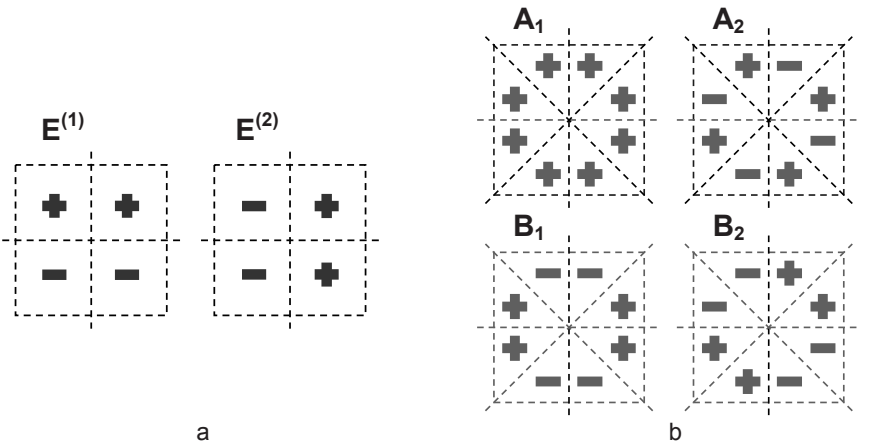


Figure 14.14 The six symmetry classes of the unit cell of Fig. 14.13. The modes depicted in a) are degenerate (in the square lattice, but not necessarily in other lattices) and they couple to plane waves at normal incidence. The six types of modes in b) are non-degenerate and do not couple plane waves at normal incidence (after [5]).

By comparing the symmetries of the mode types of Fig. 14.14 and the plane waves of Fig. 14.15, we observe the following: Vertically polarized plane waves couple to modes of type $E^{(1)}$ and horizontally polarized plane waves couple to modes of type $E^{(2)}$, but all other combinations lead to zero over lap and therefore no coupling. The consequence is that the modes of Fig. 14.15b) do not play any role in the transmission and reflection of plane waves at normal incidence.

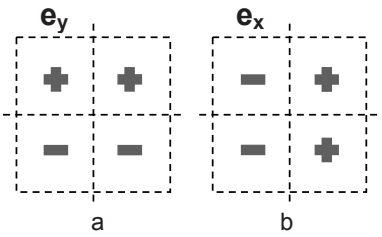


Figure 14.15 Symmetry of plane waves at normal incidence polarized vertically in the y-direction (a) and horizontally in the x-direction (b).

The usefulness of this observation is that the PC designer may bring in one or more of the modes in 14.15b) by breaking the symmetry of the desired modes.

This can be done by changing the angle of incidence[c], or by modifying the unit cell with a small perturbation designed to break a specific symmetry. This later option is very useful, because it allows control of the coupling through the pattering of the unit cell, i.e. the guided-resonance coupling, and therefore the life time, is directly controlled by lithography.

A third method for breaking the symmetry of the unit cell is to configure the crystal to deform under the influence of some measurand, e.g. pressure or acceleration, such that the deformation breaks the symmetry in a prescribed fashion. The mesurand can then be determined by its influence on the transmission and reflection of the crystal. Examples of such sensors are discussed in the next chapter on Photonic Crystal microsystems.

14.4 Comparison of Photonic and Electronic Crystals

The energy bands and gaps of Figs. 4.4, 4.5, and 4.6 are very similar to those of electronic states in natural crystals. A 1-D model similar to the one we defined in Chapter 14.2.1 can be defined and solved for electrons, yielding results that are almost identical. The extensions to two and three dimensions are completely analogous. In terms of band calculations, the only real difference is that electromagnetic fields may take one of two polarizations, so a vector model is required in the general case, while the wave function for electrons is a scalar. The consequences of the bands and the band gaps for device physics are, however, very different for electronics and photonics.

The most obvious difference between electronic and photonic crystals is that natural crystals have a well-defined number of electrons, equaling the number of atoms multiplied by their atomic number. The crystal can of course be charged, but the relative difference in the number electrons that can be sustained in a macroscopic crystal is small. Photonic Crystals have no lower or upper limit on how many photons they may contain.

In electronic crystals, each band also has a well defined number of states. A single, non-degenerate band has twice as many states as there are atoms in the crystal, i.e. each atom can contribute two electrons to each band. Photons, on the other hand, may occupy the same state, so there is no limit on how many photons can be in a given band, or even in a single state within one band.

[c] Note that a slight rotation of the PC with respect to the incident light around an axis, e.g. the *x*-axis, breaks the mirror symmetry about the same plane, because the hole not only appears elliptical, but is also tilted with respect to the incident plane wave.

A third very important difference is that electrons exchange energy with each other and with the lattice so that the electron energy is not conserved. On the contrary, an electron in a highly-excited (high-energy) state will typically decay towards the available state with the lowest energy. The electron approaches this state through a process of thermal interactions (thermalization) with the lattice and with other electrons. Electrons can also receive thermal energy from the lattice and thereby be excited to a higher-energy state. Photons, on the other hand, do not interact with each other directly, and most of their interactions with the PC lattice are elastic, so the photons can change state, but not energy. In other words, photons may move horizontally in the energy diagrams of Figs. 4.4, 4.5, and 4.6, but only under exceptional circumstances will they move vertically by changing energy[d].

The electron's ability to exchange thermal energy with its surroundings leads to the well known situation where the number of electrons in a band can be controlled by introducing "impurities", i.e. atoms with more or fewer electrons than the crystal material itself. In addition, we can shift the absolute energy of the bands by connecting it to a body with a controlled potential. The underlying mechanism for control of electron energy and band occupancy is thermalization, which means that all that is needed to change the energy of an electron is to provide an available state with lower energy.

Once the energy and the number of electrons of an energy band can be controlled, then materials with different energies and different band occupancies can be combined to create rectifying diodes and, most importantly, transistors that are the cornerstones of modern information technology. Electronic transistors are highly nonlinear devices that can perform many different functions, including amplification, binary switching, and, in a more abstract sense, rule-based information reduction.

Photons do not interact strongly with their surroundings so their energy cannot be controlled by simply applying potentials and providing lower-energy states. Photonic Crystals therefore does not enable implementations of photonic transistors or other photonic switching devices the way natural crystals do for electronics. Certainly it is possible to make optical transistors, and PCs might simplify their implementation and miniaturization and thereby make them more practical, but it would not be the PC itself that enables the transistor operation. That role would have to be filled by some optical non-linearity that is integrated into the PC.

What PCs do is to provide a means for control of photonic position and direction. The periodicity of the PC enhances the (weak) interaction between the photons and the solid, such that we may direct and localize photons in much smaller vol-

[d] These exceptions include wavelength shifts in acousto-optic modulators and second-harmonic generation and other multiphoton effects that are important in many microoptical systems, e.g. multi-photon microscopes.

umes and with much more flexibility that can be achieved with traditional optical technologies. PCs therefore provide an ideal means for communication. Signals can be sent and localized in miniaturized "circuits" (or over long distances in the case of PC fibers) that complement, rather than compete with, electronics.

Photonic Crystals also have unique advantages in sensing systems. The PC structure enhances the interaction with photons, and enables sensors in which the behavior (direction) of the photons is exquisitely dependent on the exact dimensions and materials of the PC. Sensing systems in which the measurand changes the material (e.g. bio sensors) or structure (e.g. mechanical sensors like pressure sensors, accelerometers, and gyros) of a PC can therefore be designed to have extremely high sensitivity. The fact that the sensor signals are optical is an advantage in applications where the communication of the sensor signal from the sensor to the decision-making facility is challenging. Examples of this include sensors operating in high temperature, corrosive, or other difficult environments where electronic communication will be compromised, and remote sensors that can be connected with free-space optics or optical fibers, but not electronic cables.

There are two characteristics of PCs that give them their ability to control photons; photonic band gaps and localized modes. The band gaps allow us to completely avoid electromagnetic wave propagation in certain volumes and certain directions, so that the flow of photons can be controlled. Localized modes allow us to set up interferences and control the spectral transmission and reflection of the crystal. Both these effects are important in PC devices.

We can sum up the differences between electronic and photonic crystal bands this way: Electronic system conserve the numbers of electrons and states, and photonic systems conserve photon energy. The consequences of these differences are that electronics is ideal for information-reducing circuits and photonics is ideal for information-communicating circuits. The two technologies therefore complement each other. Electronics and photonics both have unique characteristics that make them ideal for certain sensing systems. The choice of electronic or photonic sensors will depend on the specific application and environment.

14.5 Summary of PC Fundamentals

The first parts of the chapter, i.e. sections 14.1 and 14.2, contain a brief introduction to Photonic Crystals in general. The purpose is to prepare the reader for the more detailed discussion of the optical properties of 2-D Photonic Crystal Slabs in Section 14.3. Photonic Crystal slabs support guided resonances that couple to external radiation, i.e. plane waves, in ways that (1) profoundly change the optical properties of the slabs, and (2) can be precisely controlled by the structure and dimensions of the slabs. The PC slabs are modeled in a simple, coupled-mode theory that shows that these structures can be designed to have optical characteristics

that enable a variety of optical components and systems, some of which are described in the next Chapter 15. The analytical model is also a very useful tool that allows the designer to develop an intuition for how to make the correct adjustments to their crystals based on calculated or measured performance. The last section of the Chapter gives a comparison of Photonic Crystals and natural crystals, pointing out how the differences between photons and electrons, in particular their energy-conserving properties, lead to very different usage for these two technologies.

Exercises

Problem 14.1 - PC Mirrors

The broad-band mirror is one of the simplest and most robust, yet potentially a very useful, application of 2-D photonic crystals.

a. Use the model of Chapter 14.3.3 to design a broad band mirror that covers the 1.55 um wavelength range. Choose the resonance frequencies and life times to maximize bandwidth. Make reasonable assumptions for the acceptable reflectivity and variations of reflectivity within the band.
b. If possible, push the design of (a) until the mirror covers an octave. Are the chosen values of the resonance frequencies and life times reasonable?
c. Comment on the fabrication tolerances of PCs for broad-band applications compared to those of PCs for narrow band applications.

Problem 14.2 - PC Multiplexers

a. Use the model of Chapter 14.3.3 to design a filter with passband at 1.3 and 1.55 um wavelength. Choose the resonance frequencies and life times to minimize transmission at intermediate wavelengths.
b. Design a filter that passes 1.3 um and reflects 1.55 um wavelengths. Optimize the out –of-band rejection.

Problem 14.3 - 3-D PC Filters

a. How can you use a 3-D PC to create a narrow-band optical filter?
b. What are the advantages of 3-D PC filters over 2-D filters for narrow-band applications?
c. What type of applications favor 2-D PCs?

Problem 14.4 - PC Polarization Optics

a. How can you extend the models of Chapter 14.3 to polarization optics?

b. Use extended models to design a PC-based quarter-wave plate at a specific wave length.
c. Maximize the bandwidth of the quarter wave plate.
d. How does the bandwidth compare to that of a traditional quarter-wave plate? Explain the observed differences ands similarities.

Problem 14.5 - Optical "Transistors"

Explain how Photonic Crystals can enhance optical non-linearities and how that can be used for simple optical signal processing.

References

1 E. Yablonovitch, "Inhibited spontaneous emission in solid-state physics and electronics", Physical Review Letters; 18 May 1987; vol.58, no.20, p.2059-62.
2 J.D. Joannopoulos, P.R. Villeneuve, S. Fan, "Photonic crystals: putting a new twist on light", Nature; 13 March 1997; vol.386, no.6621, p.143-9.
3 J.D. Joannopoulos, R.D. Meade, J.N. Winn, "Photonic crystals: Molding the Flow of Light", Princeton University Press, New Jersey, 1995.
4 S.G. Johnson, J.D Joannopoulos, "PHOTONIC CRYSTALS The Road from Theory to Practice", Kluwer Academic Publishers, Boston, 2002.
5 K. Sakoda, Optical Properties of Photonic Crystals (Springer-Verlag, Berlin, 2001).
6 J.C. Knight, T.A. Birks, P.St.J. Russell, D.M. Atkin, "Pure silica single-mode fiber with hexagonal photonic crystal cladding", presented at the Conference on Optical Fiber Communication (OFC), San Jose, CA, Mar. 1996, Postdeadline Paper PD3.
7 P.St.J. Russell, "Photonic-Crystal Fibers", Journal of Lightwave Technology, Volume 24, Issue 12, Dec. 2006, pp. 4729-4749.
8 D.P. Aryal, K.L. Tsakmakidis, C. Jamois, O. Hess, "Complete and robust bandgap switching in double-inverse-opal photonic crystals", Applied Physics Letters; 7 Jan. 2008; vol.92, no.1, p. 011109-1-3.
9 R.C. Schroden, M. Al-Daous, C.F. Blanford, A. Stein, "Optical properties of inverse opal photonic crystals", Chemistry of Materials, vol.14, no.8, p.3305-15, 8 Aug. 2002.
10 T.G. Euser, A.J. Molenaar, J.G. Fleming, B. Gralak, A. Polman, W.L. Vos, "All-optical octave-broad ultrafast switching of Si woodpile photonic band gap crystals", PHYSICAL REVIEW B, vol. 77, pp. 115214-1-6, 26 March 2008.

11 S.G. Johnson, J.D. Joannopoulos “Three-dimensionally periodic dielectric layered structure with omnidirectional photonic band gap”, Applied Physics Letters, Vol. 77, No. 22, 27 November 2000, pp. 3490-3492.
12 S. Venkataraman, G. Schneider, J. Murakowski, S. Shi, D. Prather, “Fabrication of three-dimensional photonic crystals using silicon micromachining”, Applied Physics Letters, Vol. 85, p. 2126, 2004.
13 S. Basu Mallick, S. Kim, S. Hadzialic, A. Sudbø, O. Solgaard, “Double-layered Monolithic Silicon Photonic Crystals”, Conference on Lasers and Electro-Optics (CLEO) 2008, San Jose, CA, Paper CThCC7, May 4-9, 2008.
14 Charles Kittel, “Introduction to Solid State Physics”, Wiley, New York, 1996.
15 http://en.wikipedia.org/wiki/Bloch_state
16 S. Fan, J. D. Joannopoulos, “Analysis of guided resonances in photonic crystal slabs”, Physical Review B, 65, p. 235112, (2002).
17 K.M. Ho, C.T. Chan, C.M. Soukoulis, “Existence of a Photonic Gap in Periodic Dielectric Structures”, Physical Review Letters, vol. 65, No. 25, 17 December 1990, pp. 3152-3155.
18 K.M. Ho, C.T. Chan, C.M. Soukoulis, “Existence of a Photonic Gap in Periodic Dielectric Structures”, Physical Review Letters, vol. 65, No. 25, 17 December 1990, pp. 3152-3155.
19 E. Yablonovitch, T.J. Gmitter, K.M. Leung, “Photonic Band Structure: The Face-Centered-Cubic Case Emplying Nonsperical Atoms”, Physical Review Letters, vol. 67, no. 17, 21 October 1991, pp. 2295-2298.
20 H.S. Sözüer, J.P. Dowling, “Photonic band calculations for woodpile structures “, Journal of Modern Optics, 1994, Vol. 41, No. 2, pp.231-239.
21 K.M. Ho, C.T. Chan, C.M. Soukoulis, R. Biswas, M. Sigalas, “Photonic Band Gaps in Three Dimensions: New Layer-By-Layer Periodic Structures”, Solid State Communications, Vol. 89, No. 5, pp. 413-416, 1994.
22 K.S. Kunz, R.J. Luebbers, “The Finite Difference Time Domain Methods for Electromagnetics”, CRC Press, Boca Raton, 1993.
23 A. Taflove, S.C. Hagness, “Computational Electrodynamics: The finite difference time domain method”, Artech House, Boston, 2000.
24 S.G. Johnson, J.D. Joannopoulos, "Block-iterative frequency-domain methods for Maxwell's equations in a planewave basis," Optics Express 8, 173, 2001.
25 W. Suh, M. F. Yanik, O. Solgaard, and S.-H. Fan, “Displacement-Sensitive Photonic Crystal Structures Based on Guided Resonance in Photonic Crystal Slabs,” Appl. Phys. Lett., 82, 2003, pp. 1999-2001.
26 V. Lousse, W. Suh, O. Kilic, S. Kim, O. Solgaard, S. Fan, “Angular and polarization properties of a photonic crystal slab mirror”, Optics Express, 12, 2004, pp. 1575-1582.
27 O. Kilic, S. Kim, W. Suh, Y.-A. Peter, A. S. Sudbø, M.F. Yanik, S. Fan, O. Solgaard, “Photonic crystal slabs demonstrating strong broadband suppres-

sion of transmission in the presence of disorders", Optics Letters, 29, 2004, pp.2782-2784.

28 K.B. Crozier, V. Lousse, O. Kilic, S. Kim ,W. Suh, S. Fan, O. Solgaard, "Air-bridged photonic crystal slabs at visible and near-infrared wavelengths", Physical Review B (Condensed Matter and Materials Physics); 73, no.11, p.115126-1-14, 2006.

29 J.S. Ye, Y. Kanamori, F.R. Hu, K. Hane, "Self-supported Subwavelength Gratings with a Broad Band of High Reflectance Analysed by the Rigorous Coupled-wave Method", Journal of Modern Optics, 53, 2006, p.1995-2004.

30 E. Bisaillon, D. Tan, B. Faraji, A.G. Kirk, L. Chrowstowski, D.V. Plant, "High reflectivity air-bridge subwavelength grating reflector and Fabry-Perot cavity in AlGaAs/GaAs", OPTICS EXPRESS, 3 April 2006, Vol. 14, No. 7, pp. 2573-2582.

31 C.F.R. Mateus, M.C.Y. Huang, Y.Deng, A.R. Neureuther, C.J. Chang-Hasnain, "Ultrabroadband Mirror Using Low-Index Cladded Subwavelength Grating", IEEE Photon. Technol. Lett. 16, 2004, pp.518–520.

32 C.F.R. Mateus, M.C.Y. Huang, L. Chen, C.J. Chang-Hasnain, Y. Suzuki, "Broad-band mirror (1.12-1.62 um) using a subwavelength grating", IEEE Photonics Technology Letters, 16, July 2004, pp.1676-8.

33 L. Chen, M.C.Y. Huang, C.F.R. Mateus, C.J. Chang-Hasnain, Y. Suzuki, "Fabrication and Design of an Integrable Subwavelength Ultrabroadband Dielectric Mirror", Applied Physics Letters, 88, 2006, pp. 03110.

34 S. Boutami, B. Ben Bakir, H. Hattori, X. Letartre, J.-L. Leclercq, P. Rojo-Romeo, M. Garriguew, C. Seassal, P. Viktorovitch' "Broadband and Compact 2-D Photonic Crystal Reflectors with Controllable Polarization Dependence", IEEE Photonics Technology Letters, 18, 2006.

35 C. Lu, M.C.Y. Huang, C.F.R. Mateus, C.J. Chang-Hasnain, Y. Suzuki, "Fabrication and design of an integrable subwavelength ultrabroadband dielectric mirror", Applied Physics Letters; 16 Jan. 2006; vol.88, no.3, p.31102-1-3

36 S. Fan, W. Suh, J. D. Joannopoulos, "Temporal coupled-mode theory for the Fano resonance in optical resonators", Journal of the Optical Society of America A, vol. 20, no. 3, March 2003, pp. 569-572.

37 W. Suh, Z. Wang, S. Fan, "Temporal coupled-mode theory and the presence of non-orthogonal modes in lossless multimode cavities," IEEE Journal of Quantum Electronics, vol. 40, no. 10, October 2004, pp. 1511-1518.

15: Photonic Crystal Devices and Systems

15.1 Introduction to PC Devices and Systems

Photonic Crystals give us a whole new tool box for manipulating electromagnetic radiation. Their band gaps and localized states enable many optical devices, some of which are simply miniaturized versions of traditional optical components and others that are impossible or impractical to implement with conventional optical fabrication technologies. Photonic Crystals therefore will continue to have important impact in all fields of optics.

In this chapter we focus on the opportunities and challenges of PC devices and systems. We will start by describing fabrication techniques that are compatible with MEMS and IC technologies. To take full advantage of the opportunities that Photonic Crystals provide will require developments in MEMS technology. The very same properties of photonic crystals that make them useful for optical devices also make them extremely sensitive to pattern irregularities and surface defects. Practical and commercial development will therefore require improved MEMS surface treatments and much better lithography than is commonly used for commercial MEMS today.

From fabrication technology, we will move on to show how PCs enable new optical devices and systems. Much of the focus will be on two types of PC-MEMS that are expected to have significant technical and economical impact: Photonic Crystals that can be actuated to create tunable optical devices, and microsensors that utilize Photonic-Crystal interactions to provide superior performance.

This chapter is different from the rest of the book in that the treatment is conceptual and qualitative. The reason is partly that the topic of PC devices and systems is too large to cover in detail in a single chapter; it requires a whole book! The field is also so new that many of the device concepts are still very much on the forefront of research. Device design is rapidly changing and still a long way from reaching maturity, and, consequently, simple and elegant quantitative descriptions have not yet been developed.

15.2 IC Compatible Photonic Crystals

Photonic Crystals come in many different shapes and forms as shown in Fig. 14.1 of the previous chapter. Some of these are much better suited to chip-scale implementations than others. The holey fiber is useful for delivering optical signals to ICs, but its manufacture and form are not compatible with IC fabrication technologies. Multi-layered Bragg reflectors are used as lasers mirrors for Vertical Cavity Surface Emitting Laser (VCSELs) and other resonant optical devices, and are therefore essential for III-IV semiconductor photonics. They are, however, difficult to incorporate in standard silicon IC technology, because of material and processing incompatibilities. Differences in thermal expansion coefficient between the different layers of the Bragg gratings also lead to built-in stress that compromise free-standing structures like micromirrors.

Planar 1-D gratings and 2-D PC slabs, on the other hand, are well suited for integration into ICs. They can be created using standard IC materials and their dimensions are compatible with the film thicknesses and lithography capabilities of modern IC manufacture. Some of the fabrication technologies for PC slabs can also be extended to multilayered structures and even towards fully 3-D PCs.

In this section we will cover the most important materials and processes for fabricating PC slabs with emphasis on silicon IC and MEMS compatibility. The purpose is to present a set of PC structures that are available for use in Photonic Microsystems, and to show how existing techniques can be extended to create more complex, but still IC compatible, crystals with better performance in a wider range of applications.

15.2.1 Silicon Compatible 2-D Photonic Crystals

In Chapter 14 we saw that to create large band gaps and control guided resonances, it is beneficial to use lossless materials with high dielectric constants. Silicon is therefore close to the ideal PC material for fiber-optic applications with its high refractive index and low loss in the *1100* nm to *2* um wavelength range, that includes the important *S* (Short wavelength), *C* (Conventional), and *L* (Long wavelength) fiber-optic bands. Crystalline Silicon is preferable to poly-crystalline or amorphous materials, partly because it tend to give more reproducible crystals with smoother surfaces that are better suited for optical applications, and partly because it has lower built-in stress.

Silicon has too high loss to be a good PC material at wavelengths below 1100 nm. The natural choice for applications in the visible and near IR is therefore Silicon Nitride. It is a standard material of IC fabrication, and it has a reasonably high dielectric constant. The refractive index is about 2 for stoichiometric nitride (Si_3N_4), and it can be adjusted downwards by incorporating extra (beyond the stoichiomet-

ric balance) silicon in the material. The index of 2 is sufficient to create very useful PC devices for visible-wavelength applications [1]. Nitride has the advantage that it typically is in tensile stress when deposited on silicon substrates, so it tend to be held flat by internal stress when free-standing.

Beyond silicon and silicon nitride, there are a number of candidate materials that are not so well developed, but that have special features that make them interesting for PCs. Many polymeric materials can be shaped into complex 3-D structures by two-photon processes. In principle that enables the creation of 3-D PCs with well-controlled point defects and line defects that defines resonators and wave guides. The drawback of these materials is their low index contrast limit the type of optical functionality they can support.

Among inorganic materials, alumina or aluminum oxide (Al_2O_3 – natural crystals of this material are called Ruby or Sapphire depending on the coloring, which is caused by impurities) and silicon carbide (SiC) are the most promising. These materials have the high index and low absorption required for formation of good PCs at visible and near-IR wave lengths. This, combined with their thermal, mechanical, and chemical robustness, make alumina and silicon carbide excellent materials for sensor applications in harsh environments.

We saw in Chapter 14 that the period of Photonic Crystals is, roughly speaking, on the order of the wavelength of light. This means that state-of-the-art optical lithography has more than enough resolving power to create the patterns required for PC formation down to visible wave lengths. Still, electron-beam lithography and focused-ion-beam milling are often used for their flexibility, particularly in research environments. Nano-imprint lithography represents a compelling alternative to optical lithography for the production of PCs on a large scale.

Single-layer Photonic Crystals are straightforward to fabricate using standard IC technology. Any high-index film that is free-standing, or placed on a lower index film or substrate, can be patterned into a 1-D or 2-D PC slab. Figure 15.1 shows a typical example. Here a square lattice of cylindrical holes is patterned in the device layer of a Silicon-on-Insulator (SOI) wafer. The thickness of the crystal is 340 nm, the lattice constant is 998 nm, and the hole diameter is 700 nm. These dimensions, which are chosen to allow the PC to operate at wave lengths just beyond silicon's indirect band gap at 1100 nm wave length, show that the PC is indeed compatible with standard film thicknesses and lithography capabilities.

The PC slab of Fig. 15.1 can be released and integrated with MEMS actuators and other MEMS devices to create micro-optical systems. The release can be performed by simply etching the oxide layer underneath in a wet etch (e.g. hydrofluoric acid) through the holes of the PC, or, if a completely free-standing structure is preferred, the substrate can also be removed. Typically that will be accomplished in an anisotropic etch (e.g. potassium hydroxide) so that a well-defined substrate cavity is formed as shown in Fig. 15.2. The specific mirror of Fig. 15.1 is de-

signed for high reflectivity when released from the substrate and oxide film. Under free-standing conditions, it gives better than 98% reflectivity at normal incidence in the wave length range from 1220 nm to 1255 nm.

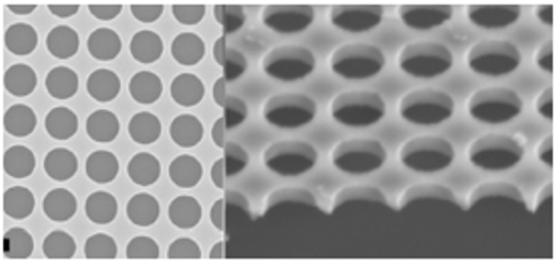


Fig. 15.1 *Scanning-electron micrographs of a high-reflectivity, single-layer, Photonic-Crystal mirror seen from the top (a) and in perspective (b). The PC is fabricated in a 340 nm thick silicon layer of a SOI wafer. The crystal has a square lattice with a 998 nm lattice constant and a 700 nm hole diameter.*

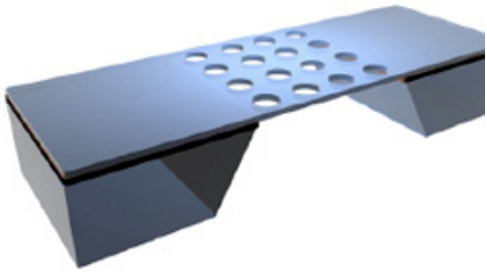


Fig. 15.2 *Free-standing PC mirror released by removal of the silicon substrate and the oxide layer of the SOI wafer through a combination of KOH and HF etching (not to scale).*

The essence of a 2-D PC is a pattered, high-index plate that is free-standing or resting on a lower index material. It can be made by deposition, or growth, of layers of alternating refractive index, but his method has limitations. As for the crystal of Fig. 15.1, a SOI wafer with a layer of crystalline silicon on silicon oxide can be used as the starting material. The drawbacks are that typically only a single layer can be used, that the silicon layer is of uniform thickness, and that the PC layer is on the starting material as opposed to being deposited at some more opportune step in the fabrication process. These problems can be solved by going to

poly-silicon, or amorphous-silicon, films that can be deposited in a range of thicknesses at any time during processing. Crystals made from such deposited films tend to suffer from crystalline defects caused by the non-uniformity of the film materials and from built-in stress caused by the thermal mismatch between the films and the underlying substrate. These effects become particularly difficult in multi-layer stacks, so the patterned, thin-film approach to building PC typically cannot be extended much beyond one or a few layers.

A compelling alternative is therefore to etch the PC directly into the silicon substrate through a sequence of etch steps [2,3]. An example of such an etch process is shown schematically in Fig. 15.3. The process starts with an oxidized, standard Si wafer. In the first step resist is spun on the wafer and patterned by optical lithography. The pattern is then transferred into the oxide masking layer and a first directional etch is performed to create holes in the silicon (step 3). After the silicon etch, the whole wafer is conformally covered by oxide that is subsequently removed in a directional etch so that a protective oxide remains on the side walls, but not on the bottom, of the silicon holes (step 5). Finally the process is completed by a directional Si etch (step 6) followed by an isotropic etch (step 7).

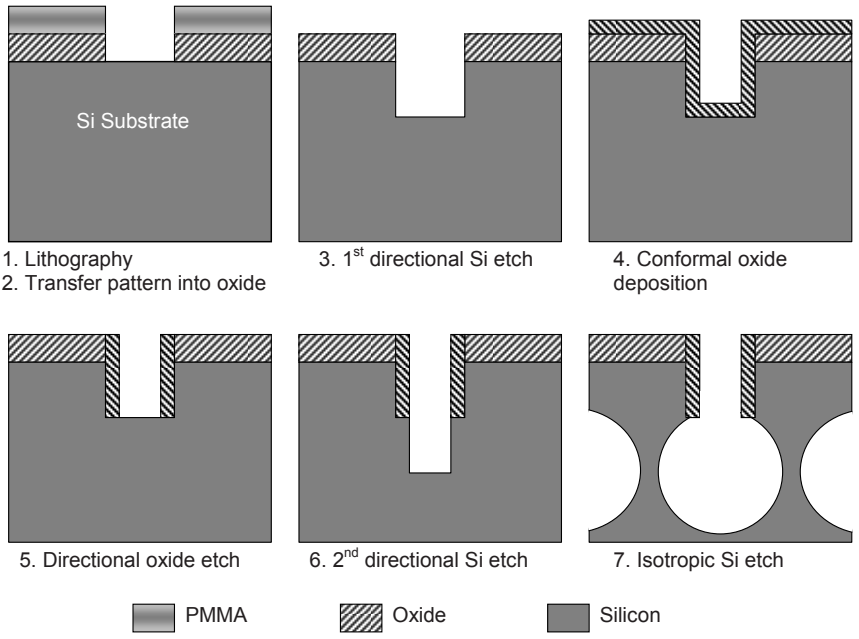


Fig. 15.3 *A single unit cell illustrating the GOPHER [3] fabrication process for etching 2-D Photonic Crystals into a monolithic silicon substrate. The purpose of the final isotropic etch is to create a lower-index layer under the Photonic Crystal. The low index layer is required to support guided resonances in the PC.*

The final isotropic etch creates a buried, low-index layer below the Photonic Crystal. Without this low-index layer the PC does not exhibit neither the guided modes below the light line, nor the guided resonances above the light line, so the low-index layer is crucial for the operation of the PC. Depending on the length of the final isotropic etch, the PC layer might be partially or fully released as shown in Fig. 15.4 a and b.

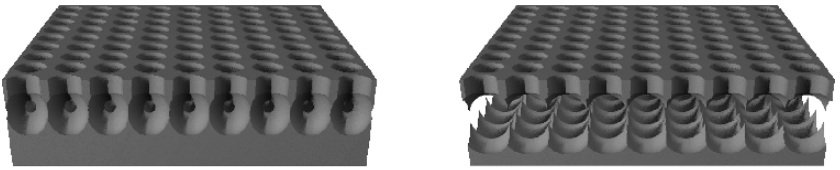


Fig. 15.4 *An isotropic etch of intermediate length results in a low-index layer with overlapping spherical holes and solid connections between the PC and the underlying substrate (a). If the isotropic etch is longer, then the PC layer is fully released (b), and may be moved by integrated MEMS actuators.*

The PCs created by the process of Fig. 15.3 can be thought of as being formed by intersecting cylindrical and spherical holes. That results in structures with very pronounced spikes and jagged interfaces as shown in Fig. 15.4b. Such high-spatial frequency interfaces will amplify the effects of fabrication errors, leading to exaggerated variations of optical properties within a single crystal and between different crystals. It is therefore useful to be able to smoothen out the jagged interfaces, using hydrogen annealing [4,5,6,7]. Figure 15.5 illustrates the remarkable changes in structure and surface quality that can be achieved with this technology.

Photonic Crystals fabricated by direct etching of silicon wafers have many advantages over PC made from deposited of grown thin films. Wafer-quality Single Crystalline Silicon (SCS) is far superior to deposited films and also better than the device layer of SOI wafers, so the construction material of etched PCs is the very best. The finished PC is essentially a monolithic structure with only small amounts of native oxide or other passivation layers. This means that the PCs have excellent chemical and mechanical properties, and that they do not experience significant material stress due to differences in thermal expansion.

The fabrication is also very flexible, and most importantly, compatible with standard MEMS and IC processing. That includes high-temperature deposition, growth, and annealing, so post processing of the PCs can be performed with the full tool set of microfabrication. Finally, the crystals are self aligned, i.e. their

structure relies on a single lithography mask. This is a key property for ensuring efficient and reliable manufacturing of any semiconductor device.

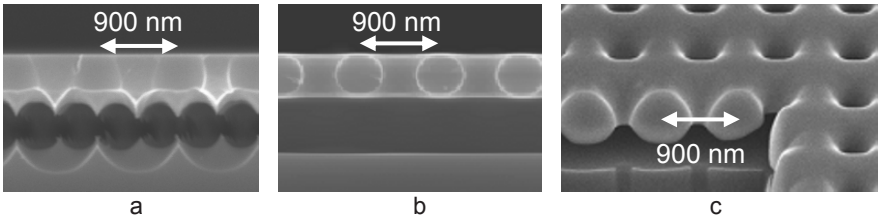


Fig. 15.5 *Scanning Electron Micrographs of Silicon PC before (a) and after (b and c) hydrogen anneal [8]. Figures a and b shows cross sections, while c shows the top surface after part of the crystal has been removed by a Focused Ion Beam (FIB). The hydrogen anneal was performed in two steps for a total of 15 minutes at 950 and 10 Torr. The period of the square-lattice crystal is 900nm and the holes are 600 nm in diameter.*

15.2.2 3-D Structuring of Photonic Crystals

At present there is no established IC-compatible fabrication technology for three-dimensional Photonic Crystals at wave lengths in the visible and near-IR. The motivation for creating 3-D PCs at these wavelengths is strong, however, so a number of methods are being investigated. The most promising developments include building log piles by layer transfer of separately patterned layers [9], layer-by-layer construction by repeated deposition, patterning, etching, back-filling of holes, and planarization [10], holographic lithography in low-index polymers [11], and guided assembly of dielectric spheres [12]. Each of these methods have their drawbacks and must be substantially refined before they are simple enough, robust enough, and well enough suited to the IC fabrication environment to be used for commercialization of 3-D PCs.

Even though IC compatible 3-D PCs remain a goal for the future, there are several methods that can be used to create multiple, aligned layers of 2-D PCs. Double-layer systems have many device applications, some of which will be described in the next section, so they represent a substantial improvement over single-layer PCs. Both layer-transfer and layer-by-layer construction can be straightforwardly applied to double-layered PC, as can direct etching as shown in Fig. 15.6, so fabrication of double-layer PC devices are well within the capability of IC technology.

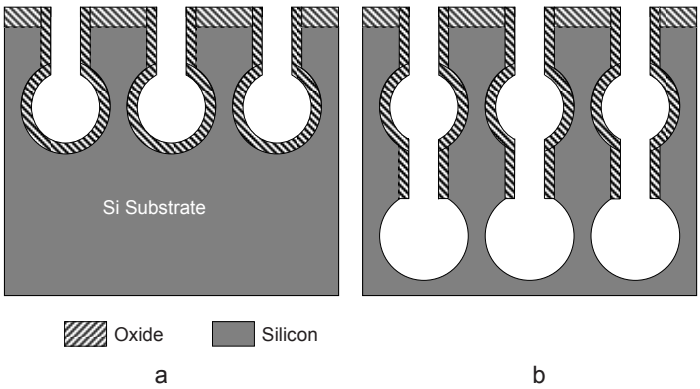


Figure 15.6 *By conformally passivating a single-layer, direct-etched PC (a), the fabrication sequence of Fig. 15.3 can be repeated to create a second, aligned PC layer (b). The process may be repeated to create additional layers. Depending on the size of the isotropically etched spheres, the layers may be connected vertically or completely released as shown in Fig. 15.4 b.*

15.3 Photonic Crystal Optical Components

The most obvious advantages of PCs are their band gaps, i.e. regions of the optical spectrum where the material rejects electromagnetic radiation, at least in certain directions. A photonic band-gap material is therefore an excellent means for producing optical waveguides. By simply surrounding an optical propagation medium with a band-gap material, we can create a structure that support guided modes.

Such PC waveguides are in principle quite different from the traditional waveguides we have described in Chapters 5 and 6. Specifically, the mode-confining mechanism is different. In PC waveguides, the surrounding band-gap material confines the propagating mode to the core of the wave guide, while in traditional guides, Total-Internal-Reflection (see Chapter 3) provides the confinement.

From a user's point of view photonic band-gap guides share many of the characteristics of regular waveguides. The most important functional difference is that PC waveguides are not constrained to having a high-index core as required by TIR waveguides. In fact, PC waveguides and fibers may have a core of low-index material of even air or vacuum. The Holey fiber of Fig. 14.1b is an example of such a PC waveguide with a low-index core.

The ability to support guided modes in vacuum is a very valuable practical advantage of PC waveguides. It means that guided modes with very low dispersion and very low non-linear optical response can be created. That is particularly useful for transport of high peak-power, femtosecond laser pulses that would be corrupted by the dispersion and non-linearity of standard fiber. Such pulses can be delivered over holey fiber with manageable pulse distortion. This ability opens up a new class of optical instruments that use the flexibility and small size of holey fibers to deliver high peak-power laser pulses in hard-to-reach places. A very promising application is in-vivo microscopy and cell surgery using femtosecond lasers [13].

In addition to being ideal confinement structures for waveguides and resonators, PC also enable a variety of other optical functions through their ability to support a wide variety of optical modes with well-defined frequency, spatial extent, and loss. The functionality of PC devices springs from the interference and coupling between these PC modes and incident and transmitted light. That is not unique to PC devices; most optical devices in loss dielectrics are based on interference, but PCs have the very useful characteristics that they create such interferences in smaller volumes than can be achieved with other technologies. Almost all basic components and building block of photonic microsystems can therefore be improved by the judicial addition of PCs.

15.3.1 Mirrors and Filters

Chapter 14.3 describes how the presence of guided resonances in Photonic Crystals changes their reflection and transmission characteristics. In particular, we demonstrate how destructive interference between a direct (plane-wave like) pathway and an indirect (via a guided resonance) pathway through the crystal leads to low transmission and high reflectivity. It is somewhat counter intuitive, but nevertheless true, that the high reflectivity can be extended to a broad range of wave lengths [14,15] and incident angles [16].

Mirrors based on this concept have mechanical and chemical properties that make them ideally suited for photonic Microsystems. First and foremost they are compact. The thickness of the 2-D PC is less than one quarter of the vacuum wavelength, and, due to the wide angle of acceptance, it is possible to focus the optical beam to a spot of a few wavelengths in radius. This makes it possible to create compact, yet mechanically very flexible, 2-D PC membranes that can be used for optical switching and sensor applications. In spite of their mechanical flexibility, such PC membranes are very robust due to their simple structure and the excellent mechanical properties of their constituent materials, which are semiconductors and high-index dielectrics.

These materials also have excellent chemical and thermal properties. They have much higher melting temperatures than metals, and in contrast to metals, they do not absorb, but scatter, the light that is not reflected. This gives the PC mirror

very good power-handling capacity. In addition, the semiconductors and dielectrics of PC mirrors have well-characterized etching characteristics when exposed to etching technologies used in IC and MEMS fabrication.

These characteristics make PC mirrors superior to metal mirrors and dielectric stacks in Photonic Microsystems. Metal mirrors have lower reflectivity, tolerate lower optical powers, absorb the power that is not reflected, and have inferior mechanical, thermal, and chemical robustness. The simplicity of metal mirrors will make up for these drawbacks in some applications, but in others they are unacceptable.

Dielectric-stack, or Bragg, mirrors are much thicker than PC mirrors. Even high index contrast Bragg mirrors typically have on the order of five pairs of quarter-wave layers, making these mirrors an order of magnitude thicker than PC mirrors at the same wavelength. This thickness increase leads to additional size increase of supporting structures. The total size increase will vary from device to device. If for example, we want to make a diaphragm of a given compliance, then a ten-fold thickness increase will require a comparable increase of the diameter of the diaphragm.

High-reflectivity Bragg mirrors also suffer from optical field penetration of dielectric stacks, while PC mirrors achieve high reflectivity in a single sub-wavelength layer. This means that resonators based on PC mirrors can have effective optical lengths that are shorter than is possible for resonators based on Bragg mirrors. PC mirrors are therefore more flexible and enable a wider range of configurations for optimization of optical systems for optical generation, modulation, and sensing.

Finally, PC mirrors are more easily integrated with other micro-optical components. The most commonly used Bragg-mirror materials are not compatible with IC and MEMS technology, and the thermal-expansion mismatch between the layers of the stack leads to stress and stress gradients. Even more difficult is the sequential deposition of multiple layers. This adds a number of fabrication steps and makes the fabrication process impractical.

The treatment in Chapter 14.3 also shows that 2-D PCs are well suited to filter applications. Both reflective and transmissive filters can be designed to have broadband and narrow-band spectral features as desired. Symmetry breaking as described in Section 14.3.4, is a practical and useful tool for accurate control of guided-resonance life times. Such control enables the creation of sharp interferences in transmission and reflection spectra for narrow-band filters.

15.3.2 Photonic Crystal Fabry-Perot Resonators

Two PC mirrors can be combined to a very compact optical resonator as shown in Fig. 15.7. This device has two distinct operating regions. If the two PCs are sepa-

rated by more than half a wavelength, then the evanescent, or near-field, coupling between the two is insignificant [17]. In the absence of direct coupling between the PCs, the resonator acts conceptually as a traditional F-P, i.e. it exhibits transmission maxima when the PC separation equals an integer number of half wavelengths. The fact that the PC mirrors can be designed to provide high reflectivity from a single sub-wavelength layer makes it possible to create high-finesse resonators as short as half a wavelength. This ability is unique to PC resonators[a] and makes it possible to create highly sensitive, yet stable, displacement sensors as described in detail in Chapter 12.2.4.

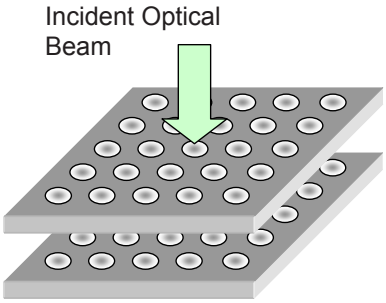


Figure 15.7 *Schematic drawing of a photonic-crystal resonator. If the PC mirrors are separated by half a wavelength or more, then the coupling between the modes of the two crystals is insignificant, and the resonator behaves like a traditional Fabry-Perot. If the PC mirrors are close, then the near-field coupling between them makes the transmission extremely sensitive to lateral and vertical displacement of the plates.*

15.3.3 PC Tunneling Sensors

When the separation between the two PC mirrors in Fig. 15.7 is on the order of a quarter wavelength or less, then the evanescent coupling between the guided resonances of the two become significant. In this regime, the evanescent fields may facilitate transmission of the optical power between the PCs. On other words, the photons are tunneling between the plates. The coupling of the guided-resonances of the two plates depends critically on the symmetry and configuration of the combined crystal, so the tunneling through the structure is extremely sensitive to the relative position of the two plates [17]. This effect has been successfully demonstrated and applied to acoustic-pressure sensing [18,19].

[a] Metal mirrors enable short cavities, but not high finesse, while Bragg mirrors provide high reflectivity, but do not allow short cavities due to the layered structure of the mirrors.

15.3.4 PC Polarization Optics

Symmetry can be used to control reflection and transmission of 2-D PC. By changing the periodicity along orthogonal axes or by breaking the symmetry of the unit cell itself (see Section 14.3.4), we can create PCs with different reflection and transmission for different polarization states [20]. In addition to the usual advantages of being compact and compatible with IC and MEMS fabrication technology, this type of polarization optics has the advantage that the form birefringence is determined by lithography. This allows tremendous flexibility in design and fabrication, such that almost any conceivable piece of polarization optics can be created.

15.3.5 PC Index Sensors

The frequency and lifetime of PC modes themselves, as well as the coupling between modes, depend, among other things, on the evanescent fields of the modes outside the PC. The effect that these modes have on optical characteristics like reflection and transmission is therefore sensitive to the refractive index in the vicinity of the crystal. This dependence can be exploited to create index sensors that are sensitive to index changes in a thin layer on or around the crystal.

An important application of this index-sensing principle is the detection of bio-molecular associations [21], as shown schematically in Fig. 15.8. The sensor consists of a 2-D PC with a bio-molecular thin film on its surface. The bio-film is designed to bind a specific molecule of set of molecules. The molecular association can be antibody-antigen binding, DNA hybridization, or some other protein-protein binding.

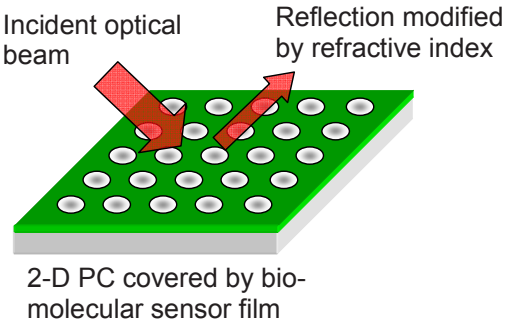


Figure 15.8 *Photonic Crystal index sensor designed for detection of bio-molecular associations. The PC is functionalized with a sensor film that recognizes a specific bio-molecule. When the target agent binds to the sensor film, the index in the vicinity of the PC changes and so does the reflection.*

Once the sensor is exposed to the molecule it is designed for, the association takes place, and the refractive index of the biosensor film changes proportionally to the number of bound molecules. This changes the resonance frequency of one or more PC modes, which in turn modifies the optical characteristics of the PC. The modified property can be reflection, as shown in Fig. 15.8, or some other quantity like transmission, diffraction angle, polarization, or phase.

More sensitive index measurements can be made by confining the light to create a high-finesse optical resonator as shown in Fig. 15.9. Here the bio-sensor film is applied in the central defect of the 2-D PC. The defect creates an optical resonator with one mode that has a resonance frequency that is very sensitive to the refractive index of the sensor film. The mode is excited through a single-mode waveguide, and the resonance frequency of the defect mode is determined by measuring the transmission of the incident light to the output waveguide.

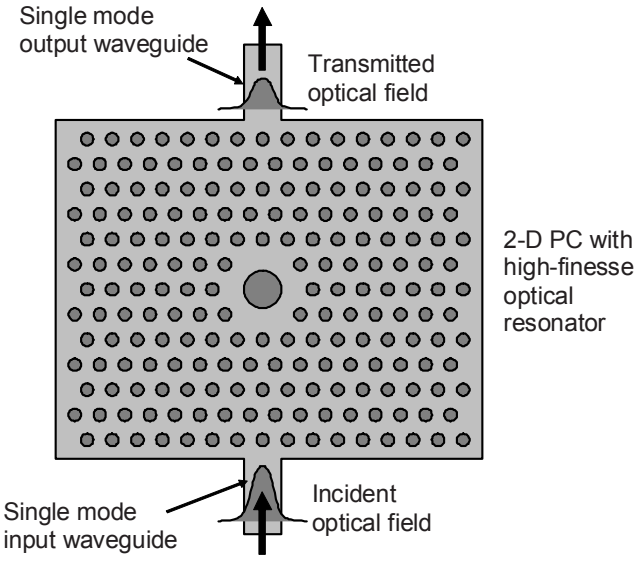


Figure 15.9 High-finesse optical resonator for detection of bio-molecular associations with single-molecule sensitivity. The central defect is functionalized with a bio-sensor film that changes the resonance frequency when a specific bio-molecular binding takes place. The state of the sensor is determined by transmission measurements (after [24]).

This resonator can be thought of as a waveguide Fabry-Perot: The section of Photonic Crystal between the input waveguide and the defect is highly reflecting at resonance, and the same it true for the section between the defect and the output waveguide. These two PC sections correspond to the first and second mirrors of a F-P, and the defect corresponds to the volume between the mirrors.

This comparison to a traditional F-P resonator helps explain the high sensitivity of the structure when it is used as an index sensor. On resonance, the fields build up in the defect mode until the fields are so high that the input coupling equals the output coupling, and the transmission approaches unity. Even minor perturbations of the defect mode will change the resonance condition and significantly reduce the transmission, making the sensor extremely sensitive to index changes. Sensors based on this principle approaches single-molecule sensitivity [22,23].

15.4 Tunable Photonic Crystals

The PC components we have listed in section Chapter 15.3 are static, but in many cases it is necessary to change the PCs during operation to implement additional functions. For example, we might want to modify the PC band gap or tune the resonance frequency, life time, or coupling of a PC mode to change the state of an optical modulator, switch, or filter. In sensor application it is often desirable to change the operational characteristics to optimize the sensitivity of the sensor to a specific measurand and a specific environment.

Tuning of PC characteristics can be achieved by changing the refractive index of the PC material or its surroundings. In principle we can use any effect that modulates the refractive index, but most tunable PC devices that have been developed require relative index changes on the order of a percent or more. That excludes the Kerr effect, the electro-optic effect (which is zero in Si, but much used in devices made in Lithium niobate and III-V semiconductors), and band-gap effects like the Quantum-Confined-Stark effect and Wannier-Stark effect, and leaves us with four viable candidates; liquid-crystal tuning, thermally tuning, injection of free carriers (plasma effect), or structural changes mediated by MEMS actuators.

Liquid crystals can be incorporated into PCs as part of the fabrication process and used to change the PC index, either thermally or by electric fields [24]. This is a promising approach to large-scale production of tunable PC devices, because of the large changes in refractive index that liquid crystals can support. Significant process development is needed, however, to optimize the incorporation of liquid crystals and make them compatible with IC technology.

Thermal tuning of PCs enables index tuning as high as several percent [25] and its implementation is simple and convenient. All that is needed are integrated resistors that heat the PC through ohmic losses. The resistors can be applied as thin films or directly integrated into the material in the case of semiconductor PCs.

The challenge of thermal tuning is to localize the heat to small volumes. Ideally we would like to isolate the individual devices so that they can be heated with minimum power dissipation, minimum influence on neighboring devices, and

maximum speed. Good thermal isolation of small volumes can be achieved by utilizing MEMS technologies, like sacrificial etching, to provide air bridges and isolating cavities, but the overall size of the optical devices tend to be dominated by the isolating features rather than by the requirements of the optical functions. Thermal tuning is therefore best suited for larger devices that are meant to be uniformly heated and that can tolerate the relatively long switching times (micro seconds) required to remove heat from large volumes.

The free-carrier effect, or plasma effect, is straightforward to apply to tuning of semiconductors PCs [26,27]. The simplest method is to irradiate the parts of the PC that should have its index modified with light at frequencies above the semiconductor band gap. Direct carrier injection is also a practical possibility in PC devices that are integrated with transistors. The plasma effect tends to be fast (sub-nano second) in PCs because of the high surface to volume ratio of these structures, so relatively high bandwidth switching can be achieved through the use of this effect.

MEMS provide extra dimensions for tuning, because rather than simply change the refractive index, MEMS actuators can dynamically modify the size, shape, and position of Photonic Crystals. Flexible structures [28,29] allow the complete PC to be dynamically altered in response to applied forces. This enables tuning of all aspects of the optical response of the crystal.

It is simpler and more efficient, however, to modulate the PC by changing its position or surroundings, because that avoids the difficulty of controlling the overall structure and shape of the PC, and because PCs through evanescent coupling are extremely sensitive to their position relative to external structures. The external structures can simply be the substrates that the PCs are built on, but the position sensitivity is enhanced if the external body is another PC [30] or a nano-scaled object like the tip of an Atomic Force Microscope [31].

These effects have been used to demonstrate MEMS-actuated, tunable PC switches [32,33,34], displacement sensors [35,36,18], tunable optical filters [37,38,39,40], and tunable optical resonators [41,42]. In the following we will describe two MEMS-based PC systems; one actuator and one sensor. These are picked to illustrate the compatibility of PCs and MEMS, and the simplicity of their integration.

15.4.1 Photonic Crystal MEMS Scanners

Photonic-crystal mirrors have several advantages over traditional mirror technologies like metal films and dielectric stacks; they have low internal stress, can handle high optical intensity, and their dimensions make them compatible with modern optical lithography. This latter point is particularly advantageous in MEMS implementations, because the PC mirrors are created as an integral part of the

MEMS structure without requiring more than a few simple additional process steps.

Integration with standard MEMS is further simplified when the PC material is silicon. It makes the PC mirror chemically robust and tolerant of high-temperature processing so that the mirrors can be made at any stage of the fabrication sequence. Silicon PC mirrors are also material-compatible with IC and MEMS fabrication. This is in sharp contrast to metal mirrors and Bragg stacks that contain materials that degrade the performance of devices like transistors and photo detectors. The chemical and thermal robustness of PC mirrors also enables high-temperature, wafer-scale encapsulation that simplifies packaging. The conclusion is that PC mirrors are well-suited to the MEMS manufacturing environment and therefore significantly simpler to integrate into photonic microsystems than metal mirrors and dielectric stacks.

These advantages of fabrication, integration, and packaging make PC mirrors candidates for replacing metal mirrors in almost all traditional Optical MEMS applications. Figure 15.10 shows a simple MEMS scanner with a PC mirror. In this example the scanning surface with the PC reflector is suspended on torsional springs and rotated by electrostatic comb drives.

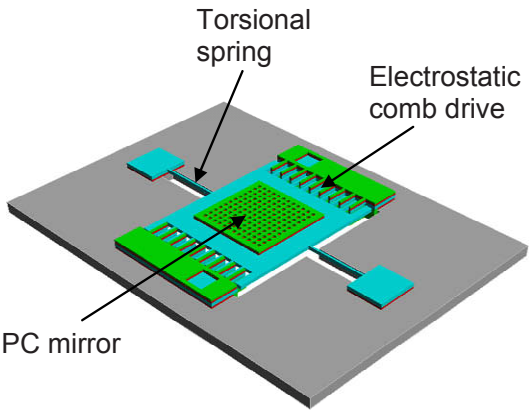


Figure 15.10 Conceptual drawing of MEMS scanner with PC mirror. The scanner is suspended on torsional springs and driven into rotation by the vertical, electrostatic comb drives on either side of the reflecting surface with the PC mirror.

A more detailed picture of the implementation of a PC scanner is shown in the cross sectional view of Figure 15.11 [43]. The MEMS components, including the scanning-mirror base, the springs, and the actuators, are made by DRIE (Deep Reactive Ion Etching) of SOI (Silicon on Insulator) wafers, while the PC reflector is made in a polysilicon layer deposited on an oxide film on the mirror base. The re-

flector can be created before or after the MEMS devices. The actuators are designed for resonant operation, with only a slight asymmetry created by the polysilicon layer. This asymmetry facilitates the starting of resonant actuation.

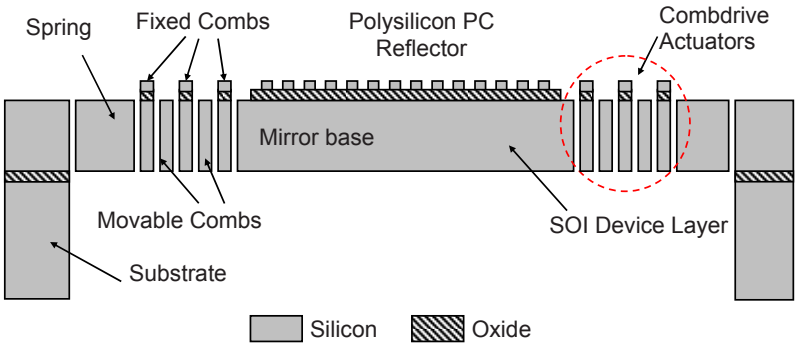


Figure 15.11 *Cross section of a PC MEMS scanner fabricated in SOI by DRIE. The PC mirror is formed in a polysilicon layer deposited on an oxide film on the mirror base, which is resonantly driven by slightly asymmetrical vertical combdrives. The substrate under the mirror is removed by DRIE.*

The PC scanner can be made monolithic by etching the PC mirror directly into the mirror base as shown in Fig. 15.12. The mirror is defined using the direct-etching process described in Section 15.2.1. If a single layer reflector is used, then it may be necessary to create a free-standing PC to ensure sufficient reflectivity, and a scattering surface on the opposite side of the mirror base may be required to avoid Fabry-Perot effects in the base.

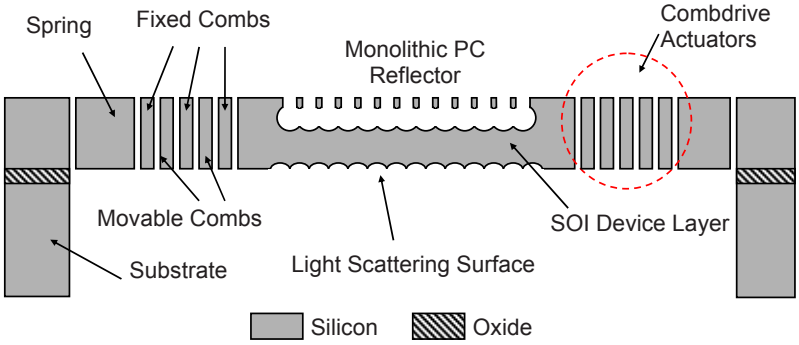


Figure 15.12 *MEMS scanner with PC mirror that is etched directly into the SOI mirror base, thus creating a monolithic scanner.*

To avoid the problems of a free-standing reflector and the extra processing required to fabricate a scattering surface, we can use a double-layer PC mirror as shown in Fig. 15.13. The double layer gives better reflectivity, even in a fully connected crystal like the one shown. This mirror is therefore more mechanically robust, and also simpler to fabricate, because the scattering surface is unnecessary.

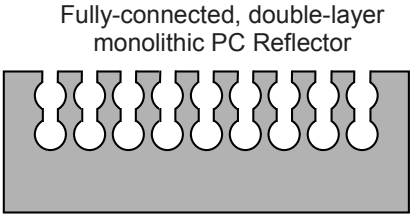


Figure 15.13 Fabrication process of a PC MEMS scanner (a) Pattern polysilicon with ebeam (b) Etch poly and oxide (c) Pattern and etch combdrives in SOI layer (d) Release substrate by DRIE (e) Release buried oxide (f) Pattern scattering surface with FIB.

15.4.2 Photonic Crystal Displacement Sensors

Displacement sensors are the basis for many MEMS sensing systems, including accelerometers, gyros, pressure sensors, and microphones. Photonic Crystals, with their high sensitivity to relative position, are good building blocks for such sensor systems. Figure 15.14 show conceptually how such a sensor can be designed.

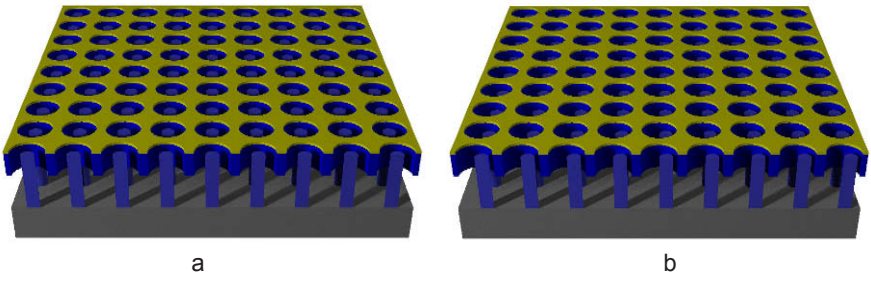


Figure 15.14 Conceptual drawing of a Photonic Crystal displacement sensor. The crystal slab is nominally perfectly centered on the pillars that serve as reference points for the PC with respect to the substrate (a). When the PC is displaced relative to the substrate (b), then the evanescent coupling of the crystal modes to the reference pillars change, and so does the optical reflection of the PC sensor.

The sensor consists of a movable PC plate that is spatially referenced to the substrate by pillars inside the PC holes. If the PC is laterally displaced with respect to the pillars, then its reflectivity changes, yielding a sensing system with nanometer sensitivity [44].

The fabrication sequence of the PC position sensor is shown in Fig. 15.15. For clarity, only a single PC hole is shown. The supporting MEMS structures are defined and fabricated simultaneously with the PC.

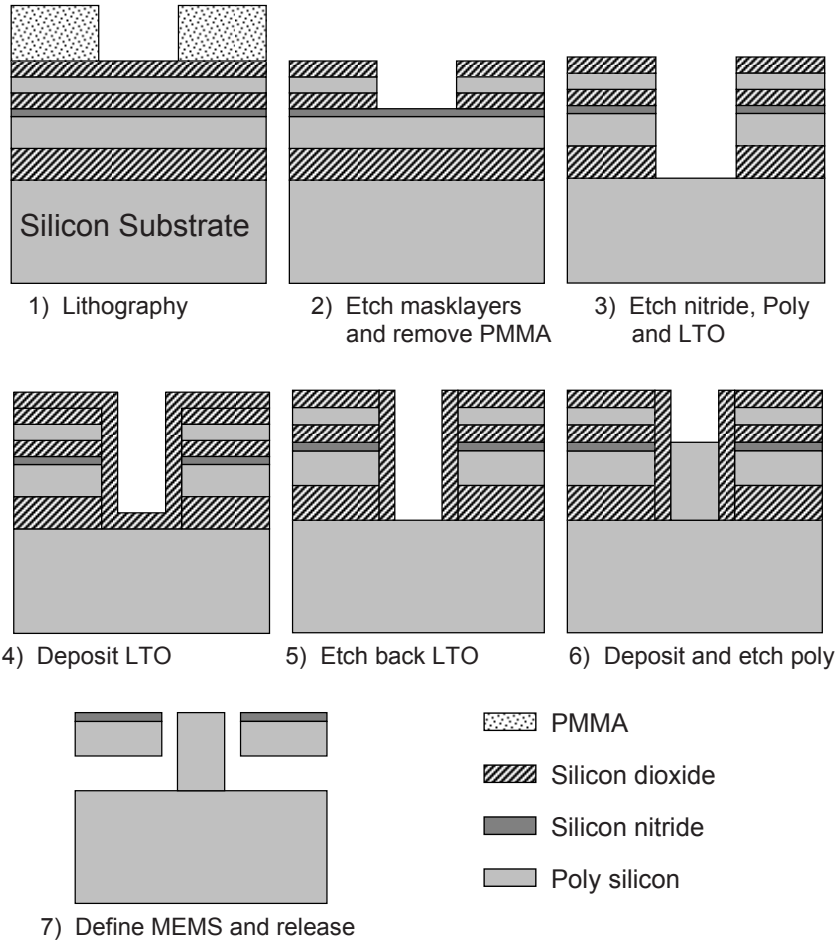


Figure 15.15 Fabrication sequence for PC displacement sensor.

The process starts with patterning of the basic PC structure in a PMMA resist layer (Step 1). The PC pattern is transferred into the masking layers (oxide/poly/oxide) and the resist is removed (Step 2). The formation of the PC and

the supporting MEMS structures is completed by etching the upper nitride layer, the poly silicon, and the underlying oxide (Step 3). A Low-Temperature Oxide (LTO) layer is then deposited (Step 4) and etched back to expose the bottom of the PC holes, while the sides of the holes are covered (Step 5). The reference pillars are then formed by deposition and etching of a poly silicon layer (Step 6). At this point the MEMS structures that supports the PC plate are defined, and finally the complete structure is released in a sacrificial oxide etch (Step 7).

15.5 Photonic Crystal Fiber Sensors

The compactness and simplicity of Photonic Crystal devices enable direct integration on single-mode optical fibers as shown conceptually in Fig. 15.16. The PC sensor is placed directly on the fiber facet and covers an area corresponding to the mode size of the fiber. The optical input to the fiber is the forward-propagating single mode of the fiber, and the output is the reflected mode. The measurand, which can be temperature, pressure, acceleration, bio-molecular associations, etc., modifies the PC or its surroundings so that the reflection is modulated. The optical readout can be over an extended spectral region, or at a single wave length.

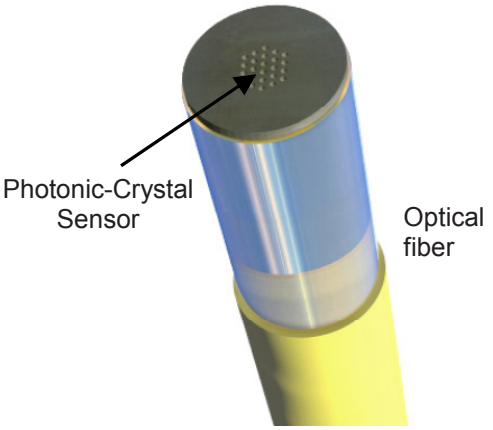


Figure 15.16 Schematic drawing of a fiber-tip sensor. The PC sensor, which may be designed to measure any one of a wide range of measurands, modifies the back reflected on the fiber. Fiber tip sensor. (Graphics courtesy Onur Kilic, Stanford University)

To work well in this configuration, the PC sensor must be designed to accept the full set of spatial frequencies of the fiber mode. As described in Chapter 5, the fiber mode can be expanded as a sum of plane waves of different directions. The PC sensor should respond similarly to all these incident plane waves to give an unambiguous output.

The sensor system of Fig. 15.16 receives its operating power over the fiber, and the fiber output signal is available at the far end, which can be as much as 100 km away. The fiber-tip sensor is therefore well suited to remote sensing. The size of the PC sensor allows it to be placed on the fiber without extending much beyond the facet. This compact solution gives access to remote locations that are inaccessible to bulkier traditional sensors.

Acoustic sensors are particularly well suited to fiber-tip implementations, because of the excellent mechanical properties of PC mirrors. The thickness of the PC mirror is on the order of a quarter of a wave length. This makes them very compliant and enables sensitive pressure sensors of the size of a standard single-mode fiber facet[b] as shown in Fig. 15.17.

The operating principle of this sensor is tried and true: The compliant PC mirror, together with the partially transmitting mirror on the fiber facet, create a Fabry-Perot resonator. Incident acoustic pressure deflects the PC mirror to modify the length, and therefore the reflection, of the Fabry-Perot resonator. The acoustic pressure can then be deduced from the magnitude of the reflected light on the fiber.

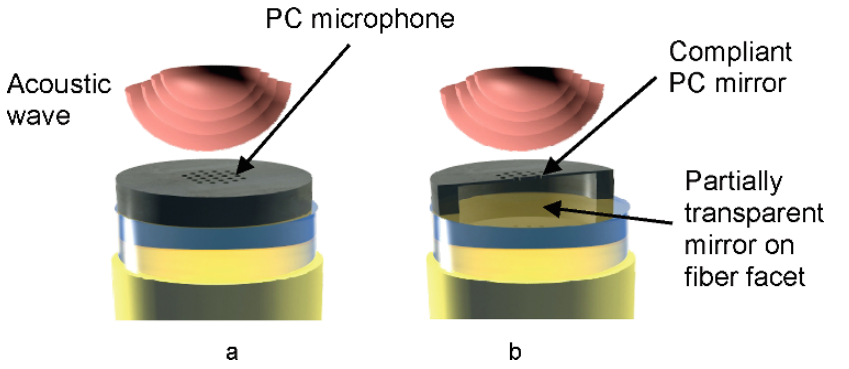


Figure 15.17 *Fiber tip microphone based on a Fabry-Perot resonator between a fiber-facet mirror and a compliant PC mirror. The PC mirror deflects under acoustic pressure to modify the back reflected light on the fiber. (Graphics courtesy Onur Kilic, Stanford University)*

Ideally both mirrors of the Fabry-Perot should be Photonic Crystals. In addition to the mechanical advantages, that allows the resonator to be short, and therefore stable, while also having high finesse and high sensitivity. The fiber-facet mirror can, however, be a thin metal film, because the fiber facet provides a stable substrate for fragile thin-film metal mirrors. That solution reduces sensitivity, be-

[b] The diameter of a single-mode fiber is standardized to *125 um*.

cause the reflectivity of the thin, partially-transmitting film is lower than for well-designed PC mirrors, but it is a practical design that avoids the complexity of creating PCs on fiber facets.

Figure 15.18 shows a simple implementation of a fiber-tip acoustic sensor. Here the Fabry-Perot is formed between a compliant PC mirror and a gold fiber-facet mirror. The PC mirror is fabricated on a Si chip and mounted on the fiber as shown. The chip is larger than the fiber facet, leading to a modest increase in size of the overall sensor system. This type of sensor has excellent characteristics both as a microphone [45,46] and as a hydrophone [47].

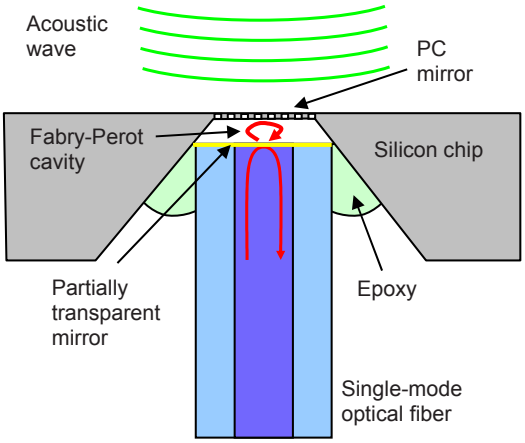


Figure 15.18 Implementation of a fiber-tip microphone/hydrophone. The PC mirror creates a low-loss, low-order F-P that combines high acoustic-pressure sensitivity with good temperature stability and a robust construction.

15.6 Summary of PC Devices and Systems

The main message of this chapter is that Photonic Crystals are very flexible and practical. They can be fabricated using IC and MEMS processing and integrated with IC and MEMS devices, as well as with fibers, and they have the flexibility to form the basis for a wide range of optical devices that are particularly well suited for miniaturized optical systems.

The chapter starts with a description of PC fabrication technology. The examples are chosen to illustrate the compatibility of PCs with IC and MEMS fabrication technology, and how the power and flexibility of the IC fabrication environment can be applied to the creation of PC devices. We then go on to list some of the PC components that are enabled by IC fabrication technology. Waveguides, waveguide devices, and optical resonators are obvious candidates for PC imple-

mentation, but the focus of our treatment is on free-space optical components including mirrors, filters, Fabry-Perot resonators, polarizers, and sensors. These components can be integrated into microphotonics systems. We illustrate that through two examples; PC MEMS scanners and PC displacement sensors. The chapter wraps up with a section on the advantages and opportunities of direct integration of PC devices on optical fibers.

The list of PC devices and systems is by no means exhaustive. In fact it barely scratches the surface, and with the rapid development in this field, the list will very soon be out dated. Novel PC devices, as well as PC-based improvements to traditional devices, are invented at an astonishing rate, and their advantages of fabrication, integration, and packaging are utilized in increasingly sophisticated systems. It is therefore expected that PC devices will become integral parts of all optical communication and sensing systems that require small size and superior performance, and that they will have significant impact on a wide range of commercial photonic applications.

Exercises

Problem 15.1 - CMOS PCs

Find a description of a standard, industrial CMOS process and show how it can be used to fabricate PC devices.

Problem 15.2 - Gopher MEMS

a. Show how you can use the GOPHER process (described in Fig. 15.3) to create MEMS electrostatic actuators. As shown in the figure, the process leaves the whole wafer connected, so you will have to add some processing to support the application of electrostatic voltages.
b. Show how your process can be used to fabricate a Fabry-Perot interferometer with PC reflectors.
c. Ditto for fabrication of a Grating Light Modulator with PC reflectors.

Problem 15.3 - Biosensors

a. Design an index sensor based on coupling to Surface Plasmons (described in Chapter 3.5.4)
b. Design a similarly-functioning index sensor based on coupling to guided resonances in 2-D PC. Use the models developed in Chapter 14.3.
c. Compare the two sensor principles. What advantages, if any, does the PC sensor have?
d. For each sensor, name two applications that it is ideally suited for.

Problem 15.4 - Dimensional Tuning of PCs

PCs can be tuned by moving them with respect to a reference as in Fig. 15.14, but they can also be tuned through changes in their dimensions. Consider a 2-D PC with a square unit cell in which there is a centered, circular hole. Explain qualitatively how the resonance frequency and life time of a guided resonance change when such a PC is

a. stretched equally in both directions in the plane
b. stretched along a row of holes
c. stretched along a diagonal
d. bowed
e. Which one of these tuning mechanisms are most efficient in terms of the force that has to be applied to establish a certain change?

Problem 15.5 - PC Fiber Sensors

a. What is the diffraction angle of the mode of a standard single-mode optical fiber at 1.55 um wavelength?
b. What does that mean for the plane-wave acceptance angle for PC devices that are designed to be placed directly on the facet of standard single-mode optical fibers?

Problem 15.6 - PC Accelerometer

Design an accelerometer based on fiber F-Ps as shown in Figs. 15.16-18. Assume that the silicon PC is *400 nm* thick, that the PC holes have negligible influence on mechanical characteristics (they are only present over the core of the fiber), and that the PC diaphragm is clamped along the periphery of the standard single-mode fiber (*125 um* in diameter). You will need a proof mass on the PC diaphragm to increase sensitivity to acceleration.

a. How would you affix the proof mass to the diaphragm so as to not interfere with the optical performance of the sensor?
b. How big must the proof mass be to create a deflection of *20 nm* in the center of the diaphragm under an acceleration of *1 g*? Use standard formulas for deflection of clamped diapraghs.
c. What are some of the applications of an accelerometer of this type?

References

1 K.B. Crozier, V. Lousse, O. Kilic, S. Kim ,W. Suh, S. Fan, O. Solgaard, "Air-bridged photonic crystal slabs at visible and near-infrared wave-

lengths", Physical Review B (Condensed Matter and Materials Physics); 15 March 2006; vol.73, no.11, p.115126-1-14.

2 S. Venkataraman, G. Schneider, J. Murakowski, S. Shi, and D. Prather, "Fabrication of three-dimensional photonic crystals using silicon micromachining", Applied Physics Letters, Vol. 85, p.2126 (2004).

3 S. Hadzialic, S. Kim, S. Basu Mallick, A. Sudbø, and O. Solgaard, "Monolithic Photonic Crystals", 2007 IEEE/LEOS Annual Meeting Conference, Orlando, Florida, 22-25 October, 2007.

4 W. W. Mullins, "Theory of thermal grooving", Journal of Applied Physics, vol. 28, 1957, pp. 333-338.

5 M.C.M. Lee, M.C. Wu, "Thermal annealing in hydrogen for 3-D profile transformation on silicon-on-insulator and sidewall roughness reduction", Journal of Microelectromechanical Systems, vol 15, no. 2, April 2006, pp.338-343.

6 D.K. Armani, T.J. Kippenberg, S.M. Spillane, and K.J. Vahala, "Ultra-high-Q toroid microcavity on a chip", Nature, vol. 421, no. 6926, Febuary 2003, pp. 925-928.

7 I. Mizushima, T. Sato, S. Taniguchi, Y. Tsunashima, "Empty-space-in-silicon technique for fabricating a silicon-on-nothing structure", Applied-Physics-Letters. 13 Nov. 2000; 77(20): pp. 3290-3292.

8 S. Kim, R. Kant, S. Hadzialic, R.T. Howe, O. Solgaard, "Interface Quality Control of Monolithic Photonic Crystals by Hydrogen Annealing", Conference on Lasers and Electro-Optics (CLEO) 2008, San Jose, CA, Paper CFY5, May 4-9, 2008.

9 S. Y. Lin, J. G. Fleming, D. L. Hetherington, B. K. Smith, R. Biswas, K. M. Ho, M. M. Sigalas, W. Zubrzycki, S. R. Kurtz, and J. Bur, "A three-dimensional photonic crystal operating at infrared wavelengths", Nature, 394, 252253, 1998. Y. Lin et al, Nature, 394, pp. 251, (1998).

10 M. Qi, E. Lidorikis, P. T. Rakich, S. G. Johnson, J.D. Joannopoulos, E. P. Ippen and H. I. Smith, "A three-dimensional optical photonic crystal with designed point defects" Nature, 429, pp. 538-542, (2004).

11 N. D. Lai, J. H. Lin, and C. C. Hsu, "Fabrication of large size photonic crystal templates by holographic lithography technique, IEEE/LEOS International Conference on Optical MEMS and Nanophotonics, pp. 155-156, Aug. 2007.

12 F Garcia-Santamaria, C López, F Meseguer, and F López, "Opal-like photonic crystal with diamond lattice", Appl. Phys. Letters 79, pp. 2309 (2001).

13 C.L. Hoy, N.J. Durr, P.Chen, W. Piyawattanametha, H. Ra, O. Solgaard, A. Ben-Yakar, "Miniaturized probe for femtosecond laser microsurgery and two-photon imaging" ,Optics Express, Vol. 16, Issue 13, pp. 9996-10005, June 20, 2008.

14 W. Suh, M. F. Yanik, O. Solgaard, and S.-H. Fan, "Displacement-Sensitive Photonic Crystal Structures Based on Guided Resonance in Photonic Crystal Slabs," Appl. Phys. Lett., Vol. 82 (13), 31 March 2003, pp. 1999-2001.
15 C.F.R. Mateus, M.C.Y. Huang, L. Chen, C.J. Chang-Hasnain, Y. Suzuki, "Broad-Band Mirror (1.12–1.62 um) Using a Subwavelength Grating", IEEE Photonics Technology Letters, vol. 16, no. 7, July 2004, pp. 1676-1678.
16 S. Kim, S. Hadzialic, A. Sudbo, O. Solgaard, "Single-film Broadband Photonic Crystal Micro-mirror with Large Angular Range and Low Polarization Dependence", Conference on Lasers and Electro-Optics (CLEO) 2007, Baltimore, MD, Paper CThP7.
17 W. Suh, O. Solgaard, S. Fan, "Displacement sensing using evanescent tunneling between guided resonances in photonic crystal slabs", Journal of Applied Physics, 98, article 033102, 1 August 2005.
18 D.W. Carr, J.P. Sullivan, T.A. Friedmann, "Laterally deformable nanomechanical zeroth-order gratings: anomalous diffraction studied by rigorous coupled-wave analysis, Optics Letters, 28, No. 18, 2003, pp. 1636-1638.
19 B.E.N. Keeler, D.W. Carr, J.P. Sullivan, T.A. Friedmann, J.R. Wendt, "Experimental demonstration of a laterally deformable nanoelectromechanical system grating transducer", Optics Letters; 1 June 2004; vol.29, no.11, p.1182-1184.
20 O. Kilic, S. Fan, O. Solgaard, "Analysis of guided-resonance based polarization beam splitting in photonic crystal slabs", Journal of the Optical Society of America A, vol. 25, no. 11, November 2008.
21 R. Magnusson, D. Wawro, "Guided-mode resonance sensors for biochemical screening", LEOS 2007. 20th Annual Meeting of the IEEE Lasers and Electro-Optics Society, 21-25 Oct. 2007, Lake Buena Vista, FL, USA; p.228-229.
22 M. Lee, P.M. Fauchet, "Two-dimensional silicon photonic crystal based biosensing platform for protein detection", OPTICS EXPRESS, vol. 15, no. 8, pp. 4530-4535, 16 April 2007.
23 M. Lee, P.M. Fauchet, "Nanoscale microcavity sensor for single particle detection", Optics Letters, vol. 32, no. 22, pp. 3284-3286, 15 November 2007. Erratum published in Optics Letters, vol. 33, no. 7, p. 756, 1 April 2008.
24 S.M. Weiss, H. Ouyang, J. Zhang, P.M. Fauchet, "Electrical and thermal modulation of silicon photonic bandgap microcavities containing liquid crystals", Optics Express, vol. 13, no. 4, 21 February 2005, pp. 1090-1097.
25 E.A. Camargo, H.M.. H. Chong, R.M. De La Rue, "2D Photonic crystal thermo-optic switch based on AlGaAs/GaAs epitaxial structure", Optics Express, vol. 12, no. 4, 23 February 2004, pp. 588-592.
26 T.G. Euser, A.J. Molenaar, J.G. Fleming, B. Gralak, A. Polman, W.L. Vos, "All-optical octave-broad ultrafast switching of Si woodpile photonic band gap crystals", PHYSICAL REVIEW B, vol. 77, pp. 115214-1-6, 26 March 2008.

27 I. Fushman, E. Waks, D. Englund, N. Stoltz, P. Petroff, J. Vuckovic, "Ultrafast nonlinear optical tuning of photonic crystal cavities", Applied Physics Letters, 2007 Applied Physics Letters, vol.90, no.9, 26 Febuary 2007, pp. 91118-1-3.
28 W. Park, J.-B. Lee, "Mechanically tunable photonic crystal structure", Applied Physics Letters, vol. 85, No. 21 22 November 2004, 4845-4847.
29 T. Takahata, K. Hoshino, K. Matsumoto, I. Shimoyama, "Photonic crystal attenuator with a flexible waveguide and nano-rods", Proceedings of the IEEE MEMS Conference. Istanbul, Turkey, 22-26 January 2006, pp. 834-837.
30 W. Suh, M. F. Yanik, O. Solgaard, and S.-H. Fan, "Displacement-Sensitive Photonic Crystal Structures Based on Guided Resonance in Photonic Crystal Slabs," Appl. Phys. Lett., Vol. 82 (13), 31 March 2003, pp. 1999-2001.
31 I. Marki, M. Salt, S. Gautsch, U. Staufer, H.P. Herzig, N. de Rooij, "Tunable microcavities in two dimensional photonic crystal waveguides", IEEE/LEOS Optical MEMs 2005, 1-4 Aug. 2005, Oulu, Finland; p.109-10.
32 X. Letartre, J. Mouette, J. L. Leclercq, P. Rojo Romeo, C. Seassal, and P. Viktorovitch, "Switching Devices With Spatial and Spectral Resolution Combining Photonic Crystal and MOEMS Structures", Journal of Lightwave Technology, Vol. 21, No. 7, July 2003, pp. 1691-1699.
33 M.-C.M. Lee, D. Hah, E.K. Lau, H. Toshiyoshi, M. Wu, "MEMS-Actuated Photonic Crystal Switches", IEEE Photonics Technology Letters, vol. 18, no. 2, January 15, 2006, pp. 358-360.
34 X. Letartre, J. Mouette, J. L. Leclercq, P. Rojo Romeo, C. Seassal, and P. Viktorovitch, "Switching Devices With Spatial and Spectral Resolution Combining Photonic Crystal and MOEMS Structures", J. Lightwave Technology, 21, 2003, pp. 1691-1699.
35 W. Suh, O. Solgaard, S. Fan, "Displacement sensing using evanescent tunneling between guided resonances in photonic crystal slabs", Journal of Applied Physics, vol. 98,issue 3, article 033102, 1 August 2005, (4 pages).
36 D.W. Carr, J.P. Sullivan, T. A. Friedmann, "Laterally deformable nanomechanical zeroth-order gratings: anomalous diffraction studied by rigorous coupled-wave analysis", Optics Letters, vol. 28, no. 18, September 15, 2003, pp. 1636-1638.
37 J. Provine, J. Skinner, D.A. Horsley, "Subwavelength Metal Grating Tunable Filter", Technical Digest of 19th IEEE International Conference on MicroElectroMechanical Systems 2006 (MEMS 2006), Istanbul, Turkey, January 22-26, 2006, pp. 854-857.
38 R. Magnusson, Y. Ding, "MEMS Tunable Resonant Leaky Mode Filters", Ieee Photonics Technology Letters, vol. 18, no 14, July 15, 2006, pp. 1479-1481.
39 W. Suh, S. Fan, "Mechanically switchable photonic crystal filter with either all-pass transmission or flat-top ref lection characteristics", Optics Letters, 28, 2003, pp. 1763-1765.

40 S. Nagasawa, T. Onuki, Y. Ohtera, H. Kuwano, "MEMS Tunable Optical Filter Using Auto-Cloned Photonic Crystal", Technical Digest of 19th IEEE International Conference on MicroElectroMechanical Systems 2006 (MEMS 2006), Istanbul, Turkey, January 22-26, 2006, pp. 858-861.

41 C. Seassal, C. Monat, J. Mouette, E. Touraille, B. Ben Bakir, H. Takashi Hattori, J.-L. Leclercq, X. Letartre, P. Rojo-Romeo, P. Viktorovitch, "InP Bonded Membrane Photonics Components and Circuits: Toward 2.5 Dimensional Micro-Nano-Photonics", IEEE J. Selected Topics in Quantum Electronics, 11, 2005, pp. 395-407.

42 F. Raineri, C. Cojocaru, R. Raj, P. Monnier, A. Levenson, C. Seassal, X. Letartre, P. Viktorovitch, "Tuning a two-dimensional photonic crystal resonance via optical carrier injection", Optics Letters, 30, 2005, pp. 64-66.

43 I.-W. Jung, S. Kim, O. Solgaard, "High Reflectivity Broadband Photonic Crystal Mirror MEMS Scanner ", Transducers &Eurosensors'07, The 14th International Conference on Solid-State Sensors, Actuators and Microsystems, Lyon, France, June 10-14, 2007, pp. 1513-1516.

44 S. Hadzialic, S. Kim, A. Sudbø, O. Solgaard, "Displacement Sensing with a Mechanically Tunable Photonic Crystal", 2007 IEEE/LEOS Annual Meeting Conference Proceedings, Lake Buena Vista, Florida, 21-25 October 2007, pp. 345-346.

45 O. Kilic, M. Digonnet, G. Kino, O. Solgaard, "External Fiber Fabry-Perot Acoustic Sensor Based on Photonic Crystal Mirror", 18th International Conference on Optical Fiber Sensors Topical Meeting (OFS-18), Cancun, Mexico, October 23-27, 2006.

46 O. Kilic, M. Digonnet, G. Kino O. Solgaard, "External fibre Fabry–Perot acoustic sensor based on a photonic-crystal mirror", IOP Publishing Measurement Science and Technology, vol, 18, 12 September 2007, pp. 3049–3054.

47 O. Kilic, M. Digonnet, G. Kino, O. Solgaard, "Photonic-crystal-diaphragm-based fiber-tip hydrophone optimized for ocean acoustics", 19th International Conference on Optical Fibre Sensors, April 14-18, 2008, Perth, Australia, Proceedings edited by David Sampson, Stephen Collins, Kyunghwan Oh, Ryozo Yamauchi, Proceedings of SPIE Vol. 7004, pp. 700405-1-4.

Appendix A: Geometrical Optics

A.1 Introduction to Geometrical Optics

This appendix summarizes the fundamentals of Geometrical Optics and introduces Geometrical-Optics model of common optical devices that are used in this book. Geometrical Optics, or Ray Optics, is based on the Law of Reflection and Snell's Law of Refraction that are described in Chapter 2. These laws are derived from Maxwell's equations and are valid for plane waves, i.e. electromagnetic waves that have no transversal variation. Geometrical Optics introduces the concept of an optical ray, which is a light beam of zero cross section, and postulates that rays propagate according to the Laws of Reflection and Refraction. This means that Geometrical Optics disregards wave diffraction, and that it is an accurate model for systems that are limited not by diffraction, but by other effects, e.g. lens aberrations. Typically, we find that Geometrical Optics is useful for optical systems with apertures that are much larger than the wavelength of light, and that we must be careful when applying this theory to miniaturized optical devices.

In the first part of this appendix, we describe the operation of lenses in the Geometrical Optics perspective. Simple considerations allow us to derive a set of first-order rules for lens-system analysis. In the last part of the appendix, we give the ABCD matrices for a number of common optical components.

A.2 Geometrical Optics Treatment of Lenses

A.2.1 Lens – Ray Picture

The operation of lenses can be understood by tracing individual rays through the lens. The rays all bend in different ways at the air-lens interface, and the shape of the lens surfaces is chosen such that each ray passes through the focus as shown in Fig. A.1.

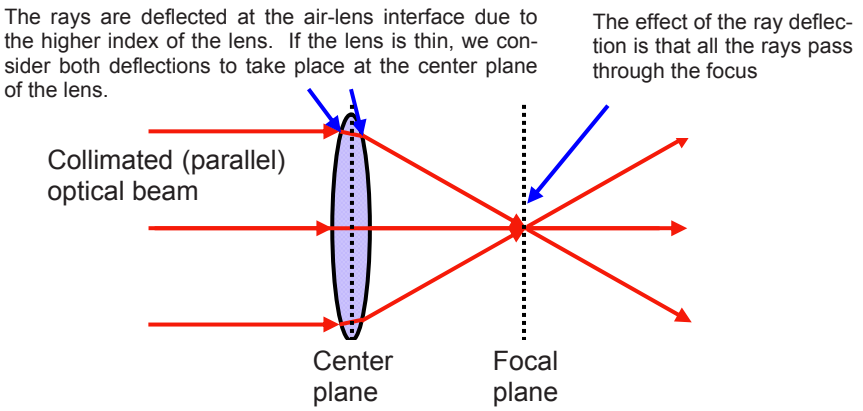


Figure A.1. *The shape of a lens is designed to refract all incoming parallel optical rays to a common point in the focal plane.*

A.2.2 Lenses – Wave Picture

The operation of a lens can also be understood by considering the wavefronts of the light passing through the lens. The lens is delaying the center part of the beam with respect to the sides such that all parts of the beam arrive at the focus in phase. In other words, all parts of the beam are interfering constructively at the focus, which leads to high intensity at this point. This wave-picture view of lens operation is illustrated in Fig. A.2.

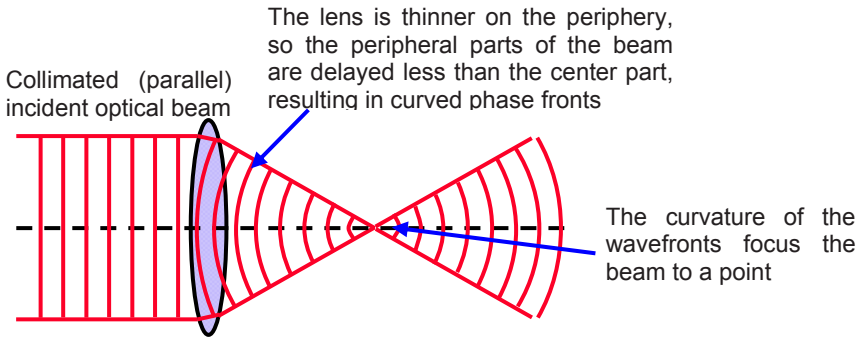


Figure A.2. *The thick center part of a convex lens creates spherical wavefronts that converge to a common point in the focal plane.*

A.2.3 Ray Tracing

To trace rays through ideal lenses we need only two simple facts: (1) Each ray goes through the focal plane at the same point as its parallel central ray, and (2) central rays (rays passing though the exact center of the lens) are not deflected. These two rules are illustrated in Fig. A.3.

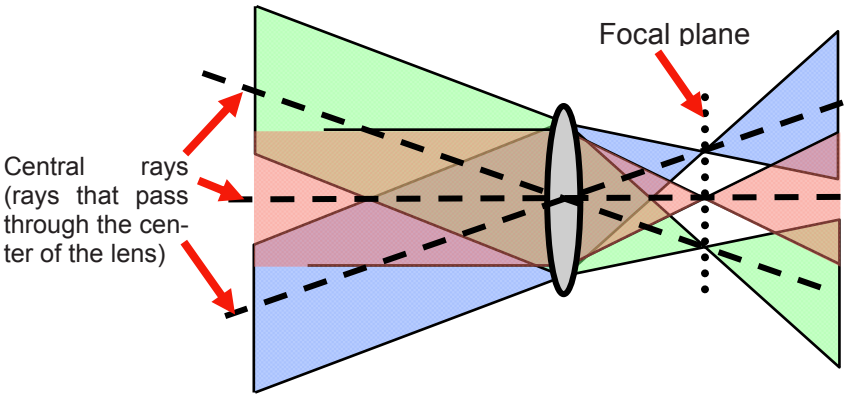


Figure A.3. The intersection of the straight central ray determines where all parallel rays meet in the focal plane.

Using these two facts we can draw the image-formation of a lens as shown in Fig. A.4. This construction is very useful for first order analysis of lens systems. Inspection of the drawing gives us the thin lens equation

$$\frac{1}{a}+\frac{1}{b}=\frac{1}{f} \tag{A.1}$$

We also see that the magnification of the imaging system is given by

$$M=\frac{b}{a} \tag{A.2}$$

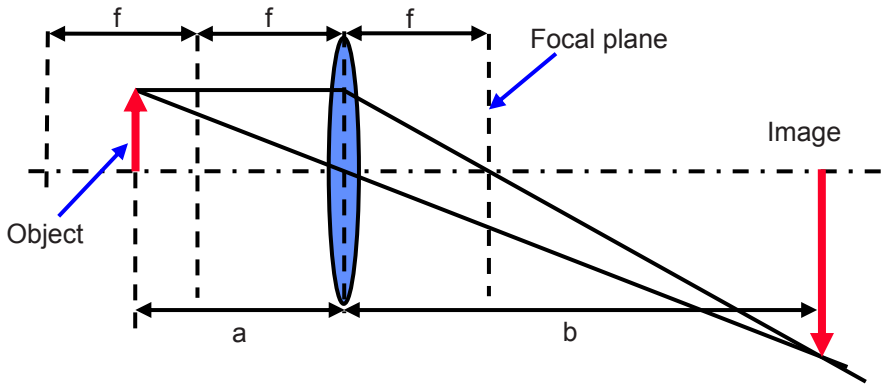


Figure A.4. Image construction with an ideal lens.

Image construction as shown in Fig. A.4 falls in one of five regimes:

1. If the object distance exceeds the focal length ($a<f$), then rays from a common object point diverge after the lens, so no image is formed.
2. If the object distance equals the focal length ($a=f$), then rays from a common object point are parallel after the lens, so we have a collimated beam, but no image.
3. If the object distance falls between one and two focal lengths ($f<a<2f$), then the lens forms an image with a magnification larger than unity ($M>1$).
4. If the object distance equals two focal lengths ($a=2f$), then the lens forms an image with a magnification of unity ($M=1$).
5. If the object distance is larger than two focal lengths ($a>2f$), then the lens forms an image with a magnification less then unity ($M<1$).

A.3 ABCD Matrices

For modeling of systems consisting of several optical elements, it is convenient to use the ABCD-matrix formalism. The trajectory of a ray through a given optical device depends on the incident position and slope. If the output position and slope are linearly dependent on the input position and slope (or approximately so), we say that the ray is paraxial. The linear dependence is equivalent to the approximation $\sin\theta \approx \tan\theta \approx \theta$. In the ABCD-matrix model an optical device, the distance and slope of rays in the output plane of the device are related to the distance and slope in the input plane

$$\begin{aligned} r_2 &= Ar_1 + Br_1' \\ r_2' &= Cr_1 + Dr_1' \end{aligned} \Rightarrow \begin{bmatrix} r_2 \\ r_2' \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} r_1 \\ r_1' \end{bmatrix} \tag{A.3}$$

This is illustrated in Fig. A.5.

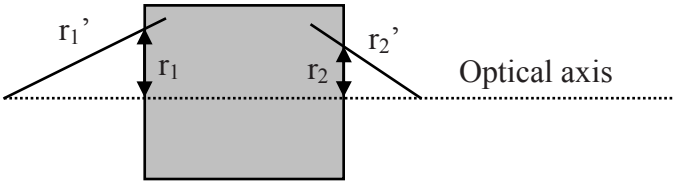


Figure A5. *The ABCD matrix of an optical device relates the distance and slope of outputs rays to the distance and slop of input rays.*

A.3.1 Free Space

Figure A.6 shows that the ABCD matrix for a fee-space propagation segment is given by

$$\begin{matrix} r_2 = r_1 + Lr_1' \\ r_2' = r_1' \end{matrix} \Rightarrow \begin{bmatrix} 1 & L \\ 0 & 1 \end{bmatrix} \tag{A.4}$$

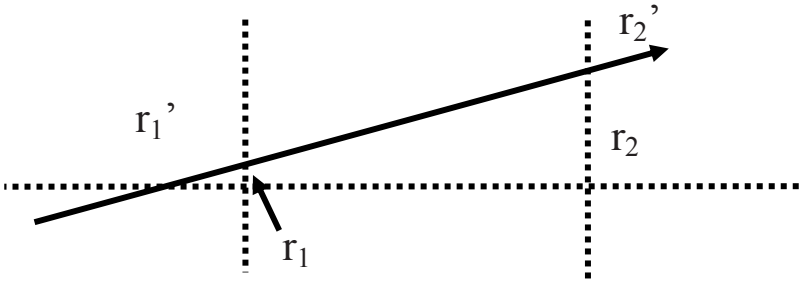


Figure A.6. *Relationship between distance and slope on the input and output of a free-space segment of an optical system.*

A.3.2 Slab of Index n

If the free-space propagation is through a slab of index n, then the distances and slopes are related as shown in Fig. A.7. We see that

$$n_{ext} \sin\left(r_{1ext}'\right) = n_{\text{int}} \sin\left(r_{1\text{int}}'\right) \Rightarrow r_{1ext}' = nr_{1\text{int}}' \tag{A.5}$$

and the ABCD matrix is

$$\begin{matrix} r_2 = r_1 + \dfrac{L}{n} r_1' \\ r_2' = r_1' \end{matrix} \Rightarrow \begin{bmatrix} 1 & \dfrac{L}{n} \\ 0 & 1 \end{bmatrix} \tag{A.6}$$

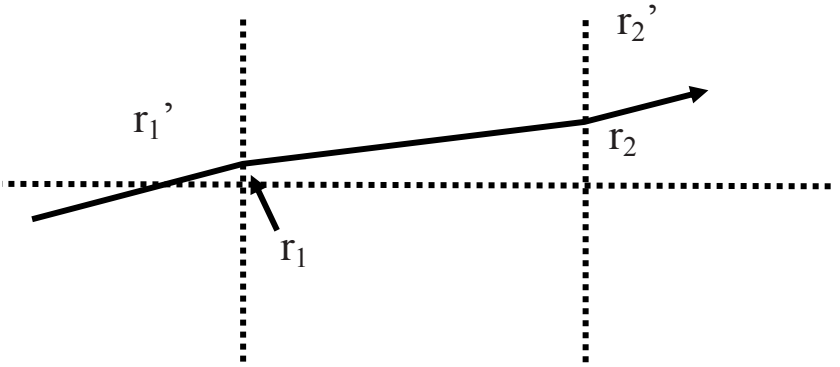


Figure A.7. Relationship between distance and slope on the input and output of a uniform slab of index n.

A.3.3 Thin Lens

We find the ABCD matrix of a thin lens from Fig. A.8. Note that for a thin lens we have $r_1=r_2$, so

$$r_1' = \frac{r_1}{a} \text{ and } r_2' = -\frac{r_2}{b} \Rightarrow -b \cdot r_2' = a \cdot r_1' \Rightarrow r_2' = -\frac{a}{b} \cdot r_1' \tag{A.7}$$

Substituting this expression into the lens equation $\left(\dfrac{1}{f} = \dfrac{1}{a} + \dfrac{1}{b} \right)$ we find

$$r_2' = a\left(-\frac{1}{f} + \frac{1}{a} \right) \cdot r_1' = \left(-\frac{a}{f} + 1 \right) \cdot r_1' = \left(-\frac{r_1}{r_1' f} + 1 \right) \cdot r_1' = -\frac{r_1}{f} + r' \tag{A.8}$$

Combined with $r_1=r_2$, this gives the ABCD matrix

$$\begin{matrix} r_2 = r_1 \\ r_2' = -\dfrac{r_1}{f} + r_1' \end{matrix} \Rightarrow \begin{bmatrix} 1 & 0 \\ -\dfrac{1}{f} & 1 \end{bmatrix} \tag{A.9}$$

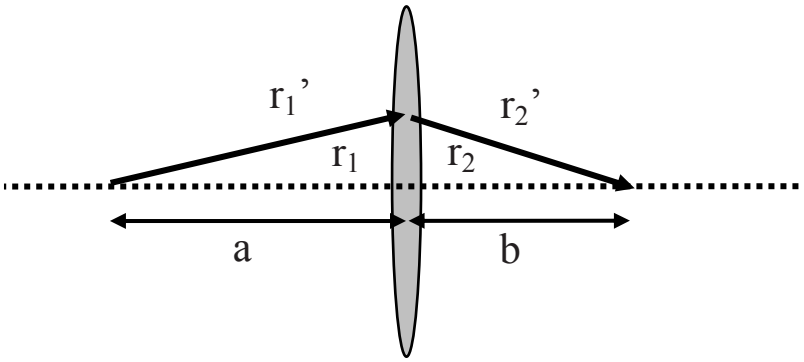


Figure A.8. Rays through a thin lens.

A.3.4 Curved Mirror

Figure A.9 shows that the focal length for a spherical mirror is $f=R/2$, where R is the radius of curvature of the mirror. By extension form the thin lens, we then find the following ABCD matrix for the spherical mirror

$$\begin{matrix} r_2 = r_1 \\ r_2' = -\dfrac{2r_1}{R} + r_1' \end{matrix} \Rightarrow \begin{bmatrix} 1 & 0 \\ -\dfrac{2}{R} & 1 \end{bmatrix} \tag{A.10}$$

Notice that in this as in the other examples we have $AC - DB = 1$.

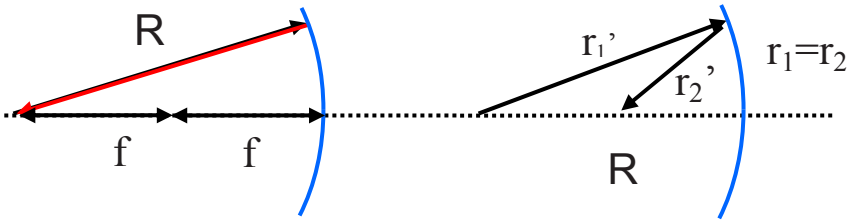


Figure A.9. Rays reflecting off a spherical mirror. As for the thin lens, there is no lateral shift of the rays at the mirror, only a change in direction.

A.3.5 Combinations of Elements

The strength of the ABCD-matrix formulation is that the matrix of a combination of optical elements is found by simple matrix multiplication as shown in the example in Fig. A.10.

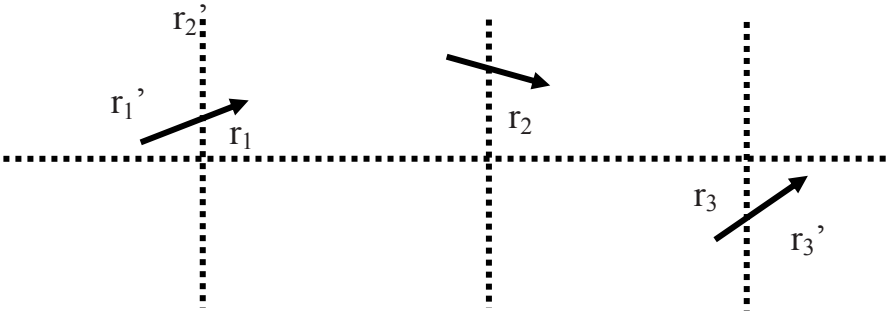


Figure A.10. The ABCD matrix of a two-element system is found by multiplying the two ABCD matrices. The extension to multiple elements is straightforward.

The position and slope at *plane 2* can be expressed in terms of the position and slope at *plane 1* in the following way

$$\begin{bmatrix} r_2 \\ r_2' \end{bmatrix} = \begin{bmatrix} A_{12} & B_{12} \\ C_{12} & D_{12} \end{bmatrix} \begin{bmatrix} r_1 \\ r_1' \end{bmatrix} \tag{A.11}$$

Similarly, the position and slope at *plane 3* can be expressed in terms of the position and slope at *plane 2*

$$\begin{bmatrix} r_3 \\ r_3' \end{bmatrix} = \begin{bmatrix} A_{23} & B_{23} \\ C_{23} & D_{23} \end{bmatrix} \begin{bmatrix} r_2 \\ r_2' \end{bmatrix} \tag{A.12}$$

By combining these we find

$$\begin{bmatrix} r_3 \\ r_3' \end{bmatrix} = \begin{bmatrix} A_{23} & B_{23} \\ C_{23} & D_{23} \end{bmatrix} \begin{bmatrix} A_{12} & B_{12} \\ C_{12} & D_{12} \end{bmatrix} \begin{bmatrix} r_1 \\ r_1' \end{bmatrix} \tag{A.13}$$

A.3.6 Reverse Transmission:

By inverting the ABCD matrix, we find the following rule for reverse transmission

$$\begin{bmatrix} r_1 \\ -r_1' \end{bmatrix} = \begin{bmatrix} A' & B' \\ C' & D' \end{bmatrix} \begin{bmatrix} r_2 \\ -r_2' \end{bmatrix} \Rightarrow \begin{bmatrix} r_2 \\ r_2' \end{bmatrix} = \begin{bmatrix} D' & B' \\ C' & A' \end{bmatrix} \begin{bmatrix} r_1 \\ r_1' \end{bmatrix}$$
$$\Rightarrow \begin{bmatrix} A' & B' \\ C' & D' \end{bmatrix} = \begin{bmatrix} D & B \\ C & A \end{bmatrix} \tag{A.14}$$

Appendix B: Electrostatic Actuation

B.1 The Parallel Plate Capacitor

Our investigation of electrostatic actuators starts with one of the most basic, but also most common MEMS devices; the parallel-plate electrostatic actuator. In our treatment we will use some simplifying assumptions about the electric field. The simplified results we obtain contain all the important physics of the correct solution. Consider the schematic of a parallel-plate capacitor shown in Fig. B.1.

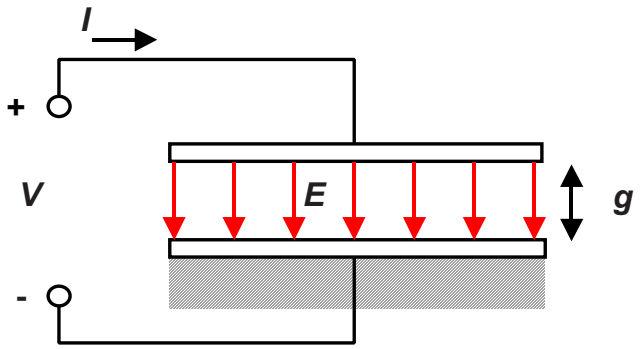


Figure B.1. Parallel plate capacitor. The lower plate is fixed, while the upper plate can move.

As a first approximation, we will assume that the electrical field is uniform between the plates of the capacitor, and zero outside. (This is of course not a completely correct solution. This electrical field distribution has non-zero curl at the edges of the capacitor plates in violation of Maxwell's equations.) The uniform electric field between the plates is then pointing down towards the lower plate, and it has the magnitude

$$E = \frac{Q}{\varepsilon A} \tag{B.1}$$

where A is the area of one capacitor plate, and Q is the magnitude of the charge on each plate.

With the signs shown in Fig. B.1, the charge is negative on the lower plate and positive on the upper. The voltage across the capacitor is simply the product of the E-field and the gap

$$V = E \cdot g = \frac{g \cdot Q}{\varepsilon A} \tag{B.2}$$

and the capacitance is the ratio of the charge and the voltage

$$C = \frac{Q}{V} = \frac{\varepsilon \cdot A}{g} \tag{B.3}$$

B.1.1 Energy Storage in Parallel-Plate Capacitors

If the capacitor plates are fixed, then the stored energy in the capacitor is given by

$$W(Q) = \int_0^Q V dQ = \int_0^Q \frac{Q}{C} dQ = \frac{Q^2}{2C} = \frac{1}{2} C \cdot V^2 = \frac{Q^2 g}{2A\varepsilon} \tag{B.4}$$

We can also find the stored energy by considering the force attracting the capacitor plates to each other. The field creates an electrostatic force that tries to bring the plates together. The magnitude of the force on each plate is:

$$F = \frac{QE}{2} = \frac{Q^2}{2A \cdot \varepsilon} = \frac{A \cdot \varepsilon \cdot V^2}{2g^2} \tag{B.5}$$

Note that when expressed in terms of the charge, Q, the force is independent of the gap between the plates!

The factor *2* in the denominator might seem surprising. You might ask: Isn't the force the product of the field and the charge, and therefore simply given by $F=EQ$? Remember that in the definition of the electric field (the classical definition of the electric field is that it is a vector field that when multiplied by the magnitude of a test charge, gives the force on that charge), we specify that the charge that is subject to the force of the electric field, is a ***test charge*** that does not itself influence the electric field.

To see how this definition leads to the factor of *2* consider the force on an infinitesimal charge on a conductor in an electric field as illustrated in Fig. B.2. The field at the conductor surface is partly due to the distant charges and partly to the charges on the conductor surface. As for all conductors, the electric field is normal to the surface, and the field is terminated on the surface (i.e. there is no field

inside the conductor). Inside the conductor the surface charges must be exactly large enough to cancel the field contribution from the distant charges. This means that the two contributions are equal. Right outside the surface these two equal contributions add up, so that the total field is twice as large as it would have been if only the distant charges were present. If we removed the surface charge and placed a test charge (i.e. one that does not change the field) there instead, it would see half the field, so the force on the surface charge on the conductor is given by

$$F = Q \cdot E_{\text{distant}} = \frac{Q \cdot E_{\text{total}}}{2} \tag{B.6}$$

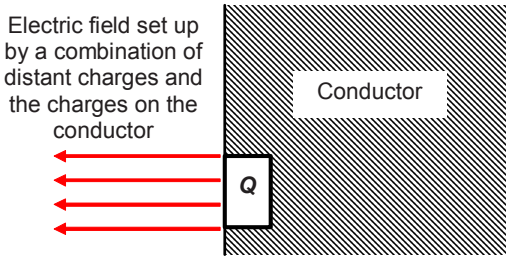


Figure B.2. Positively charged conductor in the presence of distant charges. The electric field just outside conductor is made up of contributions from the distant charges as well as the charges on the conductor.

An alternative way of explaining the factor of two in the expression for the force is to more carefully consider the product of a step function (the voltage is a step function going from zero to its constant value at the surface) and a delta function (the charge density is a delta function at the surface). If we instead use a more realistic approximation of a linearly increasing voltage across a surface region with a constant charge density, the factor of 0.5 appears naturally in the calculated force. Assume that the electric field is linearly increasing through a surface region with a thickness, Δz, and a constant charge density, ρ. The force can then be expressed

$$F = A \cdot \int_0^{\Delta z} E(z) \cdot \rho \cdot dz = A \cdot \int_0^{\Delta z} E_0 \frac{z}{\Delta z} \cdot \rho \cdot dz = A \cdot \frac{E_0 \cdot \rho}{\Delta z} \int_0^{\Delta z} z \cdot dz =$$
$$A \cdot \frac{E_0 \cdot \rho}{\Delta z} \frac{\Delta z^2}{2} = \frac{E_0 \cdot \rho \cdot A \cdot \Delta z}{2} = \frac{E_0 \cdot Q}{2} \tag{B.7}$$

The factor of 2 now appears simply as a consequence of the linearly increasing force.

The energy stored in the capacitor is equal to the energy needed to pull the two plates apart till their separation equals the final gap, g. The stored energy can then be expressed

$$W(g) = \int_0^g F dg = F \cdot g = \frac{Q^2 g}{2A\varepsilon} \tag{B.8}$$

This follows directly from the expression for the force. The force is independent of the gap size g, so we find the work required to increase the gap from zero to g as the product of the constant force and the distance over which it is applied. Note that this expression is identical to the one we found by considering electrical energy that must flow into the capacitor to increase the charge from zero to Q.

B.2 The Parallel Plate Electrostatic Actuator

This preceding simple treatment of the parallel-plate capacitor emphasizes the fact that it is both an electric and a mechanical device. It is indeed a transducer, in which electrical energy can be transformed into mechanical energy and vice versa. Usually we don't worry about the mechanical aspects of the capacitors we use in electronics, because the plates are both fixed so there is only insignificant mechanical energy storage. (In principle, of course, any real capacitor will have its plates separated by a mechanical structure with a finite compliance, so there will indeed be some stored mechanical energy).

In the practical implementation of the electrostatic actuator, however, both electrostatic and mechanical energy storage are important. Mechanical energy can be stored as potential energy, kinetic energy, or both. To model the energy storage, we include a spring in the physical model, and we attribute a mass to the moving plate. The physical model then looks as shown in Fig. B.3.

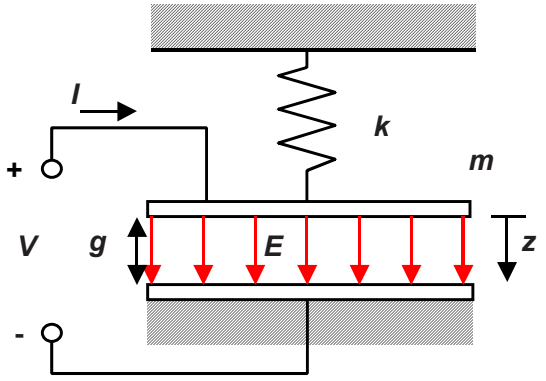


Figure B.3. *Physical model of a parallel plate capacitor with mechanical energy storage. The lower plate is fixed, while the upper plate can move. Energy can be stored in the spring (potential energy), or in the movements of the plate (kinetic energy).*

B.2.1 Charge Control

With the inclusion of the mechanical spring (we will not consider the mass and the damping until we are ready to model the dynamics of the actuator), the electrostatic actuator is modeled as shown in Fig. B.4. The electrical source in this system is a current source, which allow us to control the charge on the parallel-plate capacitor by switching the source as indicated.

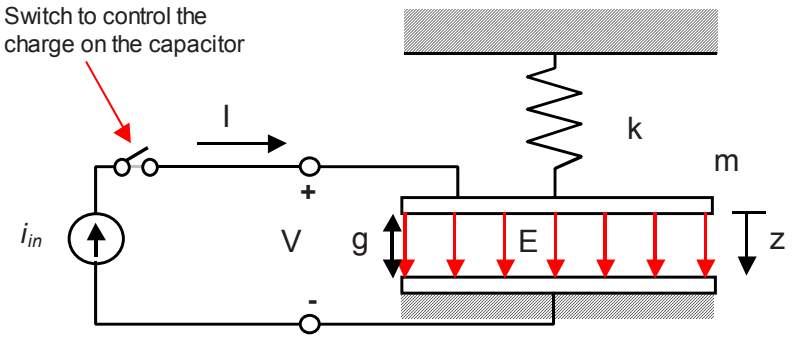


Figure B.4. *Electrostatic actuator model incorporating the two-port parallel-plate capacitor and a capacitor representing the mechanical spring.*

The charge on the capacitor is the integration of the current. Assuming that we start with an uncharged capacitor at $t=0$, we find:

$$Q = \int_0^t i_{in}(t)dt \tag{B.9}$$

The charge determines the electrostatic force on the plates. In principle, we can therefore control the force of the actuator by controlling the current as a function of time.

In equilibrium, the electrostatic force must match the spring force.

$$F = \frac{QE}{2} = \frac{Q^2}{2\varepsilon A} = k \cdot z \Rightarrow z = \frac{Q^2}{2k\varepsilon A} \tag{B.10}$$

We see that the displacement is a quadratic function of the stored charge, i.e. it is a ***monotonically increasing function that is stable throughout its range of validity***.

The gap in the actuator can be expressed as

$$g = g_0 - z = g_0 - \frac{Q^2}{2k\varepsilon A} \tag{B.11}$$

which leads to the following expression for the voltage

$$V = E \cdot g = \frac{Q}{\varepsilon A} \cdot g = \frac{Q}{\varepsilon A}\left(g_0 - \frac{Q^2}{2kA\varepsilon} \right) \tag{B.12}$$

The expression for the magnitude of the gap shows us that if we increase the charge to a sufficiently high value, the gap goes to zero. That happens when the charge reaches the value:

$$\hat{Q} = \sqrt{g_0 \cdot 2k\varepsilon A} \tag{B.13}$$

Notice that the voltage goes to zero for this value of the charge. These relationships are illustrated in Fig. B.5.

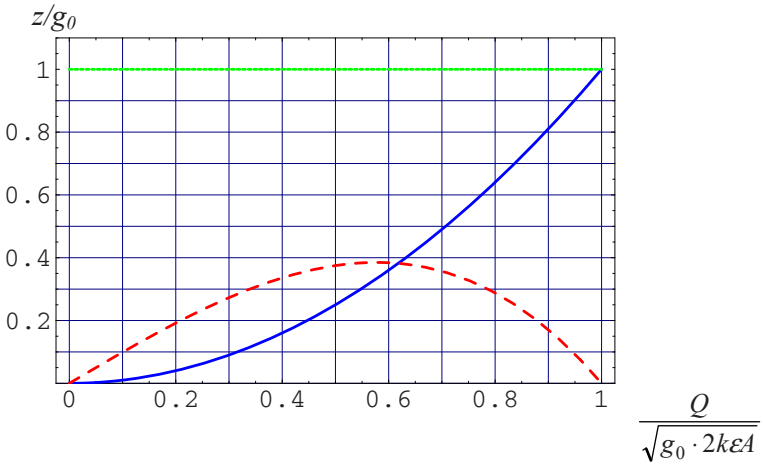


Figure B.5. Plot of normalized deflection (z/g_0, solid) and voltage ($V \cdot \frac{1}{g_0}\sqrt{\frac{A\varepsilon}{g_0 2k}}$, dashed) vs. normalized charge ($\frac{Q}{\sqrt{g_0 \cdot 2k\varepsilon A}}$) for a charge-controlled, electrostatic parallel-plate actuator.

We see that the deflection is well behaved, increasing monotonically from zero to the full value of the gap, when the charge is increased from zero to the critical value. The voltage reaches its maximum value

$$V_{\max} = \frac{2 \cdot g_0}{3} \sqrt{\frac{2 \cdot g_0}{3} \frac{k}{A\varepsilon}} = \sqrt{\frac{8 \cdot g_0^3}{27} \frac{k}{A\varepsilon}} \tag{B.14}$$

when

$$Q = \sqrt{\frac{g_0 \cdot 2k\varepsilon A}{3}} \tag{B.15}$$

as can be verified by differentiation of the voltage expression.

The charge-controlled parallel-plate actuator has many desirable characteristics. It is simple and the deflection can be controlled over the whole electrode gap. The MEMS designer faces some practical difficulties in implementing this actuator, however. Most problematic is the fact that most typical MEMS capacitors are on the order of femto-Farads, i.e. much smaller than the capacitance associated with bond pads and off-chip connections,. This means that the switching that controls the charge on the MEMS capacitor has to be on the chip. If it is off-chip, then the large stray capacitances will vary too much to allow accurate charge control.

B.2.2 Voltage Control

The voltage controlled electrostatic actuator, shown in Fig. B.6, is easier to implement, and therefore the design of choice in practice. Unfortunately, the ease of implementation comes at a cost. For many applications voltage control has less favorable characteristics than charge control.

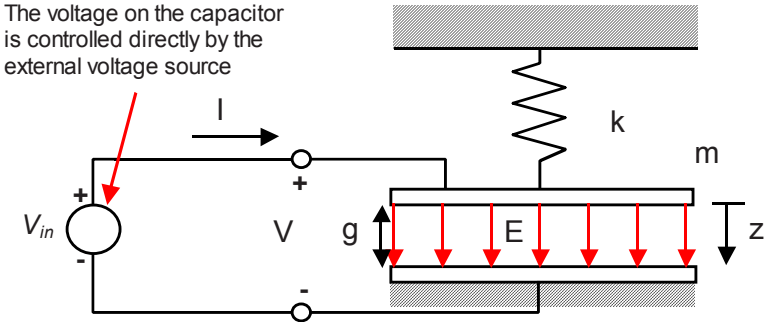


Figure B.6. Model of electrostatic actuator with voltage control.

In this case the charge on the capacitor is

$$Q = V \cdot C = \frac{VA\varepsilon}{g} \tag{B.16}$$

The charge determines the force, as before, and the electrostatic force must be matched by the spring force.

$$F = \frac{Q^2}{2\varepsilon A} = \frac{V^2 A\varepsilon}{2g^2} = k \cdot z \Rightarrow z = \frac{V^2 A\varepsilon}{2kg^2} \tag{B.17}$$

We see that z is a function of the gap size. This complicates the final expression

$$g = g_0 - z = g_0 - \frac{V^2 A\varepsilon}{2k(g_0 - z)^2} \tag{B.18}$$

To proceed, we solve this equation with respect to the voltage

$$V = \sqrt{\frac{z \cdot 2k}{A\varepsilon}}(g_0 - z) \tag{B.19}$$

This expression is plotted in Fig. B.7.

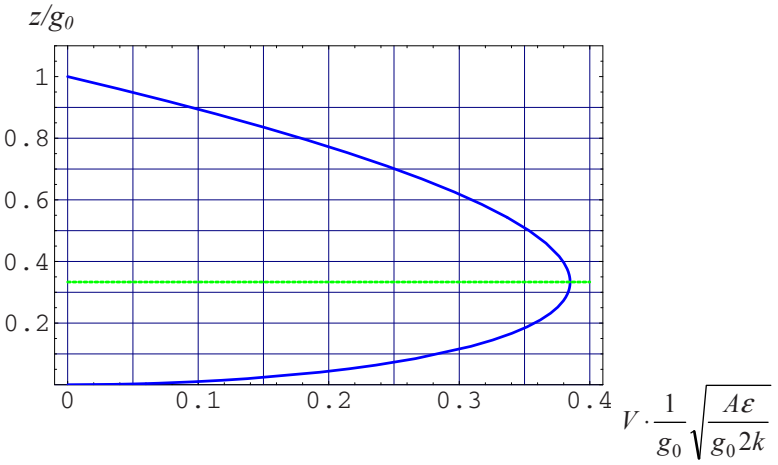


Figure B.7. *Graph showing normalized deflection, z/g_0, as a function of normalized voltage in an electrostatic, parallel-plate actuator. There are two equilibrium deflections for each value of the voltage. The solutions corresponding to the upper branch of the graph are unstable.*

We see that there are two eqilibria for each voltage. The upper branch of solutions is, however, unstable. To see that, we write an expression for the net force

$$F_{net} = -\frac{V^2 A\varepsilon}{2g^2} + k(g_0 - g) \tag{B.20}$$

and differentiate with respect to g

$$\delta F_{net} = \left.\frac{\partial F_{net}}{\partial g}\right|_V \delta g = \left(\frac{V^2 A\varepsilon}{g^3} - k\right)\delta g \tag{B.21}$$

Stability requires

$$\delta F_{net} < 0 \Rightarrow k > \frac{V^2 A\varepsilon}{g^3} \tag{B.22}$$

The edge of the stable region is defined by

$$\delta k = \frac{V^2 A\varepsilon}{g^3} \Rightarrow g = g_0 - \frac{V^2 A\varepsilon}{2kg^2} = g_0 - \frac{g}{2} \Rightarrow g = \frac{2}{3}g_0 \tag{B.23}$$

This corresponds to the maximum voltage, as can be verified by differentiation of the expression for voltage vs. deflection.

We conclude that the upper branch of the solution shown in Fig. B.7 is unstable. A real parallel-plate capacitor will therefore exhibit the snap-down characteristics shown in Fig. B.8. As the voltage is increased beyond its maximum stable value, the plates spontaneously snap together. In practice the plates will often reach a mechanical stop before they touch (which will short-circuit the voltage and lead to all kinds of unpleasant effects). In that case the capacitor has hysteresis as shown.

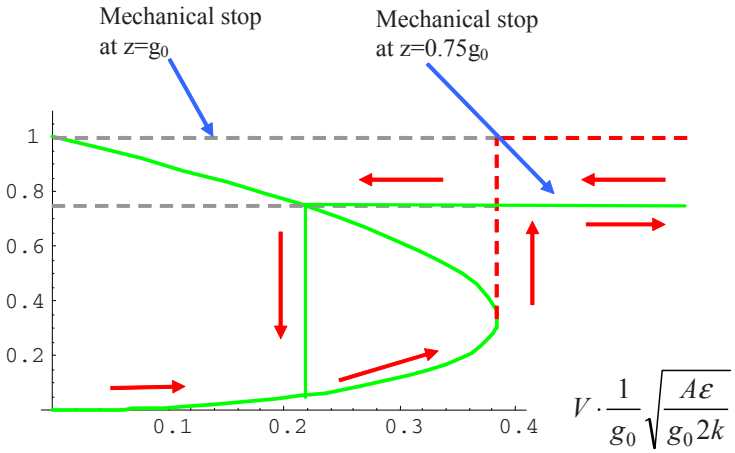


Figure B.8. Illustration of normalized deflection, z/g_0, as a function of normalized voltage in an electrostatic, parallel-plate actuator.

As the voltage applied to the parallel-plate electrostatic actuator increases, so does the deflection until the transition point between the stable and unstable regions is reached. Increasing the voltage beyond this point leads to "snap-down", or "pull-

in", i.e. the moving plate of the capacitor is accelerated until it becomes stabilized by another mechanical force.

Two cases are shown in Fig. B.8. In the first case (dashed line), the moving plate isn't stopped until it hits the lower plate. In this case, no voltage is required to hold the plate in the "snap-down" position. In the second case (solid line), the plate is stopped once it reaches a point corresponding to 75% of the original gap. Increasing the voltage further doesn't change the position of the plate. Reducing the voltage below the voltage of the unstable solution at 75% deflection makes the plate relax down to the stable branch. The result is a very open hysteresis curve.

The maximum voltage for stable operation (snap-down voltage or pull-in voltage) is given by:

$$V = \sqrt{\frac{g_0 \cdot 2k}{3A\varepsilon}}\left(g_0 - \frac{g_0}{3}\right) = \sqrt{\frac{8g_0^3 \cdot k}{27A\varepsilon}} \tag{B.24}$$

The snap-down voltage is equal to the maximum voltage for charge control. Using this expression to normalize the voltage, we find the following expression for the net force

$$\begin{aligned} F_{net} &= -\frac{V^2 A\varepsilon}{2g^2} + k(g_0 - g) = -\frac{V^2 A\varepsilon}{2\left(\frac{g}{g_0}\right)^2 g_0^2} + g_0 k\left(1 - \frac{g}{g_0}\right) \Rightarrow \\ \frac{F_{net}}{g_0 k} &= -\frac{4}{27}\frac{1}{(1-\zeta)^2}\frac{V^2}{V_{pi}^2} + \zeta \end{aligned} \tag{B.25}$$

where $\zeta = 1 - g/g_0$. The two parts of the expression for the net force is plotted in Fig. B.9.

The implication of the snap-down phenomenon is that we only can stably operate the voltage-controlled, parallel-plate, electrostatic actuator over one third of its full range of motion. The maximum force is the same as for the charge-controlled actuator, so the result is that the *force*range product*, which is an often-used figure-of-merit for microactuators, is reduced by a factor of three.

This is a substantial reduction, but the difficulties of implementing charge control for small capacitances, has made the voltage controlled actuator the more common design in practical and commercial applications. Consequently, MEMS designers have shown considerable ingenuity in coming up with actuators that extend the travel range of the simple parallel-plate actuator. We will study some of these solutions in the following.

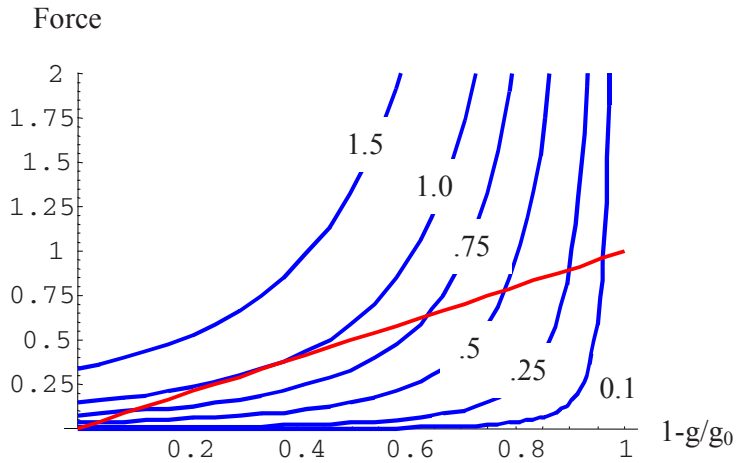


Figure B.9. *Spring force (straight line) and electrostatic force (family of curves with the applied voltage normalized to the pull-in voltage as the parameter) acting on the plates of the parallel-plate capacitor. We see that when the voltage is larger than the snap-in voltage, there are no equilibrium solutions. When the voltage equals pull-in, there is one unstable solution, and when the voltage is less than pull-in, there are two solutions, one stable and one unstable.*

B.3 Energy Conservation in the Parallel Plate Electrostatic Actuator

In the preceding sections we found the force and deflection of the parallel-plate electrostatic actuator by considering the forces set up by the electrostatic field. This straightforward approach works for this simple case, but for more complex actuators, energy methods are simpler to apply. We will develop such methods and use them to verify our calculations for the force and deflection in the parallel-plate actuator in this chapter. In the next chapter we will use these methods to investigate the characteristics of the electrostatic combdrive.

We saw earlier that if the electrodes of the parallel-plate electrostatic actuator are fixed, then the stored energy in the capacitor is given by

$$W(Q) = \int_0^Q V dQ = \int_0^Q \frac{Q}{C} dQ = \frac{Q^2}{2C} = \frac{Q^2 g}{2A\varepsilon} \tag{B.26}$$

Alternatively, we can find the same expression by considering the force attracting the capacitor plates to each other. The magnitude of the force on each plate is

$$W(g) = \int_0^g F dg = F \cdot g = \frac{Q^2 g}{2A\varepsilon} \tag{B.27}$$

The stored energy in the energy parallel-plate electrostatic actuator can be supplied either as electrical energy or mechanical energy, and the stored energy, *W(Q,g),* is a function both of the stored charge, *Q*, and the electrode gap, *g*. The differential of the stored energy can then be expressed:

$$dW(Q,g) = F \cdot dg + V \cdot dQ \tag{B.28}$$

Consequently, we can write the following expressions for the force and the voltage:

$$F = \left.\frac{\partial W(Q,g)}{\partial g}\right|_Q \tag{B.29}$$

$$V = \left.\frac{\partial W(Q,g)}{\partial Q}\right|_g \tag{B.30}$$

Using the formula we found for the stored energy, these expressions evaluate to:

$$F = \left.\frac{\partial W(Q,g)}{\partial g}\right|_Q = \frac{\partial}{\partial g}\left(\frac{Q^2 g}{2\varepsilon A}\right)\Bigg|_Q = \frac{Q^2}{2\varepsilon A} \tag{B.31}$$

$$V = \left.\frac{\partial W(Q,g)}{\partial Q}\right|_g = \frac{\partial}{\partial Q}\left(\frac{Q^2 g}{2\varepsilon A}\right)\Bigg|_Q = \frac{Qg}{\varepsilon A} \tag{B.32}$$

These expressions are valid for the charge controlled electrostatic parallel-plate actuator, and, as we would expect, we see that the results are the same as those we found by more direct methods earlier.

For the voltage controlled parallel-plate actuator we cannot use the differential above, because in this device, the voltage, not the charge, is the independent variable. In this case use the co-energy, which is a function of the voltage and the electrode gap. It is defined as:

$$W^*(V,g) = QV - W(Q,g) \tag{B.33}$$

This definition is illustrated in Fig. B.10. For a linear capacitor, the energy and the co-energy are the same, but in general these two quantities can be different.

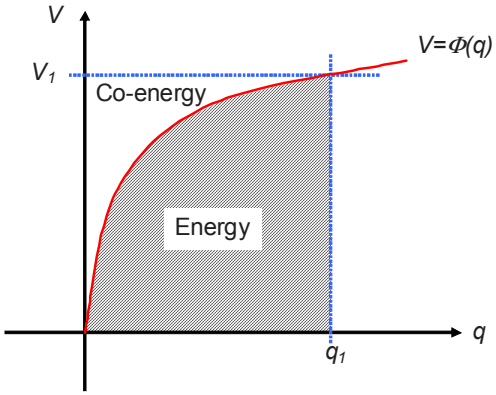


Figure B.10. Energy and co-energy of a non-linear capacitor. As the capacitor is charged, the capacitance is increased. The voltage is therefore a nonlinear function of the charge, and the energy and co-energy are different.

The differential of the co-energy is:

$$\begin{aligned} dW^*(V,g) &= Q\cdot dV + V\cdot dQ - dW(Q,g) \\ dW^*(V,g) &= Q\cdot dV + V\cdot dQ - F\cdot dg - V\cdot dQ \\ dW^*(V,g) &= Q\cdot dV - F\cdot dg \end{aligned} \tag{B.34}$$

We can now write the following expressions for the force and the charge

$$F = -\left.\frac{\partial W^*(V,g)}{\partial g}\right|_V \tag{B.35}$$

$$Q = \left.\frac{\partial W^*(V,g)}{\partial V}\right|_g \tag{B.36}$$

The equation for the force will be rewritten in terms of the capacitance for future use

$$F = -\left.\frac{\partial W^*(V,g)}{\partial g}\right|_V = -\left.\frac{\partial}{\partial g}\left(\frac{CV^2}{2}\right)\right|_V = -\left.\frac{\partial C}{\partial g}\right|_V \cdot \frac{V^2}{2} \tag{B.37}$$

The co-energy can be found by integration of the charge for a fixed gap

$$W^*(V,g) = \int_0^V Q \cdot dV = \int_0^V CV \cdot dV = \frac{1}{2}CV^2 = \frac{\varepsilon A V^2}{2g} \tag{B.38}$$

We can now evaluate the expressions for the force and charge:

$$F = -\frac{\partial}{\partial g}\left(\frac{\varepsilon A V^2}{2g}\right)\bigg|_V = \frac{\varepsilon A V^2}{2g^2} = \frac{Q^2}{2\varepsilon A} \tag{B.39}$$

$$Q = \frac{\partial}{\partial V}\left(\frac{\varepsilon A V^2}{2g}\right)\bigg|_g = \frac{\varepsilon A V}{g} = CV \tag{B.40}$$

We see that these expressions agree with the basic definitions and the formulas we found earlier by more direct methods.

To see the use of the energy and co-energy, we now plot these quantities for the parallel-plate actuator. First we plot the stored energy, including both electrical and mechanical energy, in a charge controlled parallel-plate actuator as a function of electrode spacing (gap) with the constant charge as a parameter. This is done in Fig. B.11, showing that there is one stable solution for each value of the charge, Q.

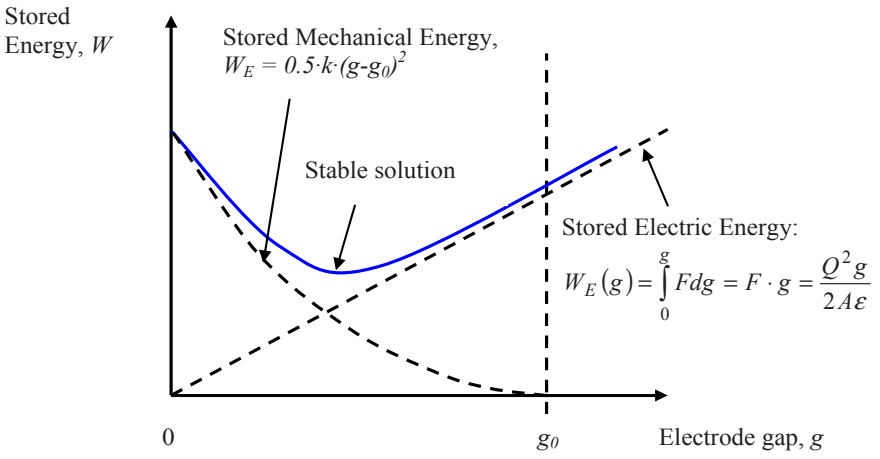


Figure B11. *The stored mechanical and electrical energy in a charge controlled parallel-plate actuator as a function of electrode spacing (gap) with the constant charge as the parameter. The minimum represents a stable solution. This figure is for illustration only. It is not to scale and should not be used for calculations.*

For the voltage controlled actuator, the situation is more complex. This is illustrated in Fig. B.12 where we have plotted the co-energy in a voltage controlled parallel-plate actuator as a function of electrode spacing (gap) with the constant voltage as a parameter. The co-energy is defined as:

$$W^* = QV - W = QV - \frac{1}{2}CV^2 - \frac{1}{2}k(g-g_0)^2 = CV^2 - \frac{1}{2}k(g-g_0)^2$$
$$W^* = \frac{\varepsilon A}{g}V^2 - \frac{1}{2}k(g-g_0)^2 \tag{B.41}$$

The graphs shows two stable solutions at the two maxima of the co-energy.

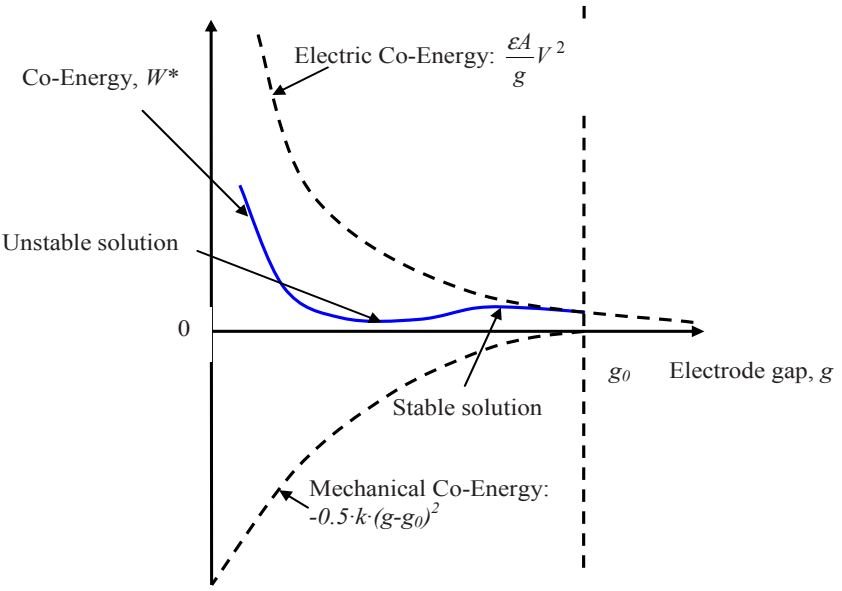


Figure B12. The co-energy in a voltage-controlled parallel-plate actuator as a function of electrode spacing (gap) with the constant voltage as the parameter. The maxima (one at g=0 and one marked on the graph) represent stable solutions. This figure is for illustration only. It is not to scale and should not be used for calculations

B.4 Electrostatic Spring

The nonlinear characteristics of the electrostatic force creates an "electrostatic spring" that leads to shifts of the natural frequency of microactuators [1], and that can be used to tune both the sense-mode frequency and the sensitivity of microsensors [2]. The basic mechanism of the electrostatic spring is illustrated in Fig. B.13.

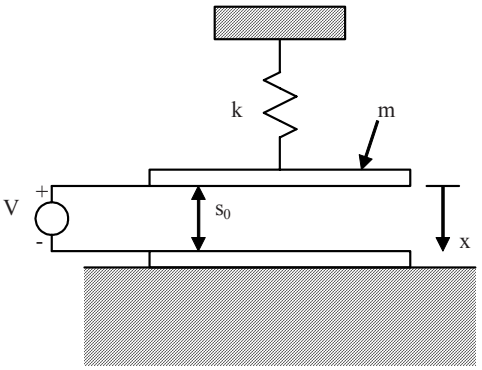


Figure B.13. Schematic drawing of parallel-plate MEMS resonator with an applied voltage that creates an electrostatic spring that modifies the spring constant, and therefore the natural frequency, of the resonator.

Neglecting fringing fields (not because they are not important, but because it is easy), and the damping, we can write the force balance for the upper plate as:

$$m\ddot{x} + kx = \frac{\varepsilon_0 V^2}{2(s_0 - x)^2} \tag{B.42}$$

where ε_0 is the dielectric constant and the other parameters are defined in Fig.B13. The nonlinear electrostatic force (the right-hand side of the force-balance equation) can be expanded in a Taylor series around a nominal displacement x_0

$$\begin{aligned} F_e &= \frac{\varepsilon_0 V^2}{2(s_0 - x)^2} \approx \frac{\varepsilon_0 V^2}{2(s_0 - x)^2}\bigg|_{x=x_0} + \frac{\varepsilon_0 V^2(-2)(-1)}{2(s_0 - x)^3}\bigg|_{x=x_0} (x - x_0) + \ldots \\ F_e &\approx \frac{\varepsilon_0 V^2}{2(s_0 - x_0)^2} + \frac{\varepsilon_0 V^2}{(s_0 - x_0)^3}(x - x_0) + \ldots = \frac{\varepsilon_0 V^2}{2(s_0 - x_0)^2}\left[1 + \frac{x - x_0}{s_0 - x_0} + \ldots\right] \end{aligned} \tag{B.43}$$

Inserting this expansion into the force balance yields

$$m\ddot{x} + \left(k - \frac{\varepsilon_0 V^2}{(s_0 - x_0)^3}\right)x = \frac{\varepsilon_0 V^2}{2(s_0 - x_0)^2}\left[1 - 2\frac{x_0}{s_0 - x_0}\right] \tag{B.44}$$

This equation is the familiar expression for a second order resonance. We see that the mechanical spring constant, and therefore the resonance frequency, is modified by the electrostatic force. The modified resonance frequency is given by

$$\omega_{res} = \left(\frac{k}{m} - \frac{\varepsilon_0 V^2}{m(s_0 - x_0)^3} \right)^{\frac{1}{2}} \tag{B.45}$$

The voltage required for a specific static deflection, x_0, is found from the force balance (without the time derivative term):

$$kx_0 = \frac{\varepsilon_0 V^2}{2(s_0 - x_0)^2} \Rightarrow \frac{\varepsilon_0 V^2}{m(s_0 - x_0)^3} = \frac{2kx_0}{m(s_0 - x_0)} \tag{B.46}$$

Inserted into the equation for the resonance frequency, this gives:

$$\omega_{res} = \left(\frac{k}{m} - \frac{2kx_0}{m(s_0 - x_0)} \right)^{\frac{1}{2}} = \sqrt{\frac{k}{m}} \sqrt{1 - \frac{2x_0}{s_0 - x_0}} \tag{B.47}$$

Solving the force-balance equation for the voltage, V, and maximizing gives us the maximum voltage, and the corresponding deflection, that the plate-spring system can support. Applied voltages larger than this maximum will lead to a spontaneous "pull-in" or "snap-down" to the substrate of the spring-supported plate. We'll call this voltage (deflection) the electrostatic instability voltage (deflection).

$$V = \sqrt{\frac{2kx_0}{\varepsilon_0}(s_0 - x_0)^2} \tag{B.48}$$

$$\frac{\partial V}{\partial x} = 0 = \frac{\frac{2k}{\varepsilon_0}\left((s_0 - x_0)^2 + x_0 2(s_0 - x_0)\right)}{2\sqrt{\frac{2kx_0}{\varepsilon_0}(s_0 - x_0)^2}} \Rightarrow x_0 = \frac{s_0}{3} \tag{B.49}$$

$$V_{snap} = \sqrt{\frac{2ks_0}{3\varepsilon_0}(s_0 - s_0/3)^2} = \sqrt{\frac{8k}{27\varepsilon_0} s_0^3} \tag{B.50}$$

$$\omega_{snap} = \sqrt{\frac{k}{m}} \sqrt{1 - \frac{2\,s_0/3}{s_0 - x_0}} = 0 \tag{B.51}$$

At the instability, the resonance frequency goes to zero.

B.4.1 Sensors Based on the Electrostatic Spring

The electrostatic-spring effect can be used to create sensors with frequency-modulated output and tunable sensitivity. Consider the schematic of a pressure sensor shown in Fig. B.14.

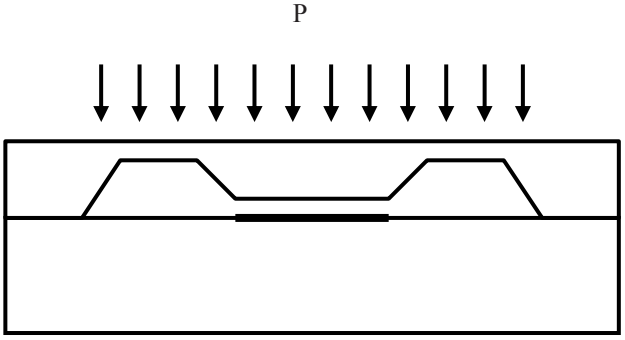


Figure B.14. Schematic drawing of pressure sensor, in which the sensitivity and the response frequency can be electrostatically tuned.

In static equilibrium we have that the mechanical spring force is equal to the pressure and the electrostatic force

$$F = \frac{\varepsilon V^2}{2(s_0 - x)^2} + P \cdot A \tag{B.52}$$

The pressure causes a static deflection

$$s = s_0 - PA/k \tag{B.53}$$

The resonance frequency then becomes

$$\omega = \sqrt{\frac{k}{m} - \frac{\varepsilon_0 V^2}{m(s_0 - PA/k - x_0)^3}} \tag{B.54}$$

The sensitivity of the sensor is defined as

$$\frac{\partial \omega}{\partial P} = \frac{-\dfrac{(-3)\varepsilon_0 V^2 (-A/k)}{m(s_0 - PA/k - x_0)^4}}{2\sqrt{\dfrac{k}{m} - \dfrac{\varepsilon_0 V^2}{m(s_0 - PA/k - x_0)^3}}} = -\frac{3A\varepsilon_0 V^2}{2mk(s_0 - PA/k - x_0)^4 \sqrt{\dfrac{k}{m} - \dfrac{\varepsilon_0 V^2}{m(s_0 - PA/k - x_0)^3}}} \tag{B.55}$$

At the instability:

$$\frac{\partial \omega}{\partial P} = -\frac{3A\varepsilon_0 V^2}{2mk(s_0 - PA/k - x_0)^4 \sqrt{\dfrac{k}{m} - \dfrac{\varepsilon_0 V^2}{m(s_0 - PA/k - x_0)^3}}} \tag{B.56}$$

$$\Rightarrow \lim_{V \to V_{snap}} \frac{\partial \omega}{\partial P} = \infty$$

We see that the electrostatic spring allows us to trade off bandwidth for sensitivity; as the voltage is increased towards the instability point, the sensitivity goes to infinity, but the resonance frequency, and therefore the bandwidth, goes to zero.

B.5 Electrostatic Combdrives

Voltage-controlled, parallel-plate, electrostatic actuators suffer from problems with snap-down and limited range of operation as discussed in the preceding chapters. A much-used electrostatic actuator that avoids these problems is the electrostatic combdrive shown in Fig. B.15.

The operation of the electrostatic combdrive is very similar to that of the parallel-plate actuator. Just like the parallel-plate actuator, the combdrive has two electrodes; one stationary, and one that is suspended by a mechanical spring so that is will move under an applied force. The force required to move the suspended electrode is created by setting up an electrostatic field between the two electrodes. This can be accomplished by controlling the charge on the electrodes, or by applying a voltage between them, as is the case for the parallel-plate actuator.

The obvious difference between the combrive and the parallel-plate actuator is in the geometry of the electrodes. The combdrive has interdigitated electrodes as shown in Fig. B.15, and that has important consequences for the characteristics of the device. As the two electrodes are pulled together, the increase in the capacitance is mostly due to the increased overlap of the teeth of the two combs. (There

is also a small contribution to the capacitance increase from the end of the teeth moving closer to the base of the opposite electrode, but this contribution is negligible in most practical designs). This capacitance increase is a linear function of the relative positions of the electrodes (this is different from the parallel-plate actuator, in which the capacitance is inversely proportional to the electrode spacing, i.e. the capacitance is a non-linear function of the relative electrode positions). The combdrive is therefore sometimes referred to as the linear electrostatic combdrive.

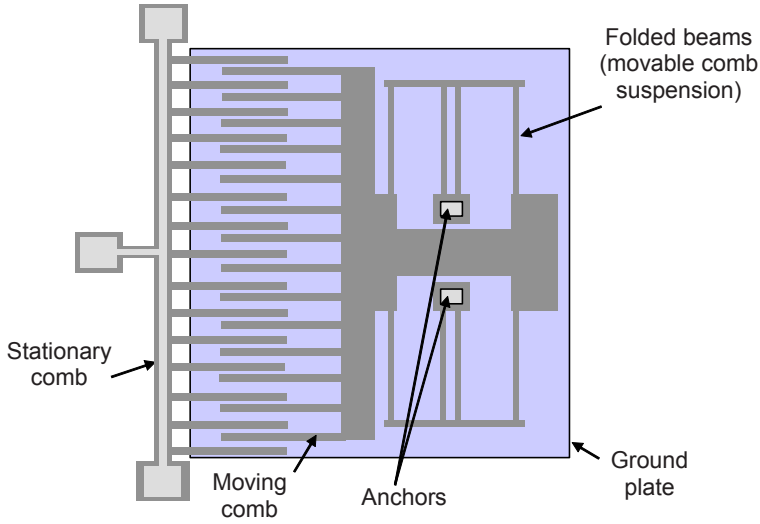


Figure B.15. Electrostatic combdrive. The voltage across the interdigitated electrodes creates a force that is balanced by the spring force in the crab-leg suspension.

To calculate the electrostatic force in the combdrive, consider the field distribution shown in Fig. B.16. We write the capacitance as a sum of two parts; one corresponding to the fringing fields, and one corresponding to the fields in the region of overlap between the electrodes:

$$C_{tot} = C_0 + C(x) \tag{B.57}$$

The force can be written (note that the coordinate x here is chosen opposite of g in the parallel-plate actuator. This changes the sign in the expression for the force):

$$F = \left.\frac{\partial W^*}{\partial x}\right|_V = \frac{\partial}{\partial x}\left(\frac{1}{2}C \cdot V^2\right)\Bigg|_V = \frac{V^2}{2} \cdot \frac{\partial C}{\partial x} \tag{B.58}$$

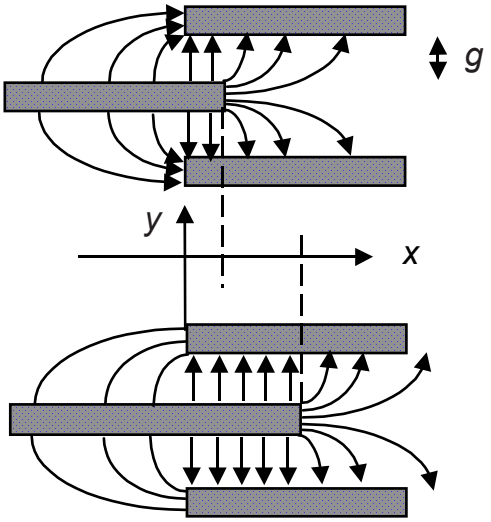


Figure B.16. Electric field distribution in comb-finger gaps. Note that the direction of the x-coordinate is chosen opposite of the parameter g in the parallel-plate actuator.

Using the same uniform-field approximation we employed for the parallel-plate actuator, we can write:

$$F = \frac{1}{2} \cdot V^2 \cdot \frac{\partial C}{\partial x} = \frac{1}{2} \cdot V^2 \cdot \frac{2N \cdot \varepsilon \cdot h}{g} = V^2 \cdot \frac{N \cdot \varepsilon \cdot h}{g} \tag{B.59}$$

where N is number of comb-fingers, h is the thickness of the comb-fingers (perpendicular to the plane in Fig. B.16), and g is the width of gap between the comb-fingers.

In many practical implementations of the electrostatic combdrive, the thickness, h, of the combteeth is comparable to the electrode gap, g. Under those conditions, the parallel-plate approximation is relatively inaccurate. The most accurate representation of the fringing field are obtained by numerical techniques, but for many purposes it is sufficient to use tabulated correction factors to compensate for the effects of the finite thickness. The force can then be expressed as:

$$F = V^2 \cdot N \cdot \left(\frac{\alpha \cdot \varepsilon \cdot h}{g^{\beta}} \right) \cdot \eta \tag{B.60}$$

where α, β, η are fitting parameters extracted from simulations [3].

We see that the force in the ***voltage-controlled*** combdrive is not a function of the displacement. This is what we found for the ***charge-controlled*** parallel-late actuator. The combdrive does therefore not suffer from snap-down in the primary de-

flection direction (x in Fig. B.16), even in the voltage-controlled case. This is one of the major reasons for the popularity of combdrives in MEMS technology.

The voltage-controlled combdrive is, however, susceptible to snap-down in the transversal direction. Figure B.16 shows clearly that each of the teeth in the movable comb is attracted sideways towards its nearest neighbors on either side. In the ideal case, the gaps on both sides are equal, so that the sideways forces exactly balance. In reality, however, the two gaps will not be exactly equal, and there will be a net sideways force in one direction or the other. It is also important to be aware that even in the ideal case, the combdrive will be unstable if the voltage and the overlap between the combteeth are too large. This happens when the voltage creates sideways forces that are so big that an infinitesimal offset from the perfectly centered position will make the comb snap sideways.

To analyze the stability of the combdrive, we need an expression for the total potential energy and co-energy. We start by generalizing the expression for the capacitance of the combdrive to the situation where the movable teeth are asymmetrically placed between the stationary teeth [4]

$$C = N \cdot \varepsilon \cdot h \cdot x \left(\frac{1}{g-y} + \frac{1}{g+y} \right) \tag{B.61}$$

The force can then be expressed

$$F_x = \frac{1}{2} \cdot V^2 \cdot \frac{\partial C}{\partial x} = \frac{1}{2} \cdot V^2 \cdot N \cdot \varepsilon \cdot h \left(\frac{1}{g-y} + \frac{1}{g+y} \right) = k_x x \tag{B.62}$$

In equilibrium, the electrostatic force must equal the mechanical spring force

$$\begin{aligned} &\frac{1}{2} \cdot V^2 \cdot N \cdot \varepsilon \cdot h \left(\frac{1}{g-y} + \frac{1}{g+y} \right) = k_x x \Rightarrow \\ &V = \sqrt{\frac{2k_x x}{N \cdot \varepsilon \cdot h \left(\frac{1}{g-y} + \frac{1}{g+y} \right)}} \end{aligned} \tag{B.63}$$

We can write a similar expression for the force balance in the transversal direction

$$F_y = \frac{1}{2} \cdot V^2 \cdot \frac{\partial C}{\partial y} = \frac{1}{2} \cdot V^2 \cdot N \cdot \varepsilon \cdot h \cdot x \left(\frac{1}{(g-y)^2} - \frac{1}{(g+y)^2} \right) = k_y y \tag{B.64}$$

Finally, we can write an equation for the total co-energy in the actuator

$$W^* = QV - W = QV - \frac{1}{2}CV^2 - \frac{1}{2}k_x x^2 - \frac{1}{2}k_y y^2 = \\ = \frac{1}{2}CV^2 - \frac{1}{2}k_x x^2 - \frac{1}{2}k_y y^2 \Rightarrow \tag{B.65}$$

$$W^* = \frac{1}{2}N \cdot \varepsilon \cdot h \cdot x\left(\frac{1}{g-y} + \frac{1}{g+y}\right) \cdot V^2 - \frac{1}{2}k_x x^2 - \frac{1}{2}k_y y^2 \tag{B.66}$$

The values of y where $\partial W^*/\partial y = 0$ are all possible equilibria, but only those that have $\partial^2 W^*/\partial y^2 < 0$ are stable

$$\frac{\partial^2 W^*}{\partial y^2} = N \cdot \varepsilon \cdot h \cdot x\left(\frac{1}{(g-y)^3} + \frac{1}{(g+y)^3}\right) \cdot V^2 - k_y \tag{B.67}$$

$$\frac{\partial^2 W^*}{\partial y^2} = \left(\frac{1}{(g-y)^3} + \frac{1}{(g+y)^3}\right) \cdot \frac{2k_x x^2}{\left(\frac{1}{g-y} + \frac{1}{g+y}\right)} - k_y \tag{B.68}$$

$$\frac{\partial^2 W^*}{\partial y^2} = 2k_x x^2\left(\frac{(g+y)^3 + (g-y)^3}{(g-y)^3(g+y)^3}\right) \cdot \frac{g^2 - y^2}{2g} - k_y \tag{B.69}$$

$$\frac{\partial^2 W^*}{\partial y^2} = 2k_x x^2 \cdot \frac{g^2 + 3y^2}{\left(g^2 - y^2\right)^2} - k_y \tag{B.70}$$

For $y=0$, this expression simplifies to:

$$y = 0 \Rightarrow \frac{\partial^2 W^*}{\partial y^2} = \frac{2k_x x^2}{g^2} - k_y \tag{B.71}$$

In the ideal case ($y=0$) we therefore have the stability criterion:

$$k_y - \frac{2k_x x^2}{g^2} > 0 \Rightarrow \frac{x}{g} < \sqrt{\frac{k_y}{2k_x}} \tag{B.72}$$

We see that sideways snap-down limits the deflection of the electrostatic comb-drive. We have to make the ratio of the transversal (y-direction) spring constant to the longitudinal (x-direction) spring constant as large as possible.

It should be noted here that the sideways snap-down that we have focused on in this treatment is only one of several possible electrostatic instabilities that must be considered in the design of combdrives. Others include rotational snap-down, and snap-down to the substrate. Creating springs with large ratios of their spring constants for unwanted vs. wanted motion is therefore an important issue in mechanical MEMS design.

We can now compare the force in the parallel-plate and the combdrive actuators. The first-order expressions for the force in these two actuators are:

$$\text{Combdrive: } F_{cd} = \frac{N \cdot \varepsilon \cdot h \cdot V^2}{d} \tag{B.73}$$

$$\text{Parallel plate: } F_{pp} = \frac{A \cdot \varepsilon \cdot V^2}{2g^2} \tag{B.74}$$

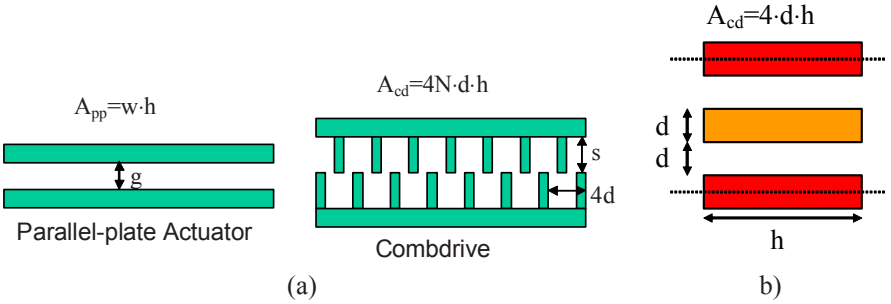


Figure B.17. Calculation of area of the combdrive. (a) shows both the parallel-plate actuator and the combdrive, and (b) shows one unit cell of a periodic combdrive.

To compare these two expressions, we write the area of the combdrive as

$$A_{cd} = 4N \cdot d \cdot h \tag{B.75}$$

where the parameters are defined in Fig. B.17. Given these definitions, we find the ratio of the force produced by a combdrive and a parallel-plate actuator of the same cross-sectional area to be

$$\frac{F_{cd}}{F_{pp}} = \frac{g^2}{2d^2} \tag{B.76}$$

We see that the combdrive can generate substantially larger forces than the parallel-plate actuator. If we make the assumptions that the parallel-plate actuator is

voltage controlled and can be operated over one third of its gap (*g/3*), and that the gap in the combdrive is determined by the lithographic resolution, we find

$$\frac{F_{cd}}{F_{pp}} = \frac{9 \cdot (range)^2}{2 \cdot (linewidth)^2} \tag{B.77}$$

In many applications we might want the total range of travel of the actuators to be one or two orders of magnitude larger than the lithographic linewidth. In these applications, the combdrive is clearly vastly superior to the parallel-plate actuator, at least if the maximum available force is important.

We should remember, however, that the range of the combdrive is also limited by snap-down as discussed above. Using the expression we found for the deflection in the combdrive $\left(x_{\max} = d \cdot \sqrt{\frac{k_y}{2k_x}} \right)$, we can rewrite the force ratio as

$$\frac{F_{cd}}{F_{pp}} = \frac{9}{2} \cdot \frac{d^2 \cdot \frac{k_y}{2k_x}}{d^2} = \frac{9}{4} \frac{k_y}{k_x} \tag{B.78}$$

Note that this equation is valid for the situation where the parallel-plate actuator and the combdrive have the same range of motion, given by the maximum range of motion possible in the combdrive. This is not necessarily the most "fair" way to compare these two actuators. We can often use leverage to trade off force and range such that their product is constant.

In many applications we therefore find that the *force*range* product is a better figure-of-merit than the force:

$$\frac{F_{cd} \cdot range_{cd}}{F_{pp} \cdot range_{pp}} = \frac{g^2}{2d^2} \cdot \frac{\sqrt{\frac{k_y}{2k_x}} \cdot d}{\frac{g}{3}} = \frac{3}{2} \cdot \frac{g}{d} \cdot \sqrt{\frac{k_y}{2k_x}} \approx \frac{3}{2} \cdot \sqrt{\frac{k_y}{2k_x}} \tag{B.79}$$

We see that this equation is not as favorable for the combdrive as the one derived earlier. If the mechanical springs are well designed ($k_y >> k_x$), however, the combdrive is still superior to the parallel-plate actuator by a substantial margin.

B.6 Summary of Electrostatic Actuation

Under the assumption of uniform field between the plates and zero field outside, we find the following expressions for the electric field, voltage, capacitance, force

(note the factor of 2 in the denominator), and stored energy in parallel-plate capacitors:

$$E = \frac{Q}{\varepsilon A} \tag{B.80}$$

$$V = E \cdot g = \frac{g \cdot Q}{\varepsilon A} \tag{B.81}$$

$$C = \frac{Q}{V} = \frac{\varepsilon \cdot A}{g} \tag{B.82}$$

$$F = \frac{QE}{2} = \frac{Q^2}{2\varepsilon A} \tag{B.83}$$

$$W(Q) = \frac{Q^2 g}{2A\varepsilon} \tag{B.84}$$

Here A is the area of each capacitor plate, g is the spacing of the plates, ε is the dielectric constant of the material (air) between the plates, and Q is the magnitude of the charge on each plate.

If we control the charge on the capacitor, we can express the force, deflection, gap, and voltage as follows:

$$F = \frac{QE}{2} = \frac{Q^2}{2\varepsilon A} \tag{B.85}$$

$$z = \frac{Q^2}{2k\varepsilon A} \tag{B.86}$$

$$g = g_0 - z = g_0 - \frac{Q^2}{2k\varepsilon A} \tag{B.87}$$

$$V = E \cdot g = \frac{Q}{\varepsilon A} \cdot g = \frac{Q}{\varepsilon A}\left(g_0 - \frac{Q^2}{2kA\varepsilon}\right) \tag{B.88}$$

The deflection is a monothonically increasing function of the charge, increasing from zero to the full gap as the charge increases from zero to $\hat{Q} = \sqrt{g_0 \cdot 2k\varepsilon A}$.

The charge controlled parallel plate actuator is stable over the whole electrode gap, but it is hard to implement because typical MEMS capacitors are very small ($\sim 10^{-15}$ *F*). In practice we therefore more often use voltage control. In this case the charge, force, and displacement of the capacitor are:

$$Q = V \cdot C = \frac{VA\varepsilon}{g} \tag{B.89}$$

$$F = \frac{Q^2}{2\varepsilon A} = \frac{V^2 A\varepsilon}{2g^2} = k \cdot z \Rightarrow z = \frac{V^2 A\varepsilon}{2kg^2} \tag{B.90}$$

The displacement is a function of the gap size:

$$g = g_0 - z = g_0 - \frac{V^2 A\varepsilon}{2k(g_0 - z)^2} \Rightarrow \tag{B.91}$$

$$V = \sqrt{\frac{z \cdot 2k}{A\varepsilon}}(g_0 - z) \tag{B.92}$$

This cubic equation in z has two solutions for $z<g_0$, but only voltages less than $V_{snap-down} = \sqrt{\frac{8g_0^3 \cdot k}{27A\varepsilon}}$ lead to stable solutions.

The differential of the stored energy can be expressed:

$$dW(Q,g) = F \cdot dg + V \cdot dQ \tag{B.93}$$

which leads to the following expressions for the force and voltage:

$$F = \left.\frac{\partial W(Q,g)}{\partial g}\right|_Q = \frac{\partial}{\partial g}\left(\frac{Q^2 g}{2\varepsilon A}\right)\Bigg|_Q = \frac{Q^2}{2\varepsilon A} \tag{B.94}$$

$$V = \left.\frac{\partial W(Q,g)}{\partial Q}\right|_g = \frac{\partial}{\partial Q}\left(\frac{Q^2 g}{2\varepsilon A}\right)\Bigg|_Q = \frac{Qg}{\varepsilon A} \tag{B.95}$$

The co-energy is defined as:

$$W^*(V,g) = QV - W(Q,g) \Rightarrow \tag{B.96}$$

$$dW^*(V,g) = Q \cdot dV + V \cdot dQ - dW(Q,g) = Q \cdot dV - F \cdot dg \Rightarrow \tag{B.97}$$

=>

$$F = -\left.\frac{\partial W^*(V,g)}{\partial g}\right|_V \tag{B.98}$$

$$Q = \left.\frac{\partial W^*(V,g)}{\partial V}\right|_g \tag{B.99}$$

The co-energy can be found by integration of the charge for a fixed gap:

$$W^*(V,g) = \int_0^V Q \cdot dV = \int_0^V CV \cdot dV = \frac{1}{2}CV^2 = \frac{\varepsilon A V^2}{2g} \tag{B.100}$$

which leads to the following expressions for the force and charge:

$$F = -\left.\frac{\partial}{\partial g}\left(\frac{\varepsilon A V^2}{2g}\right)\right|_V = \frac{\varepsilon A V^2}{2g^2} = \frac{Q^2}{2\varepsilon A} \tag{B.101}$$

$$Q = \left.\frac{\partial}{\partial V}\left(\frac{\varepsilon A V^2}{2g}\right)\right|_g = \frac{\varepsilon A V}{g} = CV \tag{B.102}$$

The force in the ***electrostatic combdrive***, can be found from the equation:

$$F = \left.\frac{\partial W^*}{\partial x}\right|_V = \left.\frac{\partial}{\partial x}\left(\frac{1}{2}C \cdot V^2\right)\right|_V = \frac{V^2}{2} \cdot \frac{\partial C}{\partial x} \tag{B.103}$$

Note that the coordinate x here is chosen opposite of g in the parallel-plate actuator. Using the uniform-field approximation, we can write:

$$F = \frac{1}{2} \cdot V^2 \cdot \frac{\partial C}{\partial x} = \frac{1}{2} \cdot V^2 \cdot \frac{2N \cdot \varepsilon \cdot h}{g} = V^2 \cdot \frac{N \cdot \varepsilon \cdot h}{g} \tag{B.104}$$

where N is number of comb-fingers, h is the thickness of the comb-fingers (perpendicular to the plane in Fig. B.15), and g is the width of gap between the comb-fingers.

The force in the ***voltage-controlled*** combdrive is not a function of the displacement, which means that the combdrive is ***stable in the longitudinal direction***. Voltage control does make the combdrive unstable in the transversal direction. This limits the range of motion even in the perfectly aligned combdrive

$$\frac{x}{g} < \sqrt{\frac{k_y}{2k_x}} \tag{B.105}$$

The ratio of the forces of combdrives and parallel-plate actuators is

$$\frac{F_{cd}}{F_{pp}} = \frac{g^2}{2d^2} \tag{B.106}$$

If the range is the same for both actuators, this can be written

$$\frac{F_{cd}}{F_{pp}} = \frac{9 \cdot (range)^2}{2 \cdot (linewidth)^2} = \frac{9}{2} \cdot \frac{d^2 \cdot \frac{k_y}{2k_x}}{d^2} = \frac{9}{4}\frac{k_y}{k_x} \tag{B.107}$$

The ratio of the *force*range* products is:

$$\frac{F_{cd} \cdot range_{cd}}{F_{pp} \cdot range_{pp}} = \frac{g^2}{2d^2} \cdot \frac{\sqrt{\frac{k_y}{2k_x}} \cdot g}{\frac{g}{3}} = \frac{3}{2} \cdot \frac{g^2}{d^2} \cdot \sqrt{\frac{k_y}{2k_x}} \approx \frac{3}{2} \cdot \sqrt{\frac{k_y}{2k_x}} \tag{B.108}$$

Conclusion: The combdrive is superior to the parallel-plate actuator, particularly for long-range operation.

References

1 Y.He, J. Marchetti, C. Gallegos, F. Maseeh, "Accurate fully-coupled natural frequency shift of mems actuators due to voltage bias and other external forces", Proceedings of MEMS 99.

2 W.A. Clark, R.T. Howe, R. Horowitz, "Surface micromachined z-axis vibratory rate gyroscope", Proceedings of the Solid-Sate Sensor and Actuator Workshop, pp. 283-287, Hilton Head, North Carolina, June 1996.

3 W.C-K. Tang, "Electrostatic Comb Drive for Resonant Sensor and Actuator Applications", PhD thesis, Department of Electrical Engineering and Computer Science, University of California, Berkeley, 1990.

4 J.D. Grade, "Large-Deflection, High-Speed, Electrostatic Actuators for Optical Switching Applications", PhD thesis, Department of Mechanical Engineering, Stanford University, September 1999.

Index

E

F

G

H

I

K

L

M

N

O

P

R

S

T

U

V

W

Y