Optical Antenna-Enhanced Light-Emitting Diodes and Inverse Electromagnetic Design

Sean Hooten

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2022-18
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2022/EECS-2022-18.html

May 1, 2022

Optical Antenna-Enhanced Light-Emitting Diodes and Inverse Electromagnetic Design

by

Sean Hooten

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering – Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Eli Yablonovitch, Chair
Professor Ming C. Wu
Professor Feng Wang

Spring 2021

Optical Antenna-Enhanced Light-Emitting Diodes and Inverse Electromagnetic Design

Abstract

Optical Antenna-Enhanced Light-Emitting Diodes and Inverse Electromagnetic Design

by

Sean Hooten

Doctor of Philosophy in Engineering – Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Eli Yablonovitch, Chair

After its invention in 1960, the laser caused a paradigm shift in long-haul telecommunications, delivering unparalleled communications bandwidth at high power. However, next-generation integrated on-chip optical data communications require different specifications, favoring efficient nanoscale transmitters operating at low-power and high direct-electrical modulation speed. This dissertation argues that a novel device called the optical antenna-enhanced light-emitting diode (antenna-LED) can meet these requirements, as an alternative to conventional semiconductor lasers.

A detailed comparison between LEDs and lasers for on-chip optical links is provided in the first part of the dissertation. We demonstrate novel methods to quantify the stimulated emission carrier lifetime (τ_{st}) and spontaneous emission carrier lifetime (τ_{sp}) in lasers and LEDs respectively, ultimately finding $\tau_{st} = 6$ps for a laser at saturation and $\tau_{sp} = 1$ns for a heavily-doped LED. While exploring the limits of τ_{sp}, we reject the standard BNP model of the LED recombination rate. We go on to show that optical antennas can enhance the rate of spontaneous emission from LEDs by several orders of magnitude. The resulting antenna-LED can reach the needed carrier lifetime due to (enhanced) spontaneous emission of 6ps, rivaling the corresponding lifetime in lasers. This allows us to quantify the direct-electrical modulation rate of antenna-LEDs versus lasers. In doing so, we reject the standard small-signal modulation approximation in favor of the large-signal digital modulation that would be required in low-power on-chip interconnects. We find that antenna-LEDs and lasers are both limited by their respective carrier lifetimes in this modulation format – indicating that antenna-LEDs can be as fast as lasers. Finally, we show that antenna-LEDs are capable of achieving practical internal quantum efficiency, provided that surface treatment processes for III-V semiconductors are improved. Putting it all together, our analysis demonstrates that an antenna-LED with 6ps carrier lifetime, 10^4cm/s surface recombination velocity, and 50% overall efficiency can reach a direct-electrical modulation speed exceeding 50Gbit/s while emitting 500 photons/bit. We argue that this is sufficient signal for next-generation receivers.

Tangentially, we demonstrate novel ways that the antenna-LED radiation efficiency and waveguide coupling efficiency can be maximized. For example, by utilizing dielectric nanostructures in the antenna gap, a known tradeoff between antenna enhancement and antenna efficiency can be overcome. Furthermore,

we present the design and simulation of an optical antenna-LED with 94% coupling efficiency to a single-mode waveguide, potentially enabling efficient integrated optical interconnects.

In the final part of the dissertation, we discuss inverse electromagnetic design – computational tools and methods that can be used to efficiently optimize electromagnetic devices consisting of arbitrary numbers of geometric parameters. We provide two inverse design tutorials: (1) inverse design via the adjoint method for electromagnetic devices that satisfy Maxwell's equations, and (2) a novel semi-analytical transfer-matrix method for the design of 1D interference filters. We then apply these techniques to conventionally difficult electromagnetic design problems. Using (1) we design 65nm-CMOS-compatible perfectly vertical grating couplers with an industry competitive simulated insertion loss of -0.52dB. Using (2) we demonstrate distributed Bragg reflectors (DBRs) with $>$99% reflectivity over an extremely broad spectrum. We conclude with a brief look at emerging inverse design methods that are aided by neural networks.

To my parents

Contents

List of Figures

List of Tables

Acknowledgments

I have had the privilege of working with a number of passionate individuals over the course of my Ph.D whose influences have shaped the researcher that I am today.

First and foremost, I would like to thank my advisor Professor Eli Yablonovitch. His uncanny ability to peer through the web of intricacies in search of fundamental science has taught me the invaluable lesson of seeking clarity and comprehension in my work. I will always be grateful for the confidence and independence he has instilled in me as an academic and a person.

Additionally, I thank Professor Ming C. Wu, who has mentored me throughout the majority of my Ph.D. My interactions with him were integral to much of the research presented in this thesis, and I greatly appreciate his experience and insight.

I thank each of the members of the Yablonovitch, Wu, and Kante cohorts who I have collaborated with and become genuine friends with over the years: namely Zunaid Omair and Sri Krishna Vadlamani for insightful discussions of physics, optimization, and any given controversy of the week; Nicolas Andrade for many scientific collaborations and endless deliberations of LED physics, often by text; Andrew Michaels for teaching me inverse design and the adjoint method; Patrick Xiao, Greggory Scranton, and Luis Pazos-Outon for entertaining lunch discussions, scientific or otherwise (usually otherwise); and finally, Seth Fortuna and Kevin Han for collaboration on the antenna-LED project.

Furthermore, I thank Dr. Thomas Van Vaerenbergh and Dr. Peng Sun who advised me during my internship at HPE Labs, and who I will continue to interact with in the future. Additionally I thank Dr. Richard Schatz who graciously provided his laser simulation software to me.

They say friends are the family you choose, but I have never had to question the lifelong friendships that continue to enrich my life. So much has happened over the last 5 years, and I would not have stayed sane without your support: *you know who you are*.

Finally, my academic journey and personal development would never have been possible without my family. Most importantly, I thank my parents Stacy and Grant Clemens who have offered nothing but unwavering love and support; and I thank my father, whose memory lives on.

Chapter 1

Introduction

Whether by fire, telegraph, radio, or optical fiber, there are few technologies that are as historically pervasive and disruptive as telecommunications by electromagnetic waves. For example, after the invention of the laser in the 1960s, optical fiber became the dominant telecommunications medium to foster the scale and ubiquity of the Internet – a technology that is now integral in our day-to-day lives and to the function of civilization as a whole. Incidentally, the proliferation of web-connected devices require powerful computing systems on the opposing end to satiate the endless hunger for more data and information. This hunger has changed the incentives for the innovation of new communication technologies. Namely, telecommunications of the past sought to increase the distance and speed of communications (for which optical fiber technologies were revolutionary). By contrast, data communications of the present seek to decrease the viable distance for optical communications, with the hope of improving data bandwidth and energy efficiency. This has led to the replacement of electrical wires with optics and photonics within data centers, facilitating ultra-fast and efficient data communications for high-performance computing tasks. As the continued down-scaling of optical communications ensues, the big question is how much shorter of a scale is optics viable, continually replacing electronics along the way? More specifically, can optical communications be brought on-chip, facilitating intercommunications between silicon logic?

In the last three decades, the computational design of electromagnetic devices has been enabled by improved simulation techniques and high-performance computers. This has facilitated the physical investigation of micro- and nano-scale electromagnetic devices by direct solutions to Maxwell's Equations – oftentimes providing remarkable agreement with experiment. Recently, investigators have sought to engineer devices that perform at the absolute limits by leveraging computational design. Unfortunately, complex designs can require thousands to millions of simulations to exhaustively explore a design space, and exhaustive searches on this scale remain computationally precluded (even with the impressive speed of modern machines). Notwithstanding, recent revelations in computer science and mathematics (such as machine learning) have enabled optimization techniques with unprecedented computational efficiency. Can these techniques be applied to electromagnetic design?

1.1 Thesis outline

This dissertation seeks to answer the rhetorical questions posed above in two parts. In Chapter 2, we will argue that on-chip optical communications are feasible, but will require several integrated photonic devices. In particular, the central topic of this thesis is a comparison between lasers and light-emitting diodes (LEDs) – the two most viable sources for on-chip optical communications. This will require a detailed discussion of the fundamental physics of LEDs and lasers, discussed in Chapter 3. We will find that regular LEDs are not as fast lasers because of the intrinsically slow speed of spontaneous emission compared to stimulated emission. However, Chapter 4 shows that the speed of LEDs can be boosted by several orders of magnitude using optical antennas. Therefore, we argue that optical-antenna enhanced spontaneous emission can be as fast as stimulated emission. We will conclude Chapter 4 with a discussion of direct electrical modulation and efficiency in optical sources. Chapter 5 presents two supplementary topics relating to the ultimate efficiency of optical-antenna enhanced light-emitting diodes: waveguide coupling and metal-dielectric antennas. In Chapter 6 we switch topics and discuss inverse electromagnetic design via the adjoint method – a technique that can be used to dramatically reduce simulation requirements in the design of complex electromagnetic structures. We go on to apply the adjoint method to CMOS-compatible grating couplers. Finally, in Chapter 7 we briefly discuss novel topics in inverse design, including a new semi-analytical transfer-matrix method for the design of thin-film interference filters such as distributed Bragg reflectors. We conclude Chapter 7 with a brief look at emerging inverse design methods that leverage machine learning.

To maintain brevity in the main body of this thesis, several mathematically-heavy topics have been relegated to the appendices. More detailed discussion of many topics, especially from Chapters 2-5 can be found there.

Chapter 2

On-Chip Optical Data Communications: Incentives, Requirements, and Our Proposal

Global internet traffic has risen exponentially in the last decade, with 280Tb/s of information communicated on average in 2016 [92], or about 1 zettabyte/year (10^{21} bytes/year). Data centers and high-performance computing (HPC) systems make up the backbone of internet traffic and backend computation, with as much as 10^6 bits communicated locally per bit communicated externally [8]. Indeed the global data center IP traffic[1] has increased from 1 zettabye/year in 2010 to 11 zettabytes/year in 2018, with anticipated continued exponential growth for at least the next decade [6, 81]. Consequently, data centers consumed about 200 TWh (200×10^9kWH) of electricity in 2018, or about 1% of total electricity produced globally [59, 81]. Frightening projections from 2015 indicated that the total data center electricity consumption could rise to as much as 8,000TWh (>20% of anticipated global electricity production) by 2030 [6]. However, continuous improvements to data center energy efficiency as well as the adoption of "hyperscale" infrastructure for massively distributed tasks such as social media and cloud computation have appeared to curb such projections and maintain only a small rise in total energy consumption despite massive growth in usage, according to a recent report [81].

Optical interconnects have historically and continue to enable highly efficient data communications and computation. Within the last few decades, optical fibers have replaced long-range electrical wires on the scale of meters to kilometers connecting server racks in data centers. Optical data communication provides several major benefits over conventional wires including smaller attenuation, smaller dispersion, and higher data bandwidth [94]. For instance, Luxtera's transceiver in 2017 could transmit as much as 100 Gb/s over 2km without repeaters at about 4W [53], a metric unmatched by coaxial cables or ethernet cables with ranges on the order of 100m and 100x smaller bandwidth and similar power requirements. More recently, silicon photonics – an integrated microscale optical circuitry platform that is compatible with conventional CMOS silicon-on-insulator (SOI) fabrication technology – is bringing optical communication to the server- and chip-level scale in order to meet ever expanding data bandwidth requirements while reducing power consumption [92, 132]. For example, Beausoleil et al's [11] proposed

[1] This metric accounts for communications between data centers and external users, as well as intercommunications between data center compute nodes. Communications within the nodes themselves are unaccounted in this metric, and would increase this figure by several orders of magnitude.

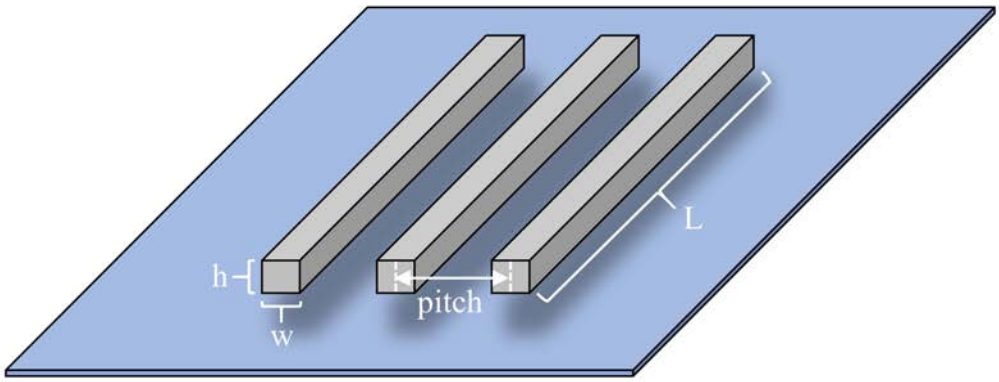

Figure 2.1: Nearest-neighbor on-chip electrical wires.

architecture for data centers and high-performance computing (HPC) clusters would employ photonic interfaces between individual compute nodes and memory with optical communication bandwidths exceeding 20Tb/s enabled by dense wavelength division multiplexing (DWDM) at energy requirements on the order of 100fJ/bit [11]. Seok et al [119] proposed a low-loss, integrated photonic switch with bandwidth and port density that are unmatched by conventional electronic gigabit switches. Tangentially, photonic platforms have recently been explored as analog accelerators for deep learning and other computationally difficult tasks [142, 144, 122]. These disruptive innovations in optics and photonics leave only one optical interconnect length scale unexplored: on the silicon chip itself with communication distances of the order of a few centimeters or less between clusters of transistors or other physical electronic devices.

2.1 Losses of conventional electrical interconnects

Currently, the on-chip data communications are facilitated by micro-scale electrical wires. Every bit of information communicated requires an electrical wire to be charged, which incurs energy per bit (energy/bit) losses. Consider the simple depiction of on-chip electrical wires in Fig. 2.1. A single wire may have width, w, and height, h, with some pitch (defined lithographically) between adjacent wires. Missing from this picture are many vertical layers of such wires for a typical chip (in order to facilitate interconnections between potentially billions of transistors) as well as vias between vertical layers. For simplicity, we may assume that the wires float between infinite ground planes above and below. To send a bit of information across the wire, the wire must be charged. The total energy to send a single bit requires one charge and one discharge of the wire capacitance resulting in a total energy requirement of,

$$\left(\frac{\mathrm{E}}{\mathrm{bit}}\right)_{\text{capacitive losses}} = CV^2 \tag{2.1}$$

where C is the wire self- and mutual- capacitances, and V is the supply voltage. According to the ITRS, modern ICs operate with approximately 0.7V supply voltage. If we take the wires to have square cross-sections with edge width of $w = h = \frac{1}{2}$pitch, and assume that the distance between the wire faces and ground planes is also $\frac{1}{2}$pitch, then the wire mutual capacitance is given by,

$$C_{\text{mutual,wire}} \approx 4 \times \varepsilon \frac{A}{d} = 4 \times \varepsilon \frac{L \cdot \frac{1}{2}\text{pitch}}{\frac{1}{2}\text{pitch}} = 4\varepsilon L \tag{2.2}$$

Assuming that the cladding medium is SiO_2 with $\varepsilon = 3.9\varepsilon_o$, then we find an approximate mutual capacitance per unit length of $C/L \approx 1.4$pF/cm. This capacitance per unit length is independent of the wire pitch and cross-sectional dimension, indicating that it is fundamental and cannot be eliminated by scaling [92]. Thus by Eq. 2.1 we find an energy/bit/length requirement exceeding $500\frac{\text{fJ}}{\text{bit}\cdot\text{cm}}$. This is demonstrably an underestimate however, as we did not include self-capacitance, fringing capacitance, nor mutual-capacitance between further neighboring wires [2]. Moreover, we did not include losses associated with on-chip repeaters to boost the transmitted signal. Indeed, it is estimated that over >50% of transistor gates serve as repeaters [92], incurring an outstanding portion of overall on-chip losses. Thus we see that for on-chip distances approaching 1cm, we already greatly exceed 100fJ/bit just in charging the wires, with an unavoidable scaling of energy/bit with communication distance[3].

2.2 Benefits of optical interconnects

By contrast, optical interconnects are essentially lossless compared to their electrical counterparts, especially on the length scale of a chip. Indeed the best known absorption coefficient of silica is on the order of ≈1dB/km at wavelengths in the L- and C-bands, essentially nullifying any potential distance based losses on-chip (though this does not include potential losses in integrate photonics such as waveguide bends or surface roughness, but these issues are minor). Furthermore, the data bandwidth that can be carried by optical interconnects is extremely high, owing to a lack of RC time delay and large optical frequencies of nearly 200THz. Indeed, the speed of optical interconnects is limited only by the transmitters and receivers.

Ideally, the only losses expended in optical interconnects are in the transmission and reception of photons, as well as the necessary electronics to convert between electrical and optical signals (including amplifiers). If we ignore the latter considerations, a perfect shot-noise limited receiver requires about 40

[2] This simple calculation of the capacitive losses may appear somewhat deceiving, because electrical interconnects acting as ideal transmission lines (instead of DC wires) should not suffer from such reactive losses nor require frequent repeaters along the length of a chip. The details are nuanced, but at the dimensions and frequencies involved in on-chip interconnects, RC-limited behavior of electrical interconnects cannot be avoided [95, 94]. One may attempt to tread the line by optimizing the wire dimensions and transmission frequency, but then one must also deal with impedance matching, fabrication sensitivities, cross-talk, and higher chip footprint requirements – problems that are not nearly as severe when adopting optical interconnects.

[3] There has been some discussion of low-swing repeaters that can operate at lower supply voltage, and therefore lower the overall energy/bit of the wires. However, even if voltage can be scaled down, the capacitance scaling with length in Eq. 2.2 remains.

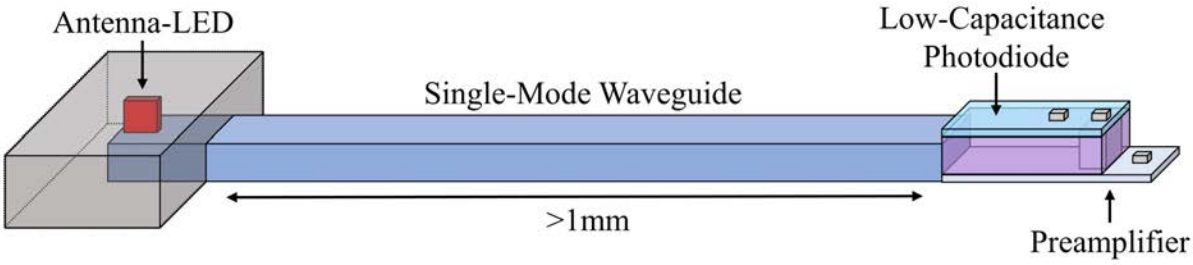

Figure 2.2: Proposed photonic link consists of an ultra-fast source (Antenna-LED), a single-mode waveguide, and a high-speed receiver (Photodiode+Preamplifier).

photons/bit (discussed in Chapter 4). All non-idealities ignored, the energy/bit required to emit 40 photons with photon energy of 1eV is only 6aJ/bit (6×10^{-18} J/bit) – representing over 5 orders of magnitude potential improvement over the ≈100 fJ/bit currently required to communicate across chip electrically. Notwithstanding, achieving this advantage implied by the use of on-chip optics remains a technologically challenging endeavour. From a cost-motivated perspective, ideal optical sources and detectors would require the heterogeneous integration of optically-active materials (such as III-V group semiconductors or 2D monolayer materials) with silicon. While there have been recent strides in these technologies [34, 9, 50, 136], high-volume production remains distant. Furthermore, from a fundamental engineering perspective, there are many obstacles towards the demonstration of a high-speed, ultra-efficient all-photonic integrated circuit with sufficient advantage over conventional CMOS.

2.3 Current conception of the photonic link, and its requirements

The nano-photonics group in the Center for Energy Efficient Electronics Science (E3S) at UC Berkeley has proposed a full optical link, depicted in Fig. 2.2. The link consists of three critical elements: (1) a nanoscale, efficient, fast, electrically-injected optical source; (2) a single-mode waveguide; and (3) a low-capacitance, high-speed receiver.

Many nanoscale optical sources have been proposed recently [109, 42, 130, 23, 43, 73, 27, 37]. The semiconductor laser is the conventional choice for an optical transmitter in a photonic link. However, we will advocate for an ultra-fast light-emitting diode called the optical antenna-enhanced light-emitting diode (antenna-LED). We will argue that antenna-LEDs can be as fast lasers while maintaining practical quantum efficiency. The antenna-LED in Fig. 2.2 consists of a narrow III-V LED ridge ($\approx$ 20nm), surrounded by a metallic antenna. Electron-injection to the LED ridge is facilitated by the antenna, while holes are injected through the waveguide (which is assumed to be electrically isolated from the antenna) [4].

The optical source transmits light through a single-mode waveguide interconnect. Targeted interconnect lengths are of the order 1mm or larger, i.e. significant distances across chip. Communications on

shorter length scales would remain relegated to electrical wires since the capacitive losses are not as severe, and denser information density is possible with multiple metallization layers. The single-mode waveguide dimensions will be on the order width $\times$ height $= 500\text{nm} \times 200\text{nm}$. This supports the fundamental TE mode and allows for high-efficiency operation of both the source and receiver.

The receiver is depicted as a high-performance bipolar phototransistor with separate photodiode and amplification regions, originally proposed by Lalau-Keraly [70]. Alternative proposals consist of low-capacitance photodiodes connected to CMOS trans-impedance amplifier circuits [120, 5]. Nevertheless, critically, the absorption and gain regions of the receiver must be separate for both high-speed and low energy/bit detection. The targeted capacitance of the photodiode for a strong photo-signal is of the order 100aF or less, which can potentially be achieved with next-generation germanium devices. One recent report claims $\approx$600aF with a 1.7μm photodiode and high responsivity [103, 104].

The optical source is the main subject of this thesis (discussed in Chapters 3-5). Our objective will be to show that the optical antenna-LED is not only a capable optical source, but will fundamentally rival the speed and efficiency of conventional semiconductor lasers. Ultimately we will find that the antenna-LED is capable of $>$50 Gbit/s direct modulation. Furthermore, with improved and scalable surface passivation, practical quantum efficiency can be achieved. The last remaining bottleneck will be optical power, where we argue that 500 photons/bit transmitted to the receiver is possible. This is below the current signal requirements of high-speed optical receivers, but next-generation receivers could potentially operate at very low power. Consequently, next-generation sub-femtojoule energy/bit on-chip optical communications is possible.

Chapter 3

Fundamentals of Semiconductor Light Emission: Lasers vs. LEDs

In the prior chapter we have advocated several reasons to implement on-chip optical interconnects, arguing primarily that optical data communications can provide increased data bandwidth and energy efficiency even on the length scale of a single chip (<1cm). Nevertheless, several optical components will need to be engineered before such a technology can be realized. Most critically, efficient and fast nanoscale light emitters and photoreceivers must be designed and properly integrated on-chip. As will be shown, these are not trivial concerns. In this thesis we are primarily interested in uncovering the best optical source in the interconnect.

Practically speaking, III-V InP/GaAs-based semiconductors, such as ternary and quaternary compounds like InGaAs and InGaAsP, are the most viable optical materials for on-chip optical communications because of their low defect density, well-known and repeatable lithographic fabrication, bright direct bandgap light emission in the low-loss telecommunications bands, high mobility, compatibility with electrical injection, and ease of epitaxial growth of lattice-matched heterostructures for good electrical confinement and mechanical properties. The downside of using III-V semiconductors is that they must be heterogeneously or monolithically integrated with silicon to facilitate communications between silicon logic. This thesis will consider heterogeneous integration out-of-scope. We are primarily interested in performing a fundamental exploration of on-chip optical sources, and whether they can fulfill the requirements of on-chip data communications. For this purpose alone, III-V semiconductors are the best choice.

Moreover, in this thesis we will advocate for the use of nanoscale light-emitting diodes (LEDs) for light emission instead of conventional semiconductor lasers. This analysis will depend heavily on the physics of semiconductor light emission. In this chapter we will introduce the fundamentals of semiconductor spontaneous and stimulated emission, and then derive a quantity related to both LED and laser speed (carrier lifetime) from experimental observations. We will find that LEDs are (unsurprisingly) optimized for modulation speed when they are heavily-doped, but the conventional knowledge and description of these physics are insufficient. By contrast, we will show a novel way to calculate laser speed from simple experimental observations. These findings will serve as a basis for comparison of the laser with the optical antenna-enhanced light-emitting diode, which will be discussed in the next chapter.

3.1 Basic concepts of semiconductor LEDs and lasers

In many ways, light-emitting diodes (LEDs) and lasers are very similar. A typical LED or laser might consist of a P-I-N diode heterostructure, where electrons are injected on the N-doped side and holes are injected on the P-doped side, with light emission occurring in the central intrinsic "active" region. For optical communications the most common active material is indium gallium arsenide ($In_{0.53}Ga_{0.47}As$) owing to its high mobility, low intrinsic defect density, and bandgap of $E_g = 0.75$eV for telecommunications in the C-band. InP is commonly chosen for the N- and P- doped heterostructure barrier regions because its lattice match to InGaAs will not cause defects and because its large bandgap $E_g \approx 1.3$eV prevents parasitic absorption. The most substantial difference between LEDs and lasers is the microscopic mechanism of light emission: spontaneous emission and stimulated emission, respectively.

A simple depiction of an (a) LED and a (b) laser is shown in Fig. 3.1. In this case, the semiconductor active regions are electrically-injected via a current source (mimicking a p-i-n diode), though in principle an external light source may be used to inject carriers. As can be seen here, the only fundamental structural difference between the LED and laser is the presence of an optical cavity (in other words, the two mirrors) for the laser. Some light remains trapped in the optical cavity and interacts with the active semiconducting material, while some light escapes the mirrors. The active material in the laser provides amplification (or gain) of the trapped light via a process called stimulated emission. The LED, on the other hand, simply emits light without any optical feedback in a process called spontaneous emission. The zoom-in insets depict the microscopic processes of spontaneous emission and stimulated emission in (a) and (b) respectively. In the case of spontaneous emission, excited electrons in the upper energy state (conduction band) spontaneously recombine with holes in the lower energy state (valence band), causing a photon emission event. In the case of stimulated emission, incident light from the left induces a recombination event of the electrons and holes, thereby emitting a photon in addition to the stimulating photon. In both cases, the output photon energy is equal to the energy of the transition, i.e. the bandgap energy ($\hbar\omega = E_g$). A mathematical description of the microscopic origin of spontaneous and stimulated emission may be found in Appendix A.

The device characteristics of LEDs and Lasers, on the other hand, can be vastly different, as qualitatively depicted in Figure. 3.2. The L-I curve in (a) provides the optical power versus current for LEDs and lasers. LEDs begin emitting light as soon as current is injected, while lasers only emit light when the optical gain exceeds the intrinsic losses in the material and optical cavity. This is manifested in the so-called threshold current, I_{th}, that must be overcome before efficient light emission begins. The speed of LEDs and lasers may be represented by the small-signal modulation bandwidth curves in (b). The speed of LEDs is roughly constant for any given current, but it is much smaller than what can be achieved by lasers at high current [1]. This is a consequence of the intrinsically slow speed of spontaneous emission versus stimulated emission. But what if spontaneous emission could be enhanced, so that efficient and fast light emission could be achieved without a threshold DC bias? This question will be the central topic of this chapter and the next two chapters, where we will be discussing the speed of lasers and LEDs in great detail. We will begin with a high-level overview of laser and LED dynamics.

[1] Note that these details, especially with regard to modulation bandwidth, are major simplifications. We will end up rejecting the notion of small-signal modulation in the next chapter.

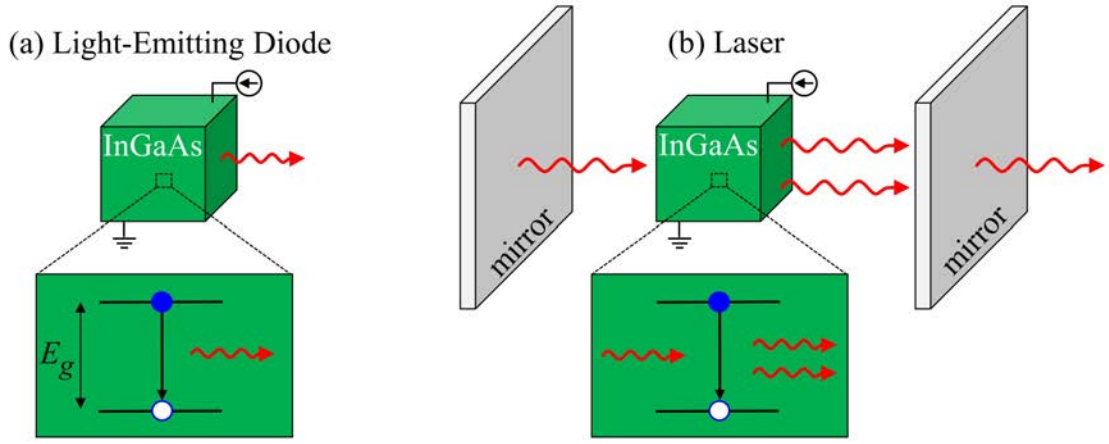

Figure 3.1: Semiconductor light-emitting diodes (a) and lasers (b) are structurally similar devices. Fundamentally, the major difference lies in the microscopic process responsible for light emission, spontaneous emission and stimulated emission for LEDs and lasers respectively (depicted in insets).

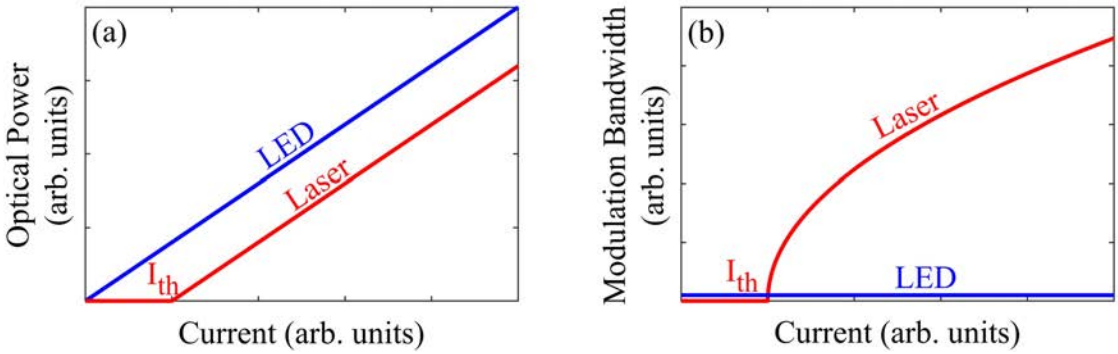

Figure 3.2: Qualitative depictions of LED and laser L-I curves (a) and small-signal modulation bandwidth (b). LEDs can emit efficiently without overcoming a threshold DC bias (I_{th}), in contrast with lasers. However, LEDs are intrinsically slow compared to lasers at large bias. Note that we have ignored quantum efficiency in (a).

3.2 A brief introduction to semiconductor carrier dynamics

In this section we will derive the carrier lifetime of LEDs and lasers, which we will argue determines the overall speed of lasers and LEDs for on-chip optical interconnects. Let the carrier concentrations of electrons and holes in a semiconductor active region be denoted by N and P respectively, both with units of carriers/cm^3. When a light-emitting diode or laser is forward-biased, electrons and holes are injected into the active region at a generation rate given by $G \propto I/qV$ where I is the current and V is the semiconductor volume. Note that the generation rate is in units $1/(\text{cm}^3 \cdot \text{s})$. By contrast, there are several mechanisms that deplete carriers in processes collectively referred to as recombination (where excited electrons and holes recombine, and are subsequently annihilated). The total recombination rate (also in units of $1/(\text{cm}^3 \cdot \text{s})$) is dominated by three co-existing processes: non-radiative recombination, spontaneous emission, and stimulated emission.

Non-radiative recombination, R_{nr}, produces heat in the semiconductor and has several contributing mechanisms such as Shockley-Read-Hall recombination, surface recombination, and Auger recombination. Radiative recombination, R_{sp}, produces light by the mechanism of spontaneous emission, and is the main light emission mechanism in LEDs. Finally, stimulated recombination, R_{st}, produces light by the mechanism of stimulated emission, and is the main light emission mechanism in lasers. Each of these recombination mechanisms have their own dependencies on the total carrier concentrations in the device as well as photon concentration in the case of stimulated emission. These dependencies will be discussed in the next few sections.

When both generation and recombination are occurring simultaneously, there is a net creation or reduction of carriers. This is reflected in the minority carrier rate equation:

$$\frac{\Delta\text{Carrier Concentration}}{\Delta\text{Time}} = \text{Generation Rate} - \text{Recombination Rate} \tag{3.1}$$

$$\frac{\partial N}{\partial t} = G - (R_{\text{nr}} + R_{\text{sp}} + R_{\text{st}}) \tag{3.2}$$

where in this case ∂N refers to the change in the excess carrier concentration, which is usually taken to be the minority carrier in the device (electrons), but in fact applies to both electrons and holes by charge conservation.

Consider the case when the generation rate is zero (Generation Rate=0), e.g. the device has been pumped to some carrier concentration N_{o} and then the current is turned off. Further, let's assume that the carrier concentration decays exponentially as $N(t) = N_{\text{o}} \exp\{-t/\tau\}$ with τ a characteristic *carrier lifetime* of decay. Plugging these assumptions into Eq. 3.2, we may solve for τ:

$$\frac{1}{\tau} \approx \frac{1}{N}(R_{nr} + R_{sp} + R_{st}) \tag{3.3}$$

Which demonstrates that the rate of decay is related to quantities R/N where R is the total recombination rate. However, note that the assumption that carriers decay exponentially with time is generally incorrect. Conventionally, the carrier lifetime is defined as,

$$\frac{1}{\tau} = \frac{\partial R}{\partial N} \tag{3.4}$$

a differential quantity that is contingent upon carrier concentration. Nevertheless, the LED speed and laser speed will be strongly related to the recombination rates R_{sp} and R_{st} respectively. In Section 3.3 we will discuss the laser recombination rate and carrier lifetime, then in Section 3.4 we will discuss the LED recombination rate and carrier lifetime. In the next chapter we will show how carrier lifetime relates to the device modulation bandwidth, or speed.

3.3 Laser speed, from observations

In Appendix A, we quantum mechanically derive the fundamental stimulated and spontaneous lifetimes for two-level systems in the presence of monochromatic light. While these lifetimes apply generally to atomic systems and reveal the fundamental relationship between stimulated and spontaneous emission, semiconductors require a slightly more nuanced treatment. This is because we must account for the continuum of states above and below the conduction and valence bands for electrons, the occupation probability of those states, and conservation of momentum and other selection rules for transitions. A full derivation is out of the scope of this thesis, but a careful accounting of all these effects produces the following net stimulated emission recombination rate:

$$R_{\text{st}} \approx \text{Group Velocity} \cdot \text{Gain} \cdot \text{Photon Density} \equiv v_g g S \tag{3.5}$$

where v_g, g, and S correspond to group velocity of the incident wave in the optical cavity, optical gain (amplification), and photon density respectively. Eq. 3.5 has a simple interpretation: photons stimulate electronic transitions, just as was indicated in Fig. 3.1(b) from the previous section. In general, the magnitude of the recombination rate depends on both a measure of how excited the material is (optical gain) as well as the density of photons available to stimulate transitions.

Using Eq. 3.4 we may estimate the carrier lifetime due to stimulated emission as,

$$\frac{1}{\tau_{\text{st}}} \equiv \frac{\partial R_{\text{st}}}{\partial N} = v_g \frac{\partial g}{\partial N} S \tag{3.6}$$

where τ_{st} is the carrier decay lifetime, and $\partial g / \partial N$ is known as the differential gain. In the remaining subsections, we will show how to obtain the stimulated lifetime by estimating these three quantities (group velocity, differential gain, and photon density) from experimental observations.

Experimental laser gain

As carefully noted in Eq. 3.5, R_{st} corresponds to the *net* stimulated emission, meaning that it is competing with the process of absorption. As famously discovered by Einstein, absorption and stimulated emission have equivalent cross sections. Therefore, when a flux of photons is incident on a semiconductor, both stimulated emission and absorption occur. However, on average there will be net stimulated emission or absorption based on the degree of population inversion in the semiconductor. This fact is reflected in the gain coefficient, g in units of 1/cm, in Eq. 3.5. This term contains information about the material matrix

element, density of states, and occupation probability of those states. For a bulk semiconductor, it may be written:

$$g = g_{\text{o}}(f_c - f_v) \tag{3.7}$$

where g_{o} is a gain coefficient that depends only on the material matrix element and density of states, and f_c, f_v are the Fermi-Dirac distributions for the electron occupation probabilities in the conduction and valence bands respectively with quasi-Fermi levels F_c and F_v. Importantly, the gain is positive when $f_c > f_v$, known as population inversion. The condition when $f_c = f_v$ is known as the Bernard-Durafforg or transparency condition, and famously occurs when $\Delta V = F_c - F_v = E_g$ where ΔV is the potential seen by the active material and E_g is the material energy bandgap.

It is important to note that the gain in Eq. 3.7 is actually a spectrum (parameterized by the photon energy $\hbar\omega$), and stimulated emission occurs at the overlap of the gain spectrum with an optical mode in the laser cavity (i.e. the least lossy mode that meets constructive interference conditions). The optical mode is represented as the photon density term (in units of photons/cm^3) and is usually assumed to be monochromatic, defined over a small energy interval [2]. For most single-mode lasers that we will be concerned with, the photon energy in the device can be assumed to overlap with the peak spectral gain, which is close to the bandgap energy of the active material. There are several empirical models of the peak gain, such as the logarithmic model:

$$g_{\text{peak}} = g_1 \ln\left(\frac{N}{N_{\text{tr}}}\right) \tag{3.8}$$

where g_1 is a logarithmic peak gain coefficient, N is the minority carrier concentration, and N_{tr} is defined as the minority carrier concentration where the transparency condition occurs (typically $N_{\text{tr}} \approx N_c$, the conduction band density of states, though it depends on a number of device parameters like doping and strain). Thus we see that when $N = N_{\text{tr}}$, the gain is zero and the gain increases with increasing carrier concentration N above transparency.

Using Eq 3.8, the differential gain is given by,

$$\frac{\partial g}{\partial N} = \frac{g_1}{N} \tag{3.9}$$

where we evaluate the differential gain at the threshold gain condition, g_{th}, with corresponding threshold carrier concentration (N_{th}). The gain threshold occurs when the gain equals the optical losses in the laser cavity. These losses include desirable loss through the mirrors as well as parasitic "intrinsic" loss such as free-carrier absorption. A typical threshold experimental threshold gain is $g_{\text{th}} = 500\text{cm}^{-1}$ for a semiconductor laser. This value can be ascertained by a simple estimate using the mirror loss in a typical Fabry-Perot laser cavity with cleaved facets, discussed below.

The differential gain depends on the lasing material and quantum confinement. Namely, quantum well lasers tend to have larger differential gain than bulk lasers because of a sharper density of states[3] and

[2] In fact, the net stimulated emission recombination rate from Eq. 3.5 should also be viewed as defined over a small energy interval, whereas the spontaneous emission recombination rate is explicitly integrated over a large distribution of energies. This is discussed further in Appendix B.

[3] This is true theoretically speaking, but in practice the absorption edge of bulk GaAs-based materials is very sharp.

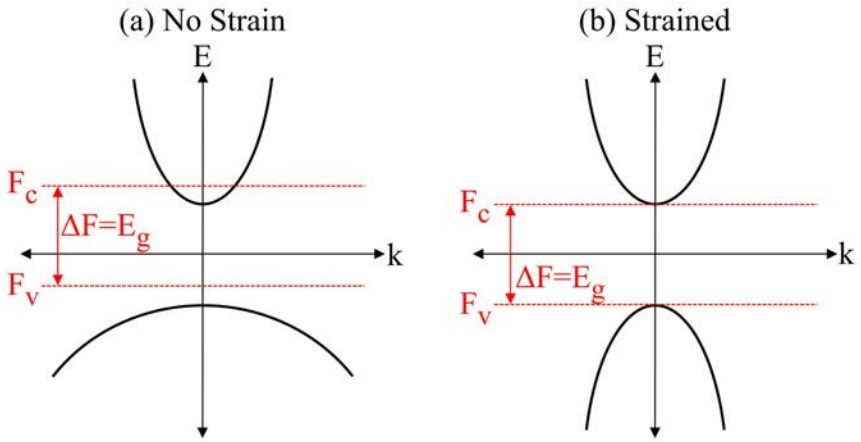

Figure 3.3: Strained semiconductors have symmetric conduction and valence bands, allowing for an overall lower carrier concentration at transparency.

strain, which can lower the transparency carrier concentration by a factor 2. The latter effect can be visualized in Fig. 3.3, where we depict the E-k diagrams of an (a) bulk and (b) strained quantum well material. Note that for illustration purposes, the materials have the same bandgap energy, but in practice a quantum well will have an effectively larger bandgap energy because of particle-in-a-box-like quantum confinement effects. As shown in Fig. 3.3(a), the bulk semiconductor has a large asymmetry between the conduction band and valence band because of a larger effective mass for heavy holes (since the heavy-hole band dominates the band edge in bulk InGaAs). Thus, at the transparency condition, $F_c - F_v = E_g$, the quasi-Fermi level for electrons is highly degenerate in order to preserve charge conservation. This is contrast to the strained quantum well in Fig. 3.3(b), which has symmetric conduction and valence bands. Strain occurs when crystals with different lattice constants are bonded or grown adjacently. Under strain, the hole effective mass in the valence band can be reduced, resembling the effective mass in the conduction band [143]. Thus, the quasi-Fermi level in the conduction band is pulled down very close to the band edge, indicating a smaller minority carrier concentration needed for transparency.

In order for the semiconductor to lase, it must be biased above transparency at the threshold condition. The carrier density at threshold depends not only on the threshold gain but also parasitic recombination mechanisms such as Shockley-Read-Hall and Auger recombination. For brevity and generality, we will make two assumptions: (1) we will take $g_1 = g_{\text{th}} = 500\text{cm}^{-1}$, and (2) we will take the threshold carrier concentration to be $N_{\text{th}} \approx 10^{18}\text{cm}^{-3}$. These are reasonable values based on typical laser parameters. Thus, we have,

$$\left.\frac{\partial g}{\partial N}\right|_{\text{threshold}} \approx \frac{500\text{cm}^{-1}}{10^{18}\text{cm}^{-3}} = 5 \times 10^{-16}\,\text{cm}^2 \tag{3.10}$$

we will use this experimental differential gain value in our estimate of the laser speed.

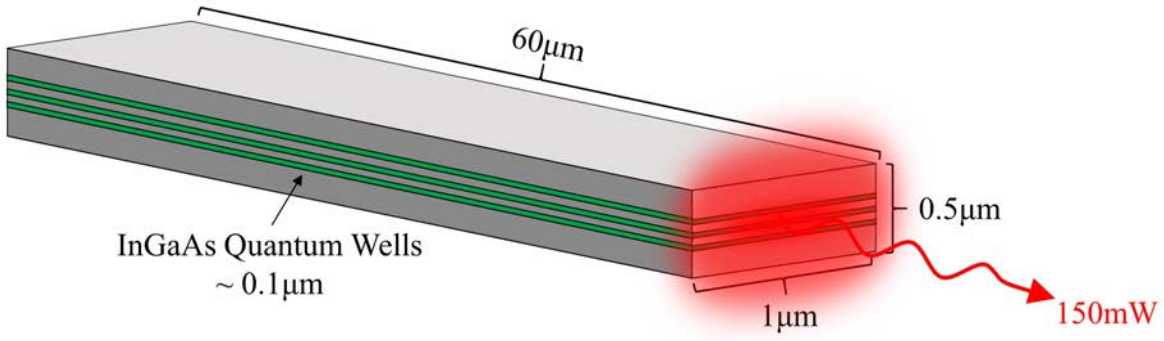

Figure 3.4: Illustration of a high-performance single-mode InGaAs multi-quantum well edge-emitting laser with one uncoated facet. We assume the other facet is perfectly reflective.

Laser photon density, from observations

Returning to the carrier lifetime due to stimulated emission from Eq. 3.6, we have two remaining unknowns: the group velocity, v_g, and the photon density, S. The group velocity refers to the propagation velocity of the laser mode, and is typically defined as $v_g = c/n_g$ where n_g is the group refractive index that depends on the waveguide materials and dimensions. We will adopt a typical value for a single-mode edge-emitting semiconductor laser, $n_g = 3$.

The photon density in a laser requires a more technical approximation. In the remainder of this subsection we will provide three methods to estimate the photon density under the assumptions that the laser is a single-mode InGaAs multi-QW edge-emitting laser with an uncoated facet. An illustration of this laser is provided in Fig. 3.4. The laser consists of a 60μm waveguide with a $1\mu\text{m} \times 0.5\mu\text{m}$ (modal) area. We assume that the confinement factor is $\Gamma = 0.2$ corresponding to to a total thickness of the InGaAs quantum wells of 0.1μm. The heterostructure barrier regions might consist of InP or some other ternary or quaternary alloy to induce strain. We assume a reasonable (but large) output power for a single-mode laser of this geometry of 150mW [83, 139]. Without loss of generality we will take only one facet to be cleaved and the other is perfectly reflective. If both facets are cleaved, one may obtain an equivalent case for these calculations by simply doubling the length and the total output optical power (150mW through each facet).

Mirror reflectivity connects external to internal photon density

The photon density may be estimated using the reflectivity of the laser and the output laser power. Consider the illustration of the internal versus external light intensity at the laser facet in Fig. 3.5, which corresponds to the emitting end of the laser from Fig, 3.4. Here we assume a cleaved facet, thus providing a reflectivity of,

$$R \approx \left|\frac{3.4-1}{3.4+1}\right|^2 = 0.3 \tag{3.11}$$

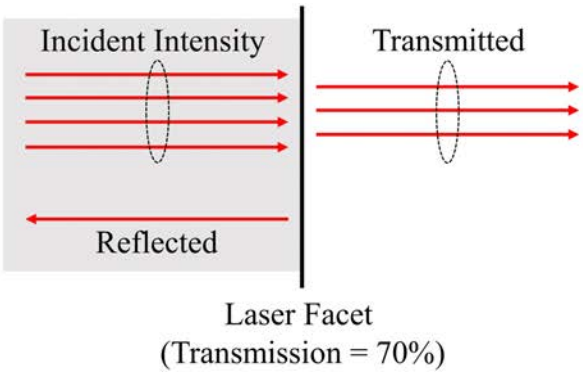

Figure 3.5: The internal laser photon density may be obtained from the facet reflectivity and the output laser intensity. See Appendix D for a detailed explanation of this calculation.

where the laser refractive index is taken to be $n = 3.4$ emitting into air $n = 1$. The incident internal laser light can either be reflected or transmitted through the facet. The transmitted light is what we observe as laser light, corresponding to 150mW of optical power. Reflected light adds to the total internal intensity, which can be converted to photon density. In Appendix D, we show that the the average internal photon density may be related to the output optical power by:

$$\text{Photon Density} = \frac{\text{Effective Refractive Index}}{2} \frac{\text{Optical Power}}{\text{Photon Energy} \cdot \text{Group Velocity} \cdot \text{Area}} \tag{3.12}$$

$$S = \frac{n_g}{2} \frac{P_{opt}}{\hbar \omega v_g A} \tag{3.13}$$

$$S = \frac{3}{2} \cdot \frac{150\text{mW}}{0.8\text{eV} \cdot 10^{10}\text{cm/s} \cdot (1\mu\text{m} \times 0.5\mu\text{m})} \tag{3.14}$$

$$S = 3.5 \times 10^{16} \frac{\text{photons}}{\text{cm}^3} \tag{3.15}$$

where the $n_g/2$ factor accounts for the incident and reflected laser intensity on the laser facet at the threshold gain condition, averaged over the laser cavity length. We assumed a photon energy $\hbar\omega = 0.8$eV in this case, since the lasing wavelength tends to occur above the bandgap energy $E_g = 0.75$eV

Matching input current with stimulated emission recombination rate

We may derive the photon density a different way by looking back at Eq. 3.2, where we described the minority carrier rate equation. Under the steady state condition, e.g. $\partial N/\partial t \rightarrow 0$, there is no net increase or decrease of the carrier density in the active material. In other words, generation exactly matches

recombination:

$$\text{Generation Rate} = \text{Recombination Rate} \tag{3.16}$$

$$\eta_i \frac{I}{qV} = R_{\text{nr}} + R_{\text{sp}} + R_{\text{st}} \tag{3.17}$$

where I is the current, V is the active region volume, and η_i is the "injection efficiency" which describes how much current injected into the device participates in generation and recombination processes. For our purposes we may take $\eta_i = 1$, assuming that the heterostructure is well-engineered. The recombination rates R_{nr}, R_{sp}, and R_{st} describe non-radiative recombination, radiative recombination due to spontaneous emission, and stimulated emission respectively. For lasers, R_{nr} and R_{sp} are considered parasitic and must be overcome for lasing to occur. They can be collected into a threshold current term I_{th} so that Eq. 3.17 may be written,

$$\frac{I - I_{\text{th}}}{qV} = R_{\text{st}} \tag{3.18}$$

The stimulated emission recombination rate may then be replaced by Eq. 3.5, $R_{\text{st}} = v_g g S$. Rewriting Eq. 3.18 in favor of the photon density, S,

$$S = \frac{I - I_{\text{th}}}{qV} \frac{1}{v_g g} \tag{3.19}$$

Previously we took the gain to be $g = 500\text{cm}^{-1}$, and from Fig. 3.4 we have the laser (active) volume $V = (1\mu\text{m} \times 0.1\mu\text{m} \times 60\mu\text{m})$. The only remaining unknown is the current. Following the logic that each electron-hole pair recombining above threshold corresponds to one output photon, then we must have that the current above threshold corresponds to $\hbar\omega(I - I_{\text{th}})/q = P_{\text{opt}}$ where P_{opt} is the optical laser power and $\hbar\omega$ is the photon energy. With $\hbar\omega \approx 0.8\text{eV}$ and P_{opt} =150mW, then $I - I_{\text{th}} \approx 187.5\text{mA}$. Plugging into Eq. 3.19,

$$S = \frac{187.5\text{mA}}{q \cdot (1\mu\text{m} \times 0.1\mu\text{m} \times 60\mu\text{m})} \frac{1}{10^{10}\text{cm/s} \cdot 500\text{cm}^{-1}} \tag{3.20}$$

$$S = 3.9 \times 10^{16} \frac{\text{photons}}{\text{cm}^3} \tag{3.21}$$

which agrees with our previous estimate in Eq. 3.15.

Using the empirical gain saturation coefficient

Since the recombination rate for stimulated emission is proportional to S, one might expect the fastest lasers to have the largest possible S. This is partially true, but really one wants the largest gain-photon density product. The gain and photon density tend to be intimately related because higher gain implies larger loss which, in turn, implies lower photon density. There is no obvious way to optimize this trade-off, but the fastest lasers experimentally tend to have threshold gain similar to what we have assumed in this work.

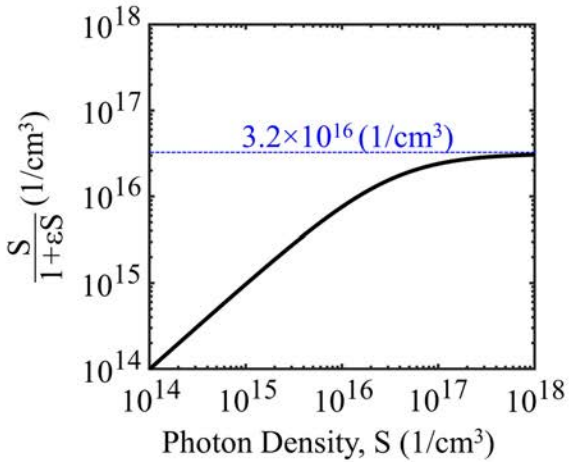

Figure 3.6: As the photon density in a laser increases, the gain-photon density product saturates. This provides us with an effective saturation photon density.

The most straightforward way to increase the gain-photon density product is just to pump the laser very hard. Unfortunately, there are limits to this approach, both fundamental and non-fundamental. Non-fundamentally, the laser may heat up or the injection efficiency may suffer [4]. Ostensibly, these issues may be relieved by using a high thermal conductivity substrate or by improved heterostructure engineering, so we will assume that they will one day be resolved. Fundamentally, lasers suffer from gain saturation at high photon density. The most important mechanism for gain saturation is somewhat elusive, and is most likely a combination of many effects, such as spectral hole burning and quantum capture time [148, 60]. Nevertheless, an empirical model for gain saturation is given by,

$$R_{\text{st}} = v_g g S \rightarrow v_g \frac{g_{\text{nominal}}}{1 + \varepsilon S} S \tag{3.22}$$

where the left hand side of the arrow is the regular expression for the stimulated emission recombination rate, while the right hand side describes the behavior at large S. We see that the gain term becomes a function of the photon density, parameterized by a "gain compression coefficient" ε which is given in units of volume. The nominal gain, g_{nominal} describes the threshold gain at low pumping. In Fig. 3.6 we plot the right hand side of Eq. 3.22 versus S. (excluding v_g and g_{nominal} because they are just constants).

We observe that at large photon density, the gain-photon density product saturates, this can be viewed another way by taking the limit of Eq. 3.22:

$$\lim_{S \rightarrow \infty} R_{\text{stim}} \rightarrow v_g \frac{g_{\text{nominal}}}{\varepsilon} \tag{3.23}$$

[4] Laser self-heating is still considered one of the primary limiters of laser speed. In fact, a recent paper from NTT demonstrated the fastest small-signal modulation bandwidth to date by bonding the laser to a silicon carbide substrate, which has a high thermal conductivity [145]

By comparison with the original stimulated emission recombination rate, we may interpret the gain compression coefficient as a saturation photon density with $S_{\text{sat}} = 1/\varepsilon$. A typical experimental value of the gain compression coefficient is $\varepsilon = 3.16 \times 10^{-17}\text{cm}^3$ [83]. The saturation photon density (for 50% gain saturation) is then equal to

$$S_{\text{sat}} = 3.2 \times 10^{16} \frac{\text{photons}}{\text{cm}^3} \tag{3.24}$$

once again agreeing well with our previous estimates [5].

Stimulated emission carrier lifetime

In the previous subsection we estimated the group velocity, differential gain, and photon density of a saturated strained quantum well InGaAs laser. We are now prepared to predict the carrier lifetime due to stimulated emission. The equation for the stimulated emission carrier lifetime is reproduced here:

$$\frac{1}{\tau_{\text{st}}} = v_g \frac{\partial g}{\partial N} S \tag{3.25}$$

We take, $v_g = 10^{10}$cm/s, $\partial g/\partial N = 5 \times 10^{-16}\text{cm}^2$, and $S = 3.5 \times 10^{16}1/\text{cm}^3$. Plugging in:

$$\frac{1}{\tau_{\text{st}}} = \left(10^{10} \frac{1}{\text{cm}}\right) \cdot \left(5 \times 10^{-16}\text{cm}^2\right) \cdot \left(3.5 \times 10^{16}\text{cm}^{-3}\right) \tag{3.26}$$

To which we find,

$$\tau_{\text{st}} \approx 6\text{ps} \tag{3.27}$$

a very fast lifetime. We will discuss how τ_{st} relates to modulation speed in the next chapter.

To summarize, in this section we derived the carrier lifetime due to stimulated emission directly from observations. Each observation provided an independent confirmation of our final carrier lifetime of $\tau_{\text{st}} = 6$ps for a conventional edge-emitting InGaAs quantum well laser. In the next section we will discuss the carrier lifetime due to spontaneous emission in LEDs.

3.4 LED speed, from observations

In this section we discuss the LED carrier lifetime, starting with the well-known empirical recombination rate due to spontaneous emission:

$$R_{\text{sp}} \approx B_{\text{o}}(NP - n_i^2) \tag{3.28}$$

where B_{o} is known as the radiative recombination coefficient in units of cm^3/s, N is the electron concentration, P is the hole concentration, and n_i is the intrinsic carrier concentration. B_{o} depends on the

[5] Note that more or less gain saturation than 50% may be tolerable, so the number chosen here is flexible.

matrix element of the material, whereas N and P take into account the electron and hole occupations as well as material doping. Note that a consequence of Eq. 3.28 is that spontaneous emission occurs in semiconductors even at small forward bias (causing a small separation in the quasi-Fermi levels, F_c and F_v, and slight excess carrier concentrations), in contrast to the population inversion condition required for lasing. This will become important in our analysis of LEDs compared to lasers later on.

Suppose we were to p-dope our LED with an acceptor concentration N_A such that $P_\mathrm{o} \approx N_A \gg n_i$. Then Eq. 3.28 suggests that the spontaneous emission recombination rate becomes $R_\mathrm{sp} \approx B_\mathrm{o} N P_\mathrm{o}$ (under the assumption that $N \ll P_\mathrm{o}$). In other words, the spontaneous emission recombination rate increases with doping. If we were to believe this model, the spontaneous emission carrier lifetime would be given by,

$$\text{Empirical BNP model: } \frac{1}{\tau_\mathrm{sp}} = \frac{\partial R_\mathrm{sp}}{\partial N} = B_\mathrm{o} P_\mathrm{o} \tag{3.29}$$

which suggests that the spontaneous emission carrier lifetime can be improved arbitrarily with doping concentration. While this behavior is true at nondegenerate doping concentrations $N_A \ll N_v$ where N_v is the effective density of states in the valence band, we will show that Eq. 3.29 fails at heavy doping concentrations. We will ultimately find that at high doping density the spontaneous emission carrier lifetime in semiconductors saturates to the fundamental two-level system spontaneous emission lifetime.

Intricate model of carrier lifetime due to spontaneous emission

Following Chuang [21], Appendix B provides a method to carefully calculate the spontaneous emission from III-V semiconductors under arbitrary doping and biasing conditions. Using this model, we solve for the radiative lifetime (i.e. the spontaneous emission carrier lifetime) of bulk InGaAs as a function of p-doping concentration in Fig. 3.7. At each doping concentration, the minority carrier concentration of electrons is taken to be constant $N = 10^{16}\mathrm{cm}^{-3}$, corresponding to a reasonable biasing condition [6]. As indicated in Fig 3.7(a)-(c) there are three characteristic regimes of the spontaneous emission carrier lifetime depending on the magnitude of doping, which we discuss below.

Region I: $N_A \ll N$

This region corresponds to Fig. 3.7(a) where the doping concentration, N_A, is smaller than the device minority carrier concentration (in this case $N = 10^{16}\mathrm{cm}^{-3}$ due to the device bias). As a consequence of charge conservation, $P \approx N$. Thus, using Eq. 3.28, $R_\mathrm{sp} \approx B_\mathrm{o} N^2$. Consequently,

$$\frac{1}{\tau_\mathrm{sp}} = \frac{\partial (B_\mathrm{o} N^2)}{\partial N} = 2 B_\mathrm{o} N \tag{3.30}$$

In other words, we find that the spontaneous emission carrier lifetime is independent of doping concentration in this regime, manifesting as a nearly slopeless line in Fig. 3.7(a).

[6] Note that under large minority carrier concentration both the BNP model and the model that is discussed in this section will break down (discussed in Appendix C). We will not consider this additional case as it would likely require a more complicated model of the semiconductor bandstructure, and is undesirable for quantum efficiency considerations.

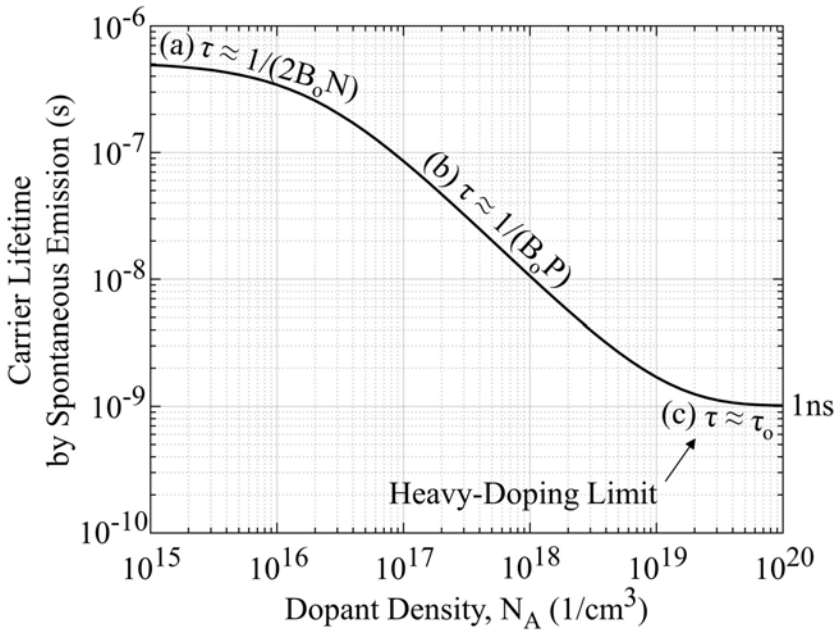

Figure 3.7: The carrier lifetime due to spontaneous emission in semiconductors saturates to a constant value at heavy doping. There are three characteristic regimes of interest: (a) the doping concentration is smaller than the minority carrier concentration, (b) the doping concentration is larger than the minority carrier concentration but nondegenerate, (c) the doping concentration is highly degenerate. (a) and (b) follow the conventional BNP model of spontaneous emission, while (c) requires a more careful consideration of the physics at play.

Region II: $N_A \gg N$ **and** $N_A \ll N_v$

This region corresponds to Fig. 3.7(b) where the doping concentration is nondegenerate (i.e. smaller than the effective valence band density of states N_v), but also exceeds the minority carrier concentration $N = 10^{16}\text{cm}^{-3}$. These conditions correspond to the empirical $B_\text{o}NP$ model in Eq. 3.29, where we can now replace $P \approx P_\text{o} = N_A$ since the doping density greatly exceeds the excess carriers. Consequently,

$$\frac{1}{\tau_\text{sp}} = \frac{\partial(B_\text{o}NP_\text{o})}{\partial N} = B_\text{o}P_\text{o} \tag{3.31}$$

to which we find that the carrier lifetime due to spontaneous emission is inversely proportional to the doping concentration with a -1 decade/decade slope in the log-log plot of Fig. 3.7(b).

Region III: $N_A \gg N$ **and** $N_A \gg N_v$

This region corresponds to Fig 3.7(c) where the doping concentration is degenerate (i.e. much larger than the effective valence band density of states N_v). A major contribution in this thesis is the insight that the

$B_{\text{o}}NP$ model from Eq. 3.28 fails in this regime of heavy doping concentration. Indeed, it is shown in Appendix C that at high P-doping concentration, the spontaneous emission recombination rate saturates to the following,

$$\text{Novel heavy-doping model: } R_{\text{sp}} = \frac{N}{\tau_{\text{o}}} \tag{3.32}$$

where N is the minority carrier concentration, and τ_{o} is the well-known spontaneous emission lifetime of a two-level system, provided here:

$$\frac{1}{\tau_{\text{o}}} = \frac{|qx_{21}|^2\omega^3 n}{3\hbar\varepsilon_{\text{o}}c^3} \tag{3.33}$$

where n is the material refractive index, ω corresponds to the frequency of emission near bandgap, and $|qx_{21}|$ is called the dipole matrix element. $|x_{21}|$ represents a dipole moment in units of length, and is specific to a given material. Eq. 3.33 is derived in Appendix C. Amazingly, Eq 3.32 implies that the spontaneous emission carrier lifetime in semiconductors is limited by the fundamental two-level system lifetime of spontaneous emission:

$$\frac{1}{\tau_{\text{sp}}} = \frac{\partial(N/\tau_{\text{o}})}{\partial N} = \frac{1}{\tau_{\text{o}}} \tag{3.34}$$

This limiting behavior can be seen by the saturation of the carrier lifetime due to spontaneous emission with heavy doping concentration $N_A > 10^{19}\text{cm}^{-3}$ in Fig. 3.7(c) to a value of approximately 1ns. This behavior occurs because the extreme doping concentration depletes all electrons on the valence band edge. Indeed the electron occupation probability is approximately 9% at the valence band edge for a doping concentration of $N_A > 10^{19}\text{cm}^{-3}$. Paired with the nondegenerate electron concentration in the conduction band, the semiconductor appears like a two-level system for photons emitting at bandgap.

Eq 3.34 effectively serves as a speed limit for the LED. One may ask whether it is possible to reach this limit in a practical LED. While possible to degenerately dope a III-V semiconductor to this level, it is usually not done in practice for several reasons, the most fundamental being parasitic Auger recombination which limits the device efficiency. We will discuss this more in the context of LED quantum efficiency in the next chapter. moreover, this behavior will be crucial for our discussion of a new device, the antenna-enhanced light-emitting diode in the next chapter.

Fig. 3.7(c) indicates that the spontaneous emission carrier lifetime in InGaAs LEDs saturates to 1ns. In the subsections below, we will show how this value of 1ns may be ascertained from experimental observations.

Dipole transition matrix element, from observations

To estimate the doping-limited carrier lifetime due to spontaneous emission in semiconductors,

$$\frac{1}{\tau_{\text{sp}}} = \frac{1}{\tau_{\text{o}}} = \frac{|qx_{21}|^2\omega^3 n}{3\hbar\varepsilon_{\text{o}}c^3} \tag{3.35}$$

we note that the only unknown is the dipole matrix element $|qx_{21}|$. We will now show three ways to find the matrix element in InGaAs.

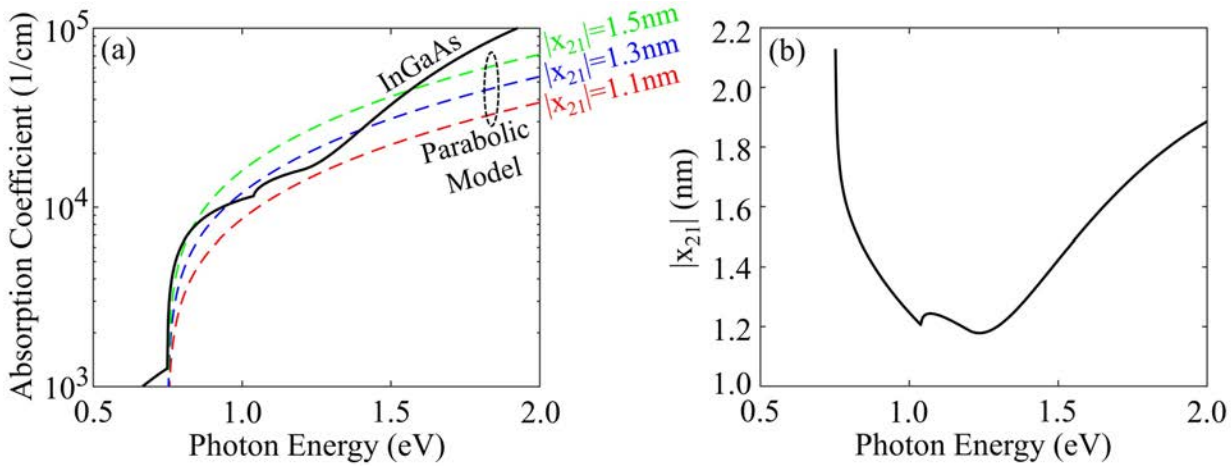

Figure 3.8: Absorption coefficient of InGaAs (a) and calculated dipole matrix element $|x_{21}|$ assuming a parabolic joint density of states (b). Note that in (a) we plot both the InGaAs absorption data from Adachi [1], along with absorption coefficient curves generated under the assumption of a parabolic joint density of states.

Using the experimental absorption coefficient

The absorption coefficient for a bulk semiconductor is given by [21],

$$\alpha_{\text{o}} = \frac{\pi\omega|qx_{21}|^2}{3\varepsilon_o cn}\rho_r(\hbar\omega - E_g) \tag{3.36}$$

where $\rho_r(\hbar\omega - E_g)$ is known as the parabolic joint density of states for electrons and holes:

$$\rho_r(\hbar\omega - E_g) = \frac{1}{2\pi^2}\left(\frac{2m_r^*}{\hbar^2}\right)^{3/2}\sqrt{\hbar\omega - E_g} \tag{3.37}$$

The only parameters needed to define the joint density of states are the bandgap $E_g = 0.74$eV and the reduced effective mass m_r^*. The reduced effective mass can be estimated from the electron effective mass in the conduction band, $m_e^* = 0.041m_{\text{o}}$, and the (heavy-hole) effective mass in the valence band $m_h^* = 0.5m_{\text{o}}$, such that:

$$\frac{1}{m_r^*} = \frac{1}{m_e^*} + \frac{1}{m_h^*} \tag{3.38}$$

$$m_r^*|_{\text{InGaAs}} = 0.038m_{\text{o}} \tag{3.39}$$

Using Eq. 3.36, we may obtain the dipole matrix element by inverting the equation in favor of $|x_{21}|$:

$$|x_{21}| = \sqrt{\alpha_{\text{o}}\frac{3\varepsilon_o cn}{\pi\omega q^2}\rho_r^{-1}(\hbar\omega - E_g)} \tag{3.40}$$

where we note that the only unknown is the experimental absorption coefficient. In Fig. 3.8 we plot the absorption coefficient for InGaAs (from Adachi [1]). In addition, we plot curves generated from Eq. 3.36 assuming a parabolic joint density of states and various candidate values of the dipole moment matrix element, $|x_{21}| = \{1.1\text{nm}, 1.3\text{nm}, 1.5\text{nm}\}$. Clearly, the parabolic joint density of states model is insufficiently expressive to capture all of the features in the experimental data. Most pertinently, the absorption coefficient along the band-edge is very sharp. Sometimes this feature is modeled with a non-parabolicity parameter, but in this case we will simply use the dipole matrix element that best captures the absorption coefficient at slightly larger photon energy where the joint density of states is approximately parabolic. This can be seen in Fig. 3.8(b) where we plot Eq. 3.40 after substituting the experimental data for α_{o}. In the photon energy range of $\hbar\omega \approx 1\text{eV}$ to $\hbar\omega \approx 1.25\text{eV}$, the calculated dipole matrix element is approximately constant with a value of:

$$|x_{21}| = 1.2\text{nm} \tag{3.41}$$

Though we note that the argument for a larger dipole moment could easily be made based on the large absorption near the band-edge.

Using $k \cdot p$ perturbation theory and the experimental effective mass

The dipole matrix element may also be estimated using second-order perturbation theory. It can be shown that $|x_{21}|$ is related to the momentum matrix element $|p_{21}|$ by [21],

$$|x_{21}| = \frac{|p_{21}|}{m_o \omega} \tag{3.42}$$

where m_{o} is the electron mass. When light is incident upon a semiconductor crystal, free particles are accelerated by the electric field, This induces a perturbation energy given by,

$$E' = \frac{|V'_{21}|^2}{E_2 - E_1} = \frac{\hbar^2 k^2 |p_{21}|^2 / m_{\text{o}}^2}{\hbar\omega} \tag{3.43}$$

where $V = -j\frac{\hbar^2 k}{m_o}\frac{d}{dx} = \frac{\hbar k}{m_{\text{o}}}p$ is the perturbation potential. The right hand side of Eq. 3.43 is commonly interpreted as an effective mass, m_{eff}, by comparison with the particle's kinetic energy:

$$\frac{\hbar^2 k^2}{2m_{\text{eff}}} \equiv \frac{\hbar^2 k^2 |p_{21}|^2 / m_{\text{o}}^2}{\hbar\omega} \tag{3.44}$$

where we may solve for m_{eff},

$$m_{\text{eff}} = \frac{m_{\text{o}}^2 \hbar\omega}{|p_{21}|^2} \tag{3.45}$$

We may now solve Eq. 3.45 for $|p_{21}|$ and combine it with Eq. 3.42 to find $|x_{21}|$:

$$|x_{21}| = \sqrt{\frac{\hbar}{2m_{\text{eff}}\omega}} \tag{3.46}$$

In this case we will take the effective mass to be the reduced effective mass of InGaAs from Eq. 3.39, $m_r^* = 0.038 m_\mathrm{o}$. Plugging this into Eq. 3.46 and taking the photon energy to be $\hbar\omega \approx 0.77$eV (slightly above bandgap), we have,

$$|x_{21}| = 1.1\mathrm{nm} \tag{3.47}$$

Agreeing reasonably with our previous estimate of 1.2nm in Eq. 3.41.

Using the experimental E_p parameter

The momentum matrix element is commonly given in terms of an E_p energy parameter by the relation [21],

$$|p_{21}|^2 = \frac{m_\mathrm{o}}{2} E_p \tag{3.48}$$

where m_o is the electron mass. Using $|x_{21}| = |p_{21}|/m_\mathrm{o}\omega$ from Eq. 3.42 we once again may convert this to the dipole matrix element, to which we find,

$$|x_{21}| = \sqrt{\frac{E_p}{2 m_\mathrm{o} \omega^2}} \tag{3.49}$$

Note that this is almost exactly equivalent to Eq. 3.46 except we have substituted the effective mass with an experimental E_p parameter. Substituting $E_p = 25.7$eV for InGaAs [21] we find,

$$|x_{21}| = 1.3\mathrm{nm} \tag{3.50}$$

which agrees remarkably well with our previous estimates.

Carrier lifetime due to spontaneous emission

Returning to Eq. 3.35, we may now evaluate the carrier lifetime due to spontaneous emission under heavily-doped conditions. Using $|x_{21}| = 1.3$nm, $n = 3.4$, and $\hbar\omega = 0.77$eV we find,

$$\frac{1}{\tau_\mathrm{sp}} = \frac{|q x_{21}|^2 \omega^3 n}{3\varepsilon_\mathrm{o} \pi \hbar c^3} \tag{3.51}$$

$$\tau_\mathrm{sp} = 1\mathrm{ns} \tag{3.52}$$

which represents the fastest radiative lifetime of a bulk InGaAs LED. Note that this value corresponds to the saturation lifetime at heavy doping seen in the full model of LED carrier lifetime in Fig. 3.7.

3.5 Outlook and needed spontaneous emission rate enhancement

Up to this point we have demonstrated the following:

1. We have derived the recombination rate and carrier lifetime due to stimulated emission in a single-mode InGaAs quantum well edge-emitting laser using experimental observations.

2. For a laser operating at saturation, we arrived at a carrier lifetime due to stimulated emission of $\tau_{st} = 6$ps.

3. We have shown a novel way to calculate the recombination rate and carrier lifetime due to spontaneous emission in bulk InGaAs LEDs. In doing so, we rejected the conventional BNP model of spontaneous emission in heavily doped LEDs, ultimately finding that the spontaneous emission recombination rate saturates to N/τ_o where τ_o is the fundamental spontaneous emission lifetime of a two-level system.

4. For LEDs doped beyond $N_A = 10^{19}\text{cm}^{-3}$, the spontaneous emission carrier lifetime converges to $\tau_{sp} = 1$ns.

Immediately we find that the lasers have a much faster carrier lifetime than LEDs. Even worse, LEDs must be heavily doped to maximize the spontaneous emission recombination rate, which will inevitably result in inefficiency due to Auger recombination. Lasers, on the other hand, can be very efficient when operating at high bias. Thus, at face value, lasers would be considered the best option for on-chip optical communications (and telecommunications at large).

But what if the carrier lifetime of spontaneous emission could be enhanced dramatically, to the extent that LED speed could compete with laser speed? In the next chapter we will demonstrate the optical antenna-enhanced light-emitting diode (antenna-LED), a device capable of enhancing spontaneous emission by the factor $\frac{1/\tau_{st}}{1/\tau_{sp}} = 166$ that would be needed for LEDs to be as fast as lasers. Additionally, we will answer some remaining fundamental questions about the implementation of optical sources in an on-chip optical link. For example, how does the carrier lifetime relate to the actual modulation speed of the device (laser or antenna-LED)? And, what are the remaining challenges for engineering a full transmitter-to-receiver optical link?

Chapter 4

The Optical Antenna-LED: Spontaneous Emission as Fast as Stimulated Emission

In the last chapter we demonstrated that semiconductor lasers are much faster than any conventional light-emitting diode, with carrier recombination lifetimes due to stimulated emission and spontaneous emission of $\tau_{st} = 6$ps and $\tau_{sp} = 1$ns respectively. In this chapter we will describe the optical antenna-enhanced light-emitting diode (antenna-LED), a device capable of boosting the intrinsic spontaneous emission rate of LEDs, to the extent that it could rival the stimulated emission rate in lasers. First, we will briefly describe the theory of spontaneous emission enhancement. Then, we will discuss the antenna-LED speed compared to the laser discussed in the previous chapter. Finally, we will provide a detailed discussion of the optical antenna-LED device metrics in the context of implementing on-chip optical communications, including an optical link analysis and remaining challenges.

4.1 Spontaneous emission enhancement

In Appendix A we provided a derivation of the fundamental spontaneous emission rate for a two-level system in a homogeneous medium (e.g. a semiconductor crystal). In doing so, we found that spontaneous emission is a consequence of a dipole perturbation potential to the quantum state, which can be expressed as the product of the zero-point electric field and the dipole moment matrix element of the excited electronic state. However, spontaneous emission can also be viewed from a much simpler semi-classical perspective. In particular, a simple optical antenna circuit model can adequately describe the intrinsic spontaneous emission rate of an excited atom. Consequently, spontaneous emission enhancement (i.e. increasing the rate of spontaneous emission) can be viewed entirely as an antenna property. This will be demonstrated below. Spontaneous emission enhancement from the purview of quantum mechanics and the Purcell factor will be discussed in the next chapter.

Spontaneous emission as an antenna property

Consider a simple oscillating dipole with length l and angular oscillation frequency ω. This might represent an excited atom, or an electron-hole pair in a semiconductor. The oscillating dipole may be treated as a very small antenna. The power radiated by a classical Hertzian dipole antenna (electrically small antenna of this nature) can be modeled as a lumped-element AC circuit model [30, 67, 20, 140, 118] with radiated power on resonance given by,

$$P_{\text{rad}} = \frac{1}{2}|I|^2 R_{\text{rad}} \tag{4.1}$$

Where $|I|$ is the peak antenna current amplitude and R_{rad} is known as the antenna radiation resistance. R_{rad} is fundamental and can be derived using the magnetic vector potential induced by the antenna current and integrated over the resulting far-field radiation [118]. For dipole antennas, R_{rad} is given by,

$$R_{\text{rad}} = \frac{2}{3}\pi Z_{\text{o}} \left(\frac{l}{\lambda}\right)^2 n \tag{4.2}$$

where $Z_{\text{o}} = 1/\varepsilon_{\text{o}} c \approx 377\Omega$ is the impedance of free space, l is the antenna length, λ is the free-space emission wavelength, and n is the refractive index of the cladding medium.

The current of the point dipole is given by $|I| = q\omega$ [30], and can be regarded as a quantum mechanical current associated with the oscillating electron-hole pair. Furthermore, we may correlate the antenna length with the dipole moment matrix element[1], with $l = 2|x_{21}|$. Plugging in these values, we find that the total power radiated of the Hertzian dipole (denoted by P_{o}) is,

$$P_{\text{o}} = \frac{1}{2}|q\omega|^2 \frac{2}{3}\pi Z_{\text{o}} \left(\frac{2|x_{21}|}{\lambda}\right)^2 n \tag{4.3}$$

Through some additional manipulation with $2\pi c/\omega = \lambda$ and $Z_{\text{o}} = 1/\varepsilon_{\text{o}} c$, we find that Eq. 4.3 may be rewritten as,

$$P_{\text{o}} = \frac{|qx_{21}]^2 \omega^4 n}{3\pi\varepsilon_{\text{o}} c^3} \tag{4.4}$$

Or, written a different way,

$$P_{\text{o}} = \hbar\omega \frac{|qx_{21}]^2 \omega^3 n}{3\pi\varepsilon_{\text{o}}\hbar c^3} = \frac{\hbar\omega}{\tau_{\text{o}}} \tag{4.5}$$

where τ_{o} is the well-known fundamental spontaneous emission lifetime of a two-level system (provided in Eq. 3.33 from the previous chapter). Thus, spontaneous emission can be thought of as the radiation of a semi-classical dipole antenna.

[1] The factor 2 comes from correlating a linear antenna with the Bohr orbit of an atom. This may be a contentious choice, but produces the correct value of the spontaneous emission lifetime.

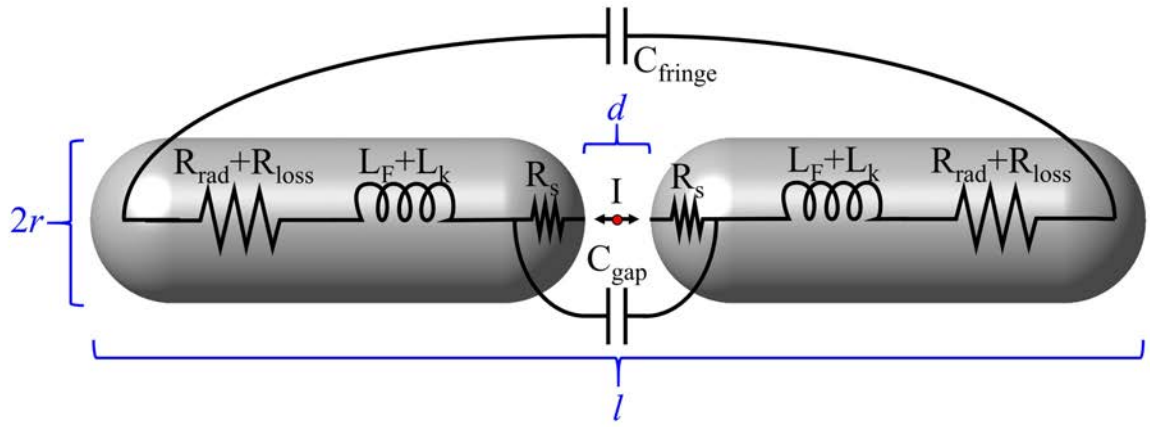

Figure 4.1: Optical dipole antenna circuit model. On resonance, the overall power radiated by the dipole point source in the central gap increases dramatically by the factor $(l/d)^2$. This manifests as an enhancement of the rate of spontaneous emission in LEDs.

Spontaneous emission enhancement as an antenna property

If spontaneous emission can be considered an antenna property, then naturally we may use the breadth of antenna theory to improve it. Indeed, by coupling a point dipole to an antenna, we may enhance (increase) the rate of spontaneous emission.

Eggleston et al [30] provided a circuit model of the optical dipole antenna for spontaneous emission enhancement. A similar circuit model is reproduced in Fig. 4.1. The antenna consists of two rounded cylindrical silver wires with radii of r, separated by a vacuum gap of width d. The total length of the antenna (including the antenna gap) is l. In this case we assume that a point dipole source, such as an atom or dye molecule, resides in the antenna gap and serves as a current source. There are a number of antenna circuit parameters in Fig 4.1 including resistance, inductance, and capacitance. The resistance terms are the radiation resistance R_{rad}, the Ohmic loss (wire) resistance R_{loss}, and spreading resistance R_{s}, respectively. The inductance terms are the Faraday (wire) inductance L_{k} and the kinetic inductance L_{k}. The capacitance terms are the gap capacitance C_{gap} and the fringe (wire) capacitance C_{fringe}. A detailed discussion of all of these circuit parameters is out-of-scope for this thesis, but some additional discussion may be found in Appendix F and a detailed description may be found in [30]. Notwithstanding, on resonance the antenna reactive impedance (inductance and capacitance) is minimized and can be approximately ignored in the dipole antenna geometry of Fig 4.1 [2].

The long metallic wires increase the total resistance seen by the point dipole source. The total power radiated on resonance is then given by,

$$P_{\text{rad}} = \frac{1}{2}|I|^2 R_{\text{rad}} \tag{4.6}$$

[2] In some cases, the reactive elements can strongly affect the antenna enhancement, this is discussed in Appendix F

Where $|I|$ is the antenna current and R_{rad} is once again the radiation resistance for a dipole antenna:

$$R_{\text{rad}} = \frac{2}{3}\pi Z_{\text{o}} \left(\frac{l}{\lambda}\right)^2 n \tag{4.7}$$

where l now refers to the full antenna length, and not the point length of the point dipole source [3].

The antenna current is provided by the oscillating point dipole source in the antenna gap. The oscillating dipole induces a current in the antenna arms according to the Shockley-Ramo effect [113], given by:

$$|I| = q\omega \frac{2|x_{21}|}{d} \tag{4.8}$$

where the original quantum mechanical current for the Hertzian dipole ($q\omega$) is scaled by a factor $2|x_{21}|/d$ to account for the dipole point charges moving in the capacitive antenna gap of width d, where $|x_{21}|$ is the dipole amplitude from the momentum matrix element [4]. Therefore, the power radiated by the antenna with variable length is given by:

$$P_{\text{rad}} = \frac{1}{2}\left|q\omega \frac{2|x_{21}|}{d}\right|^2 \frac{2}{3}\pi Z_{\text{o}} \left(\frac{l}{\lambda}\right)^2 n \tag{4.9}$$

After some additional manipulation using $2\pi c/\omega = \lambda$ and $Z_{\text{o}} = 1/\varepsilon_{\text{o}} c$, Eq. 4.9 becomes:

$$P_{\text{rad}} = \frac{|qx_{21}|^2 \omega^4 n}{3\pi\varepsilon_{\text{o}} c^3} \left(\frac{l}{d}\right)^2 = P_{\text{o}} \left(\frac{l}{d}\right)^2 \tag{4.10}$$

where P_{o} was the power radiated by the Hertzian dipole from Eq. 4.4. To find the enhancement factor of the antenna radiation, we simply take the ratio $(P_{\text{rad}}/P_{\text{o}})$:

$$\frac{P_{\text{rad}}}{P_{\text{o}}} = \left(\frac{l}{d}\right)^2 \tag{4.11}$$

thus revealing that spontaneous emission enhancement results from a simple ratio of the antenna geometrical parameters. In principle, the enhancement factor, Eq. 4.11 can be incredibly large. This is because the vacuum gap width d may be as small as 20 nanometers in size while the antenna length l may be as long as $\lambda/2$ for single-mode operation [5]. However, as we will see in the Section 4.2, the calculation of the enhancement factor for a light-emitting diode is more nuanced than suggested here; for instance, we must consider electron-hole pairs scattered spatially across the LED volume. Nevertheless, this derivation serves as a proof that spontaneous emission enhancement is a purely classical effect. This justifies the use of Maxwell's Equations solvers to calculate overall antenna-LED enhancement, which we will provide in Section 4.2.

[3] Note that this radiation resistance formula technically only applies to Hertzian dipoles. When antennas are long, approaching half-wavelength $l = \lambda/2$, one must take into account the spatial distribution of the antenna current along the wire [118]. For simplicity we will ignore this effect in this derivation, as it only amounts to a reduction factor. The full antenna circuit model provided by Eggleston et al [30] accommodates this effect.

[4] The factor 2 in the numerator accounts for the two charges (electron and hole) oscillating in the antenna gap.

[5] d is constrained to 20nm for efficiency considerations, which is discussed in depth in the next chapter.

Simple derivation of spontaneous emission enhancement

The enhancement factor in Eq. 4.11 may also be obtained by an intuitive argument, namely that the total voltage between the two antenna ends (separated by l) will drop completely across the antenna vacuum gap (with width d) because there is negligible voltage drop in the conductive metal. The spontaneous emission factor may then be estimated by the ratio of the electric field intensity in the antenna gap to that in free space [6]. Observe,

$$F \approx \frac{|E|^2_{\text{gap}}}{|E|^2_{\text{antenna}}} \approx \frac{|V/d|^2}{|V/l|^2} = \left(\frac{l}{d}\right)^2 \tag{4.12}$$

where $|E|_{\text{gap}}$ is electric field in the antenna gap, $|E|_{\text{antenna}}$ is the electric field across the antenna length, and $|V|$ is the antenna voltage. This provides the same result as above in Eq. 4.11.

4.2 Optical antenna-enhanced light-emitting diodes

In the previous section we established a theoretical basis for spontaneous emission enhancement. In particular, we found that spontaneous emission enhancement occurs when an excited electronic state (such as an atom or point dipole) is placed in a resonant electromagnetic structure such as an optical antenna. Naturally, we might consider coupling a semiconductor light-emitting diode to an optical antenna for improved LED radiation. We will refer to this device as an optical antenna-enhanced light-emitting diode, or antenna-LED for short. In this section we will introduce two viable optical antenna-LEDs that could be used in an on-chip optical interconnect, then go on to rigorously calculate the anticipated speed for one of them.

Electrically-injected antenna-LED candidates

Consider two candidate antenna-LEDs in Fig 4.2. Both antenna-LEDs consist of a p-doped InGaAs active region with a silver optical antenna on an InP-based substrate. Fig. 4.2(a) shows a conventional dipole antenna consisting of two wires with a gap d, where the optical emitter is placed. For compatibility with top-down fabrication, the wires and LED region are depicted two-dimensionally. The dipole antenna has similar enhancement scaling behavior to that discussed in the previous section. In particular, the enhancement increases proportionally to $1/d^2$ where d is the gap dimension. However, from a device fabrication point of view, the dipole antenna has a few drawbacks. For example, it is not immediately clear how one might construct p-n junction or heterostructure for good electrical contacts and confinement. Moreover, to obtain strong electromagnetic characteristics the width of the wires must be narrowly patterned and aligned to the active region. Lastly, the radiation of the dipole antenna is not directed, making it a difficult candidate for waveguide coupling for on-chip optical interconnects.

[6] As shown in Appendix A the spontaneous emission rate is proportional to the electric field squared. Antennas concentrate the zero-point electric field in the antenna gap, thereby increasing the spontaneous emission rate. This fact is also captured in the Purcell factor, which is discussed in the next chapter.

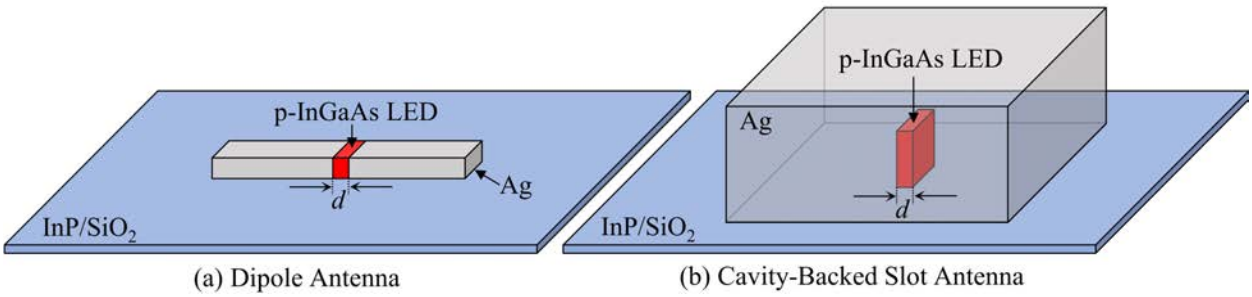

Figure 4.2: Candidate optical antenna-LED structures that are compatible with lithographic fabrication. The enhancement of the dipole antenna (a) was discussed in Section 4.1. The physics of the cavity-backed slot antenna (b) is similar [37, 36, 4]. The enhancement of each antenna is $\propto 1/d^2$.

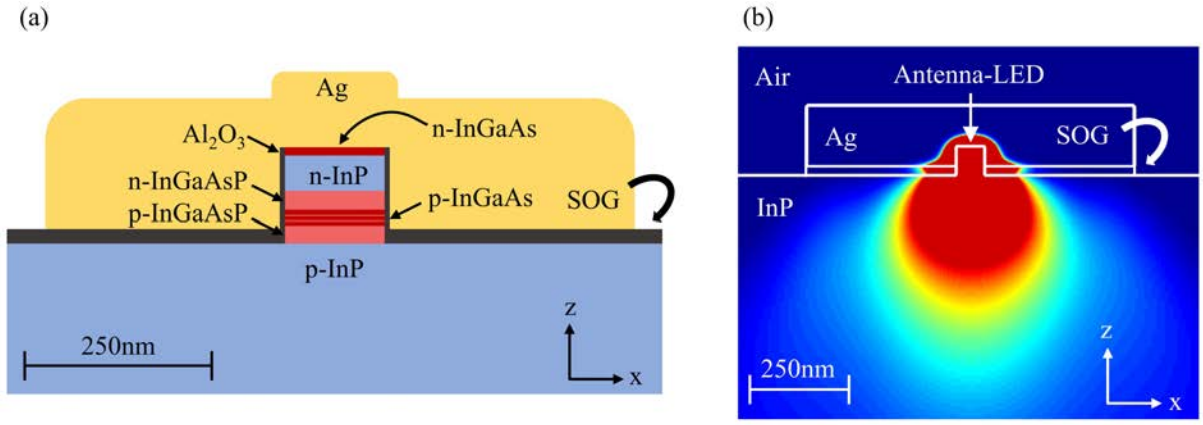

Figure 4.3: The cavity-backed slot antenna-LED is compatible with electrical-injection, and emits radiation into the substrate. Cross-sectional view provided in (a), and simulated radiation pattern (b). Figure reproduced from [4].

By contrast, the cavity-backed slot antenna-LED in Fig. 4.2(b) consists of a self-aligned LED ridge enclosed in a metallic cavity with one open face (towards the InP substrate). This is illustrated by the transparent view through the metal. The cavity-backed slot antenna also has an approximate enhancement scaling behavior of $1/d^2$ where d is the width of the narrowest ridge dimension. The advantages of cavity-backed slot antenna are illustrated more clearly in the cross-sectional views in Fig 4.3.

As shown in Fig. 4.3(a), the cavity-backed slot antenna is self-aligned to an InP/InGaAs/InP ridge. The antenna is electrically connected to the top of the ridge, where it is used as a contact to inject electrons into the n-InGaAs contact layer. The holes are injected into the p-InP layer, which is insulated from the antenna using a 40nm thick spin on glass (SOG). Finally, the InGaAs active region (in this case consisting of multiple quantum wells, but may also be treated as a bulk region) is electrically insulated from the

antenna using a 1nm thick Al_2O_3 surrounding the ridge sidewalls.

Furthermore, the height and length of the antenna-LED serve as independent degrees of freedom to tune the antenna resonance bandwidth to best match the LED material spectrum, while maximizing the radiated power for the fundamental antenna mode. This will become important for our full speed analysis below. Lastly, as depicted in Fig. 4.3(b) the radiation of the cavity-backed slot antenna is primarily directed into the substrate. In the next chapter we will show how the radiation may then be redirected into an on-chip waveguide.

These advantages make the cavity-backed slot antenna-LED an excellent candidate as a nanoscale light source for on-chip optical interconnects. Therefore, we will use it in our full-scale analysis of the antenna-LED speed below.

Detailed calculation of the radiative lifetime of the cavity-backed slot antenna-LED

In Section 4.1 we revealed the underlying physics of spontaneous emission enhancement by optical antennas. However, in that analysis we assumed that the antenna current source was a point dipole emitting at a single frequency. While this analysis provided a general prescription for how one might achieve fast antenna-enhanced LEDs, there are several nuances that were ignored. In particular, a bulk or quantum well semiconductor crystal will have spatially-distributed electron-hole pairs, represented by incoherent dipole point sources distributed throughout the semiconductor. Moreover, the intrinsic spectrum of an LED is not single-frequency, and the overlap of this spectrum with the electromagnetic antenna spectrum must be considered. Lastly, we must consider the polarization of the antenna in relation to the inherent polarization of electron-hole pairs in the semiconductor, as this will govern the physics of the dipole interaction potential matrix element.

In order to perform a detailed analysis of the total anticipated rate enhancement we must return to the fundamental physics of semiconductor light emission. For brevity, this analysis is relegated to Appendix G. After a detailed consideration of all relevant non-idealities, we arrive at the following carrier lifetime due to enhanced spontaneous emission for a heavily p-doped optical antenna-LED:

$$\frac{1}{\tau_{sp}^{*}} = \frac{F_{average}}{\tau_{sp}} \tag{4.13}$$

$$F_{average} \equiv F_{peak} \cdot \text{Polarization Average} \cdot \text{Spatial Average} \cdot \text{Spectral Average} \tag{4.14}$$

$$\frac{1}{\tau_{sp}} \approx \frac{|qx_{21}|^2 \omega^3 n}{3\pi\varepsilon_o \hbar c^3} \tag{4.15}$$

where τ_{sp} is the intrinsic spontaneous emission carrier lifetime of the heavily-doped LED (without an antenna present) and τ_{sp}^{*} is the carrier lifetime of the antenna-LED in the presence of an average enhancement factor of $F_{average}$. $F_{average}$ consists of three averaging factors: polarization, spatial, and spectral averages defined relative to the peak enhancement, F_{peak}. F_{peak} is defined as the enhancement seen by a point

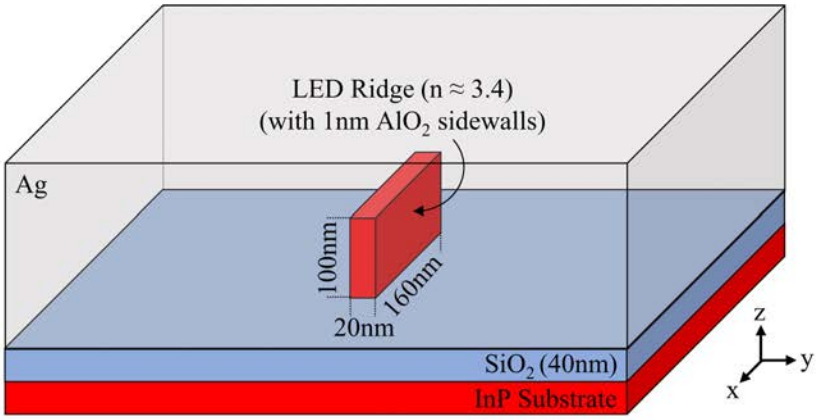

Figure 4.4: Cavity-backed slot antenna-LED used for simulation and calculation of the average enhancement factor.

dipole source placed at the optimal location in the LED and aligned with the antenna mode polarization and peak resonance frequency. Below we will perform a detailed calculation of F_{average} for a cavity-backed slot antenna.

Reference antenna-LED structure

The simulated cavity-backed slot antenna-LED for the detailed analysis is given in Fig 4.4. This is a simplified version of Fig. 4.3(a), assuming that the refractive index in the LED ridge is the same as InP ($n \approx 3.4$). Not shown is a 1nm alumina film (potentially deposited by atomic layer deposition) covering the InP ridge. This provides electrical isolation between the sidewalls of the ridge with the antenna electrode. It turns out that even a small amount of low-index material such as alumina can greatly change the antenna enhancement properties, so it was included in simulation. The total dimensions of the ridge are 20nm×100nm×160nm, which provides a single-mode antenna resonance in the C-band communications wavelength $\lambda \approx 1580$nm. The narrowest dimension is fixed to 20nm because of antenna efficiency considerations, which will be discussed in more detail in the next chapter. The height and length of the ridge serve as two parameters that can control the resonance frequency and bandwidth of the antenna. Reasonable parameters were chosen here that might be compatible with a fabrication process, but different choices could potentially yield better average enhancement factor.

Spatial average of enhancement factor

We will now solve for the Spatial Average factor from Eq. 4.14. As discussed in Appendix A, the spontaneous emission rate is a function of $|E|^2$, where E in this case refers to a zero-point electric field. In an optical antenna-LED, the optical mode (and therefore the zero-point electric field) is spatially distributed. We may approximate this electric field by exciting the antenna mode in simulation using an

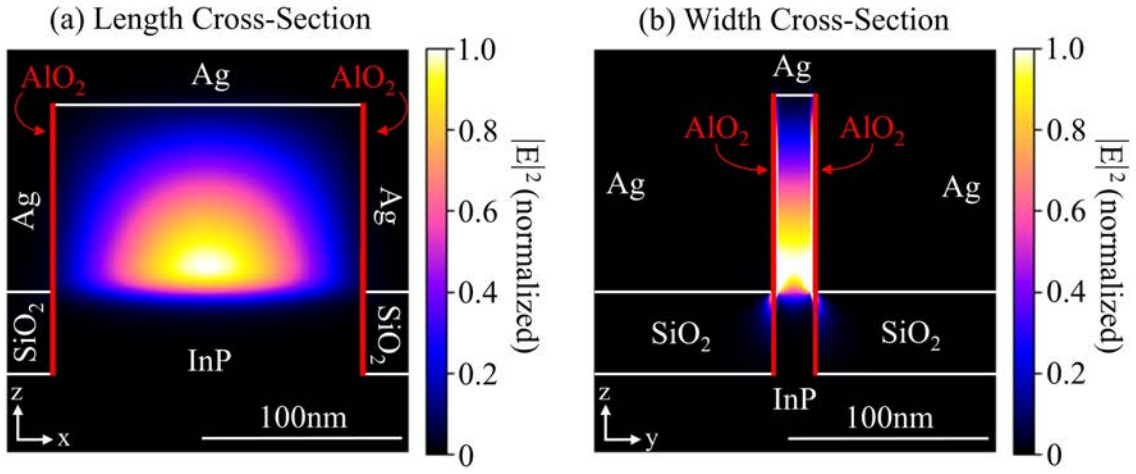

Figure 4.5: Normalized electric field intensity ($|E|^2$) within the cavity backed slot antenna along the length cross-section (a) and width cross-section (b).

incident plane wave[7]. Cross-sections of the electric field intensity ($|E|^2$ within the cavity-backed slot antenna from Fig. 4.4) are shown in Fig 4.5. The electric field intensity is normalized to peak, which occurs very near the antenna cavity opening. Interestingly, the electric field profile is approximately constant in the narrow width direction, but varies greatly in the length and height directions.

To obtain the maximum spatial average of the enhancement factor, we must target an InGaAs active region height that best overlaps with the antenna mode. This is depicted in Fig. 4.6 where we fix the width and length of the InGaAs (as these parameters would be difficult to control in fabrication), but we vary the InGaAs height which can be chosen during epitaxial growth. The average enhancement as a function of the InGaAs height can then be calculated using,

$$\text{Spatial Average} = \frac{1}{\text{Volume}} \iiint |E(x,y,z)|^2 dxdydz \tag{4.16}$$

where $|E|^2$ (assumed to be normalized to peak) varies with the 3D Cartesian directions x, y, and z (see Fig. 4.4 for the coordinate system marker). Volume represents the InGaAs volume, which is a function of the height parameter in Fig 4.6(a). The spatial average as a function of InGaAs height is given in Fig. 4.6(b). As the InGaAs height increases, the average enhancement tends to decrease because of poorer overlap with the peak of the antenna mode. However, we would like to maximize the height without sacrificing too much enhancement in order to increase the total output power of the LED (which is

[7]Note that the best way to calculate the spatial average of the enhancement factor would be to simulate individual dipole point sources scattered throughout the LED volume. This is because shining a plane wave on the structure will only excite radiative ("bright") modes in the antenna, whereas a dipole source can also excite non-radiative ("dark") modes. However, doing so is incredibly computationally expensive. Nevertheless, experimenting with both techniques for this antenna geometry provided adequate agreement. If a narrower antenna ridge were used, the general approach would likely be required.

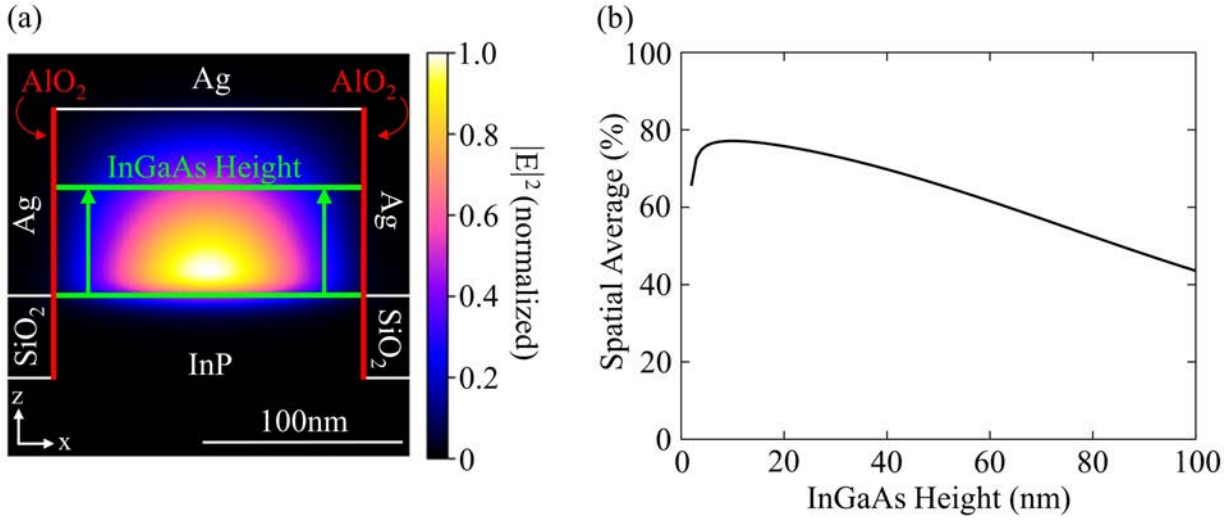

Figure 4.6: Spatial average of the enhancement factor depends on the volumetric overlap between the active semiconductor and the peak of the optical antenna mode. The electric field intensity along the length cross-section is repeated here in (a), but showing the height of the InGaAs active region relative to the cavity opening. The spatial average calculation in (b) shows that higher spatial averages can be obtained with shorter active region height, but at the cost of emitter volume.

proportional to the active volume). Therefore, we will choose an InGaAs height of 50nm, which gives:

$$\text{Spatial Average} = 65\% \tag{4.17}$$

Peak enhancement factor

Now that we know the location of the peak electric field in the antenna mode, we may simulate the peak enhancement factor, F_{peak} of a point dipole source. Fig. 4.7(a) once again shows the cross-section of the antenna mode, with the peak dipole source location explicitly noted. In particular, the peak dipole must be polarized in the narrow direction of the antenna, perpendicular to the long direction of the ridge that is shown. The resulting enhancement spectrum from this peak dipole source is shown in Fig. 4.7(b). The peak enhancement is then the scalar peak of the enhancement spectrum, with a value of

$$F_{\text{peak}} = 1280 \tag{4.18}$$

which occurs at a wavelength of $\lambda \approx 1580\text{nm}$.

Spectral average

Using the antenna enhancement spectrum in Fig. 4.7(b), we may now approximate the spectral average by finding the overlap with the intrinsic material spectrum of the InGaAs active region. The intrinsic InGaAs spontaneous emission spectrum of bulk InGaAs with an assumed doping concentration of

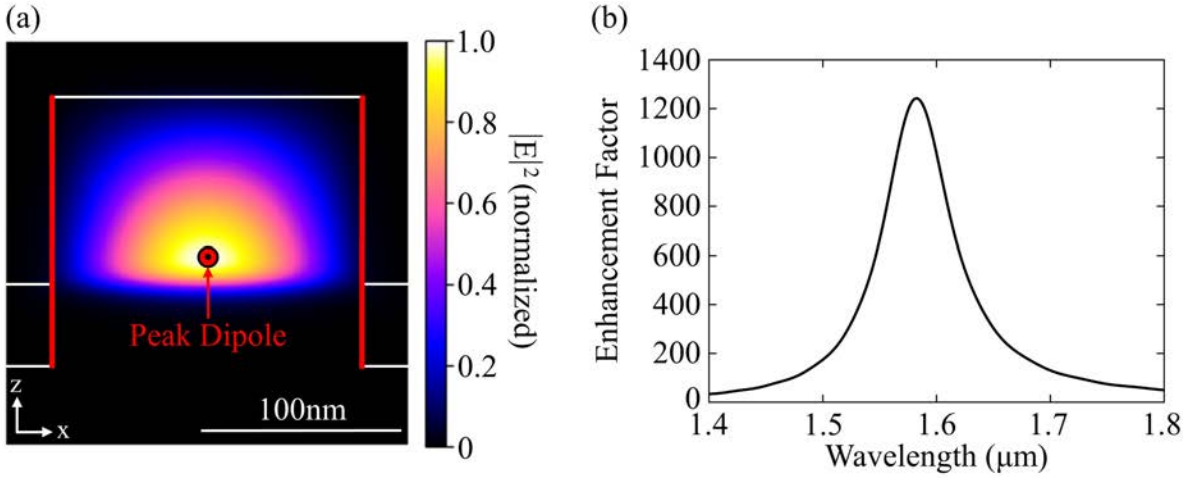

Figure 4.7: Peak enhancement occurs for a dipole source located 13nm above the opening of the cavity-backed slot antenna, polarized along the narrow antenna dimension. The peak dipole source location is shown in (a), with corresponding enhancement spectrum in (b).

$N_A = 2 \times 10^{19}\text{cm}^{-3}$ under two different biasing (pumping) conditions is shown in Fig. 4.8. These were simulated using the method provided in Appendix B. The two pumping conditions correspond to equilibrium minority carrier concentrations of $N = 10^{17}\text{cm}^{-3}$ (red) and $N = 10^{18}\text{cm}^{-3}$ (blue) respectively. The former case corresponds to low-level injection (nondegenerate conduction band occupation) while the latter case is high-level injection (degenerate conduction band occupation). Consequently, the large minority carrier concentration at $N = 10^{18}\text{cm}^{-3}$ causes bandfilling effects that are evident in the width of the spontaneous emission spectrum. By contrast, the $N = 10^{17}\text{cm}^{-3}$ case peaks very close to the bandgap wavelength of $\lambda_g = 1670\text{nm}$. Furthermore, the harder the device is pumped, the larger the peak spontaneous emission power per wavelength, as should be expected.

For our purposes, we will assume the semiconductor is pumped nondegenerately with $N \approx 10^{17}\,\text{cm}^{-3}$ or less. Under this condition, the peak optical power of spontaneous emission will increase with N, but the normalized spectral shape will be more-or-less fixed to that shown in Fig. 4.8. This will allow us to assume a fixed spectral average for our calculations, but it should be acknowledged that the spectral average will change (and in fact be variable) under high-level injection[8].

In Fig. 4.9 we plot the overlap of the spontaneous emission spectrum (corresponding to the $N = 10^{17}\text{cm}^{-3}$ and $P = 2 \times 10^{19}\text{cm}^{-3}$ case in Fig. 4.8) with the normalized antenna enhancement spectrum (from Fig. 4.7). We may obtain the spectral average by taking the weighted average of the enhancement

[8]In fact, under small-signal modulation conditions, the derivative of the spectral overlap with respect to the carrier concentration can produce an enhancement of the modulation bandwidth that the author of this thesis coined "differential enhancement" [48]. This effect has also been noted in prior work [126].

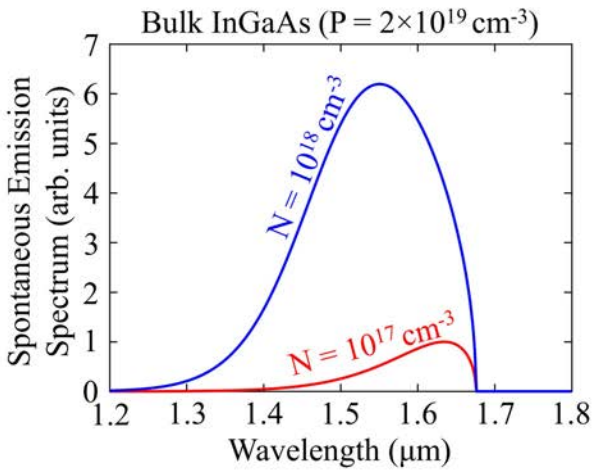

Figure 4.8: Spontaneous emission spectra of heavily-doped bulk InGaAs under two pumping conditions: nondegenerate minority carrier concentration, $N = 10^{17}\text{cm}^{-3}$, and degenerate minority carrier concentration, $N = 10^{18}\text{cm}^{-3}$.

factor with the spontaneous emission spectrum, defined as,

$$\text{Spectral Average} = \frac{1}{F_{\text{peak}}} \frac{\int F(\omega)L(\omega)d\omega}{\int L(\omega)d\omega} \tag{4.19}$$

where $L(\omega)$ is the intrinsic material spectrum of InGaAs, $F(\omega)$ is the antenna enhancement spectrum, and $F_{\text{peak}} = 1280$ is the peak of the antenna enhancement spectrum [9]. Using Eq. 4.19 we calculate:

$$\text{Spectral Average} = 49\% \tag{4.20}$$

Polarization average

Our final consideration in the total average enhancement factor seen by the LED concerns the antenna-LED polarization. As discussed in Appendix G and Refs. [21, 22], the matrix element for stimulated and spontaneous transitions of bulk active materials is typically assumed to be averaged over all three polarizations equally (resulting in a factor 1/3). However, quantum wells and quantum wires require a modified treatment, accounting for polarization selection rules within the confined geometry [21, 22]. This results in an effective polarization dependence in the relative transition strengths between electrons in the conduction band with holes in the light-hole, heavy-hole, and split-off valence bands.

[9] Note that the antenna enhancement spectrum from Fig. 4.7 has been slightly shifted to maximize the overlap integral with the spontaneous emission spectrum.

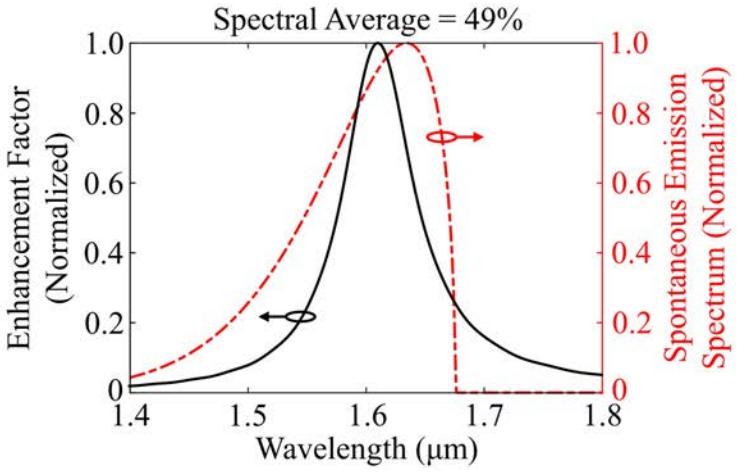

Figure 4.9: Spectral overlap of the antenna enhancement spectrum with the nondegenerate bulk InGaAs spontaneous emission spectrum from Fig. 4.8.

More specifically, the perturbation matrix element (due to the dipole transition potential induced by the incident wave) is given by $|H'_{21}|^2 = |q\vec{x}_{21} \cdot \hat{e}|^2$ where $q\vec{x_{21}}$ is the (vector) dipole moment of the electron-hole pair and $\hat{e}$ is the polarization of the incident light. When the dipole moment and light polarization are aligned, the matrix element value peaks at $|H'_{21}|^2 \equiv |qx_{21}|^2$ where $|qx_{21}|$ represents the conventional dipole moment amplitude. In a bulk crystal, the dipole moment directions are randomized, only aligning with the incident electric field 1/3 of the time on average. This factor 1/3 appears in the spontaneous emission lifetime of a two-level system from Eq. 3.33 in the previous chapter.

However, in a quantum well, the quantization of the k-vector changes the average dipole moment direction depending on the transition type and orientation of the electric field with respect to the confinement direction. For conduction band-to-heavy hole transitions, the dipole moment of electron-hole pairs are constrained to align perpendicular to the confinement direction (in the plane of the quantum well). By contrast, for conduction band-to-light hole transitions, the dipole moment lies (mostly) along the confinement direction on average. We may define a transition matrix element:

$$|H'_{21}|_i^2 = K_i |H'_{21}|^2 \tag{4.21}$$

where k_i is a relative transition strength defined for each polarization $i \in [\hat{x}, \hat{y}, \hat{z}]$, and $|H'_{21}|^2 = |qx_{21}|^2$ is the regular matrix element. These relative transition strengths are summarized in Table 4.1 for bulk and quantum well devices (reproduced from [22, 21]). Here, we assume that the quantum well confinement direction is along $\hat{z}$.

Using this information, we may now solve for the polarization average of the antenna enhancement for bulk and quantum well active region LEDs. The polarization average is a weighted average of the

	Transition Strength K_i, where $\lvert H'_{21}\rvert_i^2 = K_i \lvert H'_{21}\rvert^2$		
Polarization	Bulk (C-HH or C-LH)	Quantum Well (C-HH)	Quantum Well (C-LH)
$\hat{x}$	1/3	1/2	1/6
$\hat{y}$	1/3	1/2	1/6
$\hat{z}$	1/3	0	2/3

Table 4.1: Relative transition matrix element strengths for bulk and quantum well semiconductor crystals, reproduced from [22, 21]. C-HH and C-LH represent conduction band-to-heavy hole transitions and conduction band-to-light hole transitions, respectively.

enhancement factor and the transition matrix element element along each polarization, defined as:

$$\text{Polarization Average} = \frac{1}{F_{\text{peak}}} \frac{\sum_{i\in\hat{x},\hat{y},\hat{z}} F_i \lvert H'_{21}\rvert_i^2}{\sum_{i\in\hat{x},\hat{y},\hat{z}} \lvert H'_{21}\rvert_i^2} \tag{4.22}$$

where F_i refers to the antenna enhancement with corresponding polarization i (evaluated at the spectral and spatial peak of that polarization) in one of the three Cartesian directions, and $\lvert H_{21}\rvert_i^2$ refers to the relative transition matrix element in that same direction. Note that the denominator of Eq. 4.22 always evaluates to $\lvert H'_{21}\rvert^2$, the full matrix element, since the columns of Table 4.1 always sum to one.

In principle, one could design an antenna to have enhancement in each of the Cartesian directions. However, the most desirable configuration is like that of the dipole antenna or the cavity backed slot antenna, where one polarization is dominant. As indicated in Fig 4.7(b), the peak enhancement along the $\hat{y}$ polarization is very large with $F_{\text{peak}} = 1280$. Not shown are the enhancement factors along the other polarizations. These are provided in Appendix G, and are very small comparatively. Noting that the antenna mode is preferentially polarized in the $\hat{y}$ direction, we may then take $F_y = F_{\text{peak}}$ and $F_x = F_z = 0$. This allows us to write Eq. 4.22 as simply,

$$\text{Polarization Average} \approx \frac{\lvert H'_{21}\rvert_y^2}{\lvert H'_{21}\rvert^2} \equiv K_y \tag{4.23}$$

which is just the relative transition strength K_i in the $\hat{y}$ polarization from Table 4.1. Thus we find that for bulk transitions the Polarization Average is 1/3. By contrast, for quantum well C-HH transitions the Polarization Average is 1/2, while the C-LH transition is only 1/6. Interestingly, if we were to orient the quantum well confinement direction in the same direction as the antenna mode polarization, we could gain another polarization average boost of 2/3 for C-LH transitions – but this is a difficult configuration to achieve practically, and it would require strain. In our case, we are going to assume that the InGaAs active region is simply bulk because we will need the extra volume for optical power (as will be discussed later on)[10]. Nevertheless, it is worth remembering that an enhancement boost can be obtained by switching to quantum wells [36, 4]. Thus, to conclude, we will use:

$$\text{Polarization Average} = 33\% \tag{4.24}$$

[10]Note that the antenna geometry confines the ridge width to 20nm, so bulk is not an entirely errant assumption here. However, if one were to use a narrower LED ridge, these polarization selection rules should not be ignored.

Final average enhancement of the cavity-backed slot antenna and anticipated carrier lifetime

To summarize our progress thus far, we demonstrated the need to calculate an average enhancement factor F_{average} for antenna-LEDs with respect to the carrier lifetime of regular LEDs, which depends on the antenna mode polarization, spatial distribution, and spectral profile. For a bulk InGaAs active region LED ridge with volume=20nm×160nm×50nm doped to $P = 2 \times 10^{19}\text{cm}^{-3}$ under low-level injection conditions, we found spatial, spectral, and polarization averages of 65%, 49%, and 33% respectively, and a peak enhancement of $F_{\text{peak}} = 1280$. Thus, returning to Eq. 4.14 we find an average enhancement factor of,

$$F_{\text{average}} = 1280 \cdot 65\% \cdot 49\% \cdot 33\% \tag{4.25}$$

$$F_{\text{average}} = 135 \tag{4.26}$$

which is roughly a factor of 10 smaller than the peak enhancement factor, but still a massive rate enhancement value.

Under the assumption that the LED is heavily doped, the carrier lifetime of an unenhanced LED becomes the fundamental spontaneous emission lifetime of a two-level system. Thus, using the average enhancement value above we find a carrier lifetime by enhanced spontaneous emission of $\tau_{\text{sp}}^* \approx \tau_{\text{o}}/F_{\text{average}} = 1\text{ns}/135$, which is:

$$\tau_{\text{sp}}^* \approx 7\text{ps} \tag{4.27}$$

which is comparable to the stimulation emission lifetime of lasers given in the previous chapter of $\tau_{\text{st}} = 6\text{ps}$. Furthermore, in this analysis we made a number of conservative assumptions for the practical design of a cavity-backed slot antenna-LED. The average enhancement factor could be increased by a number of methods, for instance by using a shorter InGaAs active region consisting of quantum well heterostructures (for larger spatial and polarization averages). Or, the simplest method to increase the average enhancement would be slightly decreasing the LED ridge width d to increase the peak enhancement (at a potential cost of efficiency, which will be discussed in the next chapter). Therefore, the radiative recombination lifetime given in Eq. 4.27 should be considered a conservative estimate of what is possible, and we can conclude that spontaneous emission can be at least as fast as stimulated emission, if not faster.

4.3 Modulation dynamics of antenna-LEDs and lasers

In the previous section we demonstrated that optical antenna-enhanced spontaneous emission can be as fast as stimulated emission. But, does that actually mean LEDs can be as fast as lasers? More precisely, how does the carrier lifetime due to spontaneous emission and stimulated emission actually relate to device speed? In this section we will argue that carrier lifetime is the most important parameter that determines device speed for low-power on-chip optical interconnects applications. In particular, we will argue that large-signal direct modulation should be employed, which has different dynamics than conventional small-signal modulation. We will conclude with a brief discussion of novel laser cavities employing photon-photon resonance [145].

Optical modulation methods

There are two main ways to perform modulation of an optical signal: external and direct. As the name implies, direct modulation corresponds to modulating the injected current within the transmitter itself. External modulation assumes that the optical source is operating in the continuous-wave (CW) mode, and the optical signal is modulated by some external device.

The two major types of integrated external modulators provide either amplitude modulation or phase modulation through the electro-absorption and electro-optic effects respectively. In this thesis, we argue that external modulators should be avoided in on-chip optical interconnects in order to guarantee the smallest energy/bit operation. Our arguments against external modulation are as follows:

1. The electro-optic effect is weak. At least one analysis suggests that to achieve the necessary change in refractive index for modulation, the device must be long (>100μm) and therefore potentially lossy and slow [92]. These detriments rule out electro-optic devices.

2. Electro-absorption effects can be fast, but incur extra losses along the photonic link because of absorption.

To expand upon the second point, we will note that what actually matters at the receiver end of a photonic link is the difference between the received optical power in the off and on state, $\Delta P = P_{\text{on}} - P_{\text{off}}$. Electro-absorption devices modulate the transmission ($\Delta T = T_{\text{on}} - T_{\text{off}}$) of a CW optical signal with constant power P. Therefore, $\Delta P = P\Delta T$ in such a transmitter. To get adequate peak-to-peak signal at the receiver, one must either increase the transmission contrast ΔT or increase the optical power. In the very best case, $\Delta T = 1$, we will have $P_{\text{on}} = P$ and $P_{\text{off}} = 0$, which gives the largest power difference $\Delta P = P$. But this implies the best link efficiency that could be achieved is only 50% on average because all generated optical power is wasted in the off state or when the transmitter is idling. Note that more realistic configurations where $\Delta T < 1$ would imply even worse efficiency, because to maintain a constant ΔP for adequate signal, we must also increase the CW optical power by the factor $P = \Delta P/\Delta T$, meaning that even more optical power is wasted. In most cases (for practical device lengths and speed), the contrast will not be perfect and therefore external modulation guarantees $< 50\%$ quantum efficiency[11].

Direct electrical modulation

To maximize the efficiency of the photonic link, we will compare laser and LED direct electrical modulation. The limitation of direct modulation is, of course, that the speed of the device is limited by how quickly it can be directly modulated. The small-signal model is the conventional way of characterizing the modulation speed of lasers and LEDs. However, in this thesis we will argue that the small-signal assumption is insufficient to describe the dynamics of direct modulation for on-chip optical interconnects. First, we will briefly describe the small-signal modulation speed of lasers and LEDs. Then, we will argue why it does not apply to on-chip optical communications. Finally, we will show the limiting modulation behavior of lasers and antenna-LEDs under large-signal modulation conditions. We will find that both devices are limited by a characteristic turn-off time, which is related to the device carrier lifetime.

[11] Not to mention, changing the state of the external modulator requires energy.

Small-signal modulation

A derivation of the small-signal modulation rate of antenna-LEDs is provided in Appendix E.1 and a derivation of the small-signal modulation rate of lasers may be found in Ref. [22]. A quantity called f_{3dB} can be thought of as a limit on the small-signal modulation bandwidth. In particular, it denotes the frequency when the laser or antenna-LED is unable to respond to the oscillations in current. The 3dB frequencies of antenna-LEDs and lasers are given in Eq. 4.28 and Eq. 4.29 below,

$$f_{\text{3dB, antenna-LED}} = \frac{\sqrt{3}}{2\pi}\left(\frac{1}{\tau_{\text{sp}}^*} + \frac{1}{\tau_{\text{nr}}}\right) \approx \frac{0.28}{\tau_{\text{sp}}^*} \tag{4.28}$$

$$f_{\text{3dB, laser}} = \frac{\sqrt{1+\sqrt{2}}}{2\pi}\sqrt{\frac{1}{\tau_{\text{st}}\tau_p}} \approx \frac{0.25}{\sqrt{\tau_{\text{st}}\tau_p}} \tag{4.29}$$

where τ_{sp}^* is the (enhanced) spontaneous emission carrier lifetime, τ_{nr} is the carrier lifetime due to (parasitic) non-radiative recombination, τ_{st} is the stimulated emission carrier lifetime, and τ_p is called the photon lifetime of the laser cavity. In the second equality of Eq. 4.28 we simplify the pre-factor and ignore τ_{nr} because a fast non-radiative lifetime implies inefficiency [12]. In the second equality of Eq. 4.29 we simplify the pre-factor, revealing that the laser f_{3dB} is given by the geometric mean of τ_{st} and τ_p. The photon lifetime refers to the rate of photon loss from the optical cavity by either parasitic (intrinsic absorption or scattering) or desired (mirror transmission) mechanisms. It may be written:

$$\frac{1}{\tau_p} = v_g\alpha = \Gamma v_g g_{\text{th}} \tag{4.30}$$

where v_g is the group velocity, α is the loss (1/cm) coefficient, Γ is the confinement factor and g_{th} is the threshold gain. Evidently one observes that – under the small-signal modulation assumption – lasers can be much faster than the rate implied the carrier lifetime due to stimulated emission, τ_{st}, simply by reducing the photon lifetime, τ_p. This can easily be achieved in a non-parasitic manor by decreasing the reflectivity of the laser facet [13].

This can be illustrated by plugging-in the known values for our antenna-LED example and the cleaved-facet laser from the previous chapter. For simplicity we will assume that both the antenna-LED and laser have the same carrier lifetime due to spontaneous emission or stimulated emission respectively, $\tau_{\text{sp}}^* = \tau_{\text{st}} = 6\text{ps}$. Then, assuming $g_{\text{th}} = 500$ 1/cm, $v_g = 10^{10}$cm/s, and $\Gamma = 0.2$ we retrieve a photon lifetime of $\tau_p = 1\text{ps}$. Plugging in τ_{sp}^*, τ_{st}, and τ_p to Eq. 4.28 and Eq. 4.29 above we find that:

$$f_{\text{3dB,antenna-LED}} = 47\text{GHz} \tag{4.31}$$

$$f_{\text{3dB,laser}} = 100\text{GHz} \tag{4.32}$$

[12] Nevertheless, it is worth keeping in mind that the LED small-signal f_{3dB} can be increased simply by increasing loss

[13] The photon lifetime and stimulated emission lifetimes are not completely decoupled. For constant input current, as the photon lifetime is decreased, the stimulated emission lifetime will increase accordingly since the photon density in the cavity is reduced. Therefore, to get a speed boost one must compensate by pumping the laser harder. This will inevitably lead to heat management issues or saturation effects, but in this thesis we regard these problems as non-fundamental engineering issues. Conservatively or charitably speaking, the small-signal modulated laser is not limited by the stimulated emission lifetime.

In other words, the laser can be twice as fast as the LED even though the carrier lifetimes are exactly the same. Note that the $f_{3\text{dB}}$ for the laser found here is very optimistic. Nevertheless, an $f_{3\text{dB}}$ of 60GHz was achieved recently by the intentional reduction of the photon lifetime and improved thermal management [145].

We conclude that LEDs cannot be as fast as lasers in the small-signal modulation case. Unsurprisingly, for this and other reasons[14], lasers should be used for long-haul telecom and datacom. However, on the chip-scale, we will argue that LEDs potentially have an edge.

Large-signal modulation

On the chip-level datacom scale, the incentives change. Indeed, in the ultimate limit of scaling data communications to groups of individual logic devices, we will require a dense distribution of nanoscale light sources. Ignoring problems that arise when engineering lasers at this scale (which we will consider non-fundamental), as we reduce the active volume of the transmitters we also reduce the optical peak-to-peak power that we are capable of producing for adequate signal at the receiver. In other words, to maintain the minimum ΔP needed at the receiver, then the ratio of the modulated optical power to the CW power of a laser increases,

$$\frac{\Delta P_{\text{minimum}}}{P} \propto \frac{\Delta P_{\text{minimum}}}{V_{\text{active}}} \uparrow \text{ with } V_{\text{active}} \downarrow \tag{4.33}$$

In other words, the small-signal assumption $\Delta P \ll P$ becomes invalid as the device active volume decreases. Furthermore, as the number of transmitters on the chip multiply it becomes undesirable from an energy efficiency standpoint to maintain a large DC bias above threshold. Indeed, to maintain a device at a nominal CW operating power to increase the small-signal speed (of a laser), we are constantly incurring wasted CW optical power while idling, not to mention I^2R Ohmic losses occurring outside of the laser.

Therefore for the two reasons of (1) small-signal assumption no longer being valid for smaller footprint optical transmitters and (2) needing to avoid wasted photons while idling, we argue that nanoscale transmitters must be operated in a large-signal modulation format. But, how does the speed of lasers and LEDs change under large-signal modulation? Tucker [133] argued that under large-signal modulation, laser speed can be determined by the sum of two approximately independent times,

$$f_{\text{large-signal}} \propto \frac{1}{t_{\text{on}} + t_{\text{off}}} \tag{4.34}$$

where t_{on} and t_{off} refer to "turn-on" and "turn-off" times respectively. In his analysis the turn-on time is essentially given by the laser relaxation resonance frequency (which is the small-signal limited rate), but the turn-off time is limited by the carrier lifetime due to stimulated emission (as measured at peak

[14] Where we are not even considering optical power for signal-to-noise considerations, quantum efficiency, advanced modulation formats, coherent communications, etc.

signal)[15]. In other words,

$$t_{\text{on}} > \sqrt{\tau_p \tau_{\text{st}}} \tag{4.35}$$

$$t_{\text{off}} > \tau_{\text{st}} \tag{4.36}$$

As we showed previously, $\tau_p \ll \tau_{\text{st}}$ and therefore the laser is limited by t_{off} which is, in turn, limited by the carrier lifetime due to stimulated emission!

To illustrate this point further, we simulate the large-signal pulse response of an antenna-LED and laser side-by-side. This is provided in Fig. 4.10, where we show the input current pulse and optical responses of a laser with gain saturation, a laser without gain saturation, an antenna-LED, and the overlapped optical responses in Fig. 4.10(a)-(e) respectively. An explanation of the rate equation model used to simulate these pulse responses can be found in Appendix E and Appendix B.

The laser optical responses in (a) and (b) correspond to the cleaved-facet edge-emitting quantum well InGaAs laser from Fig. 3.4 in the previous chapter. This laser has a photon lifetime of $\tau_p = 1$ps. Both cases assume a stimulated emission recombination rate with gain saturation, but the laser in (b) has a very large saturation photon density ($S_{sat} = 3.5 \times 10^{17}\text{cm}^{-3}$ while the laser in (c) has the saturation photon density provided in the previous chapter ($S_{sat} = 3.5 \times 10^{16}\text{cm}^{-3}$). Thus, the laser without gain saturation (very small saturation), can be seen to ring immediately after the current pulse, while the laser with gain saturation is heavily damped. The ringing for the laser with large gain saturation can be seen if the laser is pumped to a smaller peak optical power, since the damping effect is less pronounced. The laser without gain saturation in (b) is pumped to a photon density such that its stimulated emission time is $\tau_{\text{st}} = 6$ps in the on-state. The laser without gain saturation in (c) is pumped with the same current[16]. Note that both lasers are taken to have a very small threshold current, but in the off-state they are pumped slightly below threshold. The antenna-LED in (d) is taken to have an enhanced spontaneous emission carrier lifetime of $\tau_{\text{sp}} = 6$ps, and we have ignored any additional non-radiative recombination. This corresponds to approximately 166$\times$ average enhancement over the nominal spontaneous emission lifetime of heavily doped InGaAs, which is a reasonable value as suggested by our analysis in the previous section for the cavity-backed slot antenna. In contrast with the two laser responses, the antenna-LED does not ring in response to the current pulse.

Of interest in Figs. 4.10(b)-(d) are the on-time t_{on} and the off-time t_{off} transients. The on-time is defined here as the time from the onset of current to when the laser or antenna-LED has reached 90% of its on-state power. The on-time for the laser responses (b) and (c) lag behind the incident current pulse by approximately 21ps. This is not fundamental, the on-time can be reduced significantly by biasing the laser above threshold in the off-state. We considered a below threshold off-current because DC biasing

[15] This is a slight simplification. Tucker's [133] analysis showed that the turn-on time is related to both the laser relaxation frequency and the ratio of the on vs. off laser power, and the turn-off time is related to the relaxation frequency plus a term that we have interpreted so far as the carrier lifetime due to stimulated emission in the laser. The laser on/off power term can be minimized if the laser is biased near threshold in the off state, so we have ignored it.

[16] However, the laser with gain saturation cannot quite reach the stimulated emission lifetime of 6ps. This is because gain saturation effectively reduces the differential gain coefficient at large internal photon density. The laser would need to be pumped to unrealistic current levels to reach the same stimulated emission lifetime. Thus, we elected to use the same input current, which corresponds to roughly the same CW optical power.

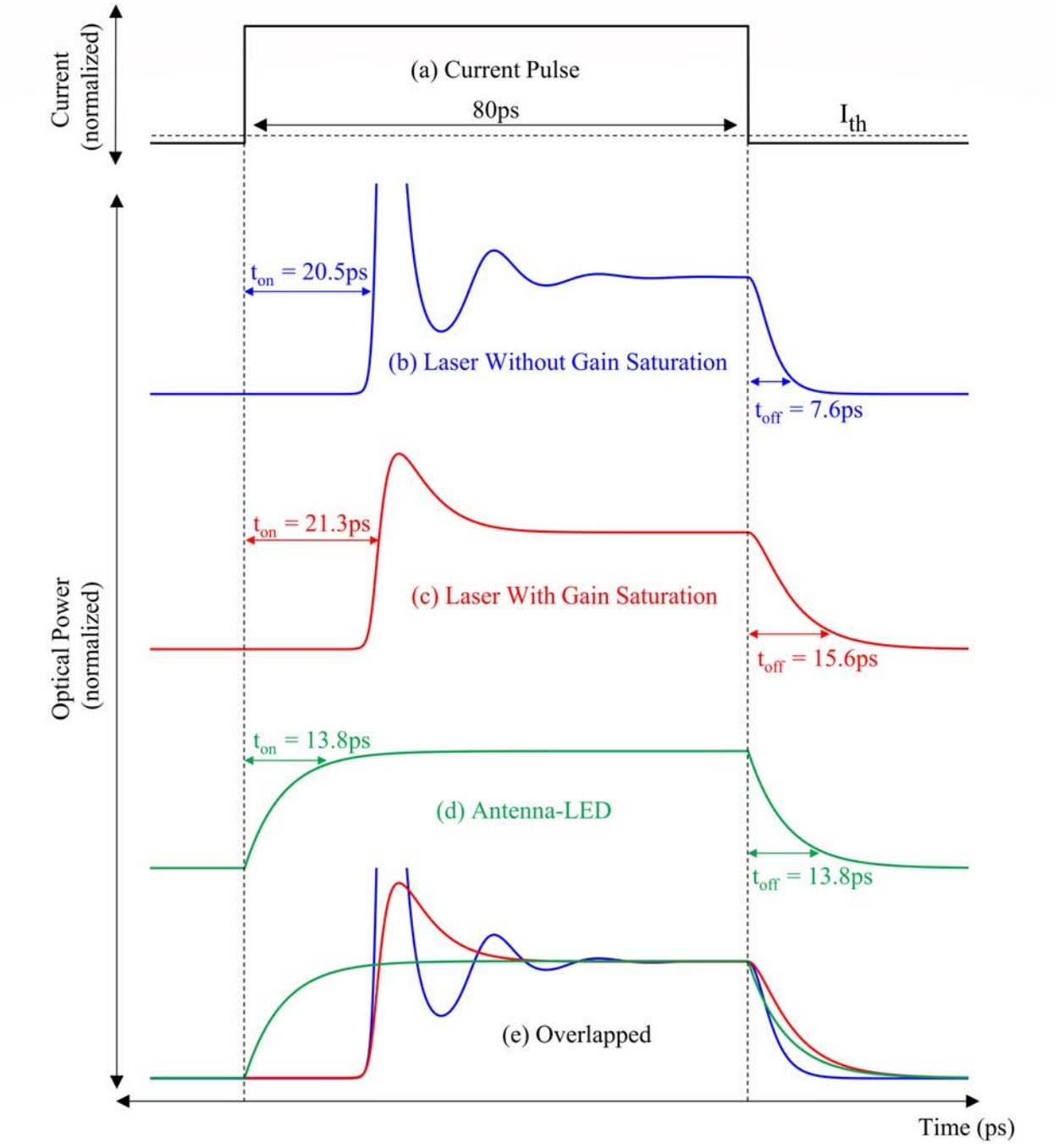

Figure 4.10: Large-signal pulse response of lasers with and without gain saturation and an antenna-LED. All three devices are limited by a characteristic off-time, which is determined by the respective carrier lifetime of each device. Note that the on-time found here is non-fundamental because it depends on DC bias and the current pulse amplitude.

nanoscale lasers above threshold could be a large source of inefficiency in on-chip transmitters. Nevertheless, in principle the on-time is only limited by the relaxation frequency of the laser which is determined approximately by the photon lifetime τ_p. The antenna-LED has a slightly shorter on-time of 13.8ps. Note that this on-time is also not fundamental, it can be shortened by using a current pulse with higher peak current, which will be demonstrated in the next section. Thus, the on-times for both the antenna-LED and lasers should be considered non-limiting for speed considerations. Note that we have not included additional confounding effects such as thermalization time (energy relaxation) for carriers injected from the contacts, "capture time" in quantum wells [60], nor hole-burning [148]. These mechanisms would likely constitute the ultimate on-time limit for these devices, with characteristic times of $\tau = 1$ps order of magnitude.

The more fundamental transient for all three optical responses is the off-time, t_{off}. The off-time is defined as the time required for the optical response to decay to 10% of its on-state value after the current pulse has been switched off. The off-times of the two laser responses differ dramatically with $t_{\text{off}} = 7.6$ps and $t_{\text{off}} = 15.6$ps for the cases without and with gain saturation respectively. For both cases t_{off} is greater than the carrier lifetime due to stimulated emission $\approx$ 6ps which was originally predicted by Tucker [133]. This can be understood intuitively in terms of the laser physics at play. When the laser current is turned off, light quickly leaves the laser facet because of the small photon lifetime. However, as light leaves the optical cavity the stimulated emission rate slows down significantly because the recombination rate $R_{\text{st}} = v_g g S$ is reduced. Then, as the laser continues to decay the carrier density decreases from its threshold value, reducing both the optical gain and the stimulated emission recombination rate further. Thus, the decay transient is largely limited by the characteristic lifetime of stimulated emission at the peak of the pulse, τ_{st} (plus an additional factor given by the laser resonance frequency, according to Tucker [133], which we are ignoring). By contrast, the antenna-LED decay is purely exponential and given by the carrier lifetime due to enhanced spontaneous emission, τ_{sp}^*. In fact, it can be easily found that $t_{\text{off}} \approx ln(10) \cdot \tau_{\text{sp}}^* \approx 2.3 \cdot 6\text{ps} = 13.8\text{ps}$ where $\ln(10)$ accounts for the 90% exponential decay time. Note that it is only coincidence that the on-time is equal to the off-time in this case. If we were to use a higher-current pulse, the on-time would be faster but the off-time would always be limited by the carrier lifetime of spontaneous emission. This latter fact will be shown in the next section.

All three cases are compared in Fig. 4.10(e) where the optical response transients are overlapped. In particular, we see that the antenna-LED off-transient lies somewhere between the two laser cases. Lasers limited by gain saturation (e.g. a strongly-pumped nanolaser) have slightly slower off-time than antenna-LEDs, while the laser not limited by gain saturation can be somewhat faster but is still limited by the carrier lifetime of stimulated emission [17]. We leave it as an experimental question whether the off-transient of real nanolasers follows the gain saturation case or not. Nevertheless, we can conclude that since both lasers and antenna-LEDs are limited by their respective carrier lifetimes of light-emission, antenna-LEDs can be as fast as lasers.

[17] Note that we have not considered gain switching in this analysis, which can potentially be used to improve laser large-signal modulation.

A brief comment on photon-photon resonance

A number of recent works have emphasized photon-photon resonance (PPR) for directly-modulated lasers [84, 145]. Photon-photon resonance is a small-signal phenomenon that is similar to injection locking, but requires no external laser. In particular, the small-signal modulation bandwidth can be enhanced by approximately the free-spectral range of a passive external Fabry-Perot cavity that provides optical feedback to the active laser cavity. In this work, we will assume that photon-photon resonance will not be useful for on-chip optical communications, because we have assumed that nanoscale optical sources will require large-signal modulation. Furthermore, photon-photon resonance is very sensitive to the phase and amplitude of optical feedback [18]. It remains to be seen whether PPR can be reliably engineered, especially on a chip-wide scale.

4.4 Optical interconnects with antenna-LEDs: system analysis and remaining challenges

Up to this point we have provided a detailed argument of the ultimate speed of light-emitting diodes in comparison to lasers. We showed that heavily-doped antenna-enhanced light-emitting diodes can have a spontaneous emission carrier lifetime that is as fast as the stimulated emission carrier lifetime in conventional index-guided lasers. Consequently, under large-signal direct electrical modulation – which we argued should be employed for nanoscale on-chip optical transmitters – antenna-LEDs can be as fast as lasers. However, speed is not the only important metric. In this section we will discuss two additional metrics, efficiency and optical power, then conclude with the final requirements and remaining challenges for the use of an antenna-LED as an on-chip optical transmitter.

Antenna-LED efficiency

There are three efficiencies that contribute to the total quantum efficiency of an optical antenna-LED in a photonic link: internal quantum efficiency, antenna efficiency, and waveguide coupling efficiency[19]. Antenna efficiency and waveguide coupling efficiency are purely electromagnetic properties, and can be calculated in Maxwell simulation. The antenna efficiency is the ratio of optical power radiated into the far field versus total power generated by the LED, $P_{\text{radiated}}/P_{\text{total}}$. More specifically, it takes into account Ohmic loss effects that occur mostly within the metal comprising the antenna. By contrast, the waveguide coupling efficiency is the waveguide mode-match ratio, $P_{\text{mode-match}}/P_{\text{radiated}}$, which takes into account mode-matching but not Ohmic loss. Thus, the product of the antenna efficiency and the waveguide coupling efficiency provides the ratio of power that is mode-matched to a waveguide divided by the total power generated by the antenna-LED source. These two efficiencies will be discussed in much greater detail in the next chapter. For now, we provide reasonable values of these efficiencies that will be used in our overall system analysis of the antenna-LED: Antenna Efficiency $\approx 70\%$ and Waveguide Coupling Efficiency $\approx 90\%$.

[18] This was tested by simulation, but not included in this thesis for brevity.

[19] We are ignoring "injection efficiency" in this analysis, assuming that an adequate heterostructure can be engineered.

On the other hand, the internal quantum efficiency takes into account the recombination lifetimes within the semiconductor source. For LEDs operating at low-level injection, this is defined as,

$$\text{Internal Quantum Efficiency} = \frac{\dfrac{1}{\tau_{sp}^*}}{\dfrac{1}{\tau_{sp}^*} + \dfrac{1}{\tau_{nr}}} \tag{4.37}$$

where τ_{sp}^* is the carrier lifetime due to enhanced spontaneous emission and τ_{nr} is the carrier lifetime due to non-radiative recombination. An enlightening re-framing of Eq. 4.37 is to write the carrier lifetime due to spontaneous emission in terms of the average enhancement factor: $1/\tau_{sp}^* = F_{average}/\tau_{sp}$ where τ_{sp} is the spontaneous emission carrier lifetime in an intrinsic bulk semiconductor. Then Eq. 4.37 becomes,

$$\text{Internal Quantum Efficiency} = \frac{\dfrac{F_{average}}{\tau_{sp}}}{\dfrac{F_{average}}{\tau_{sp}} + \dfrac{1}{\tau_{nr}}} \tag{4.38}$$

indicating that the internal quantum efficiency increases to unity with large antenna enhancement factor. This is one of the most exciting effects of spontaneous emission enhancement, because enhancement makes the LED brighter by increasing both its emission rate and quantum efficiency simultaneously.

However, in order to achieve large enhancement, generally very small LED dimensions are required. For example, for the cavity-backed slot antenna-LED discussed in Fig. 4.4, the smallest LED dimension was just 20nm. This can greatly increase non-radiative recombination effects within the LED. Typically non-radiative recombination is modeled with two empirical coefficients in the ABC recombination model:

$$\frac{1}{\tau_{nr}} \approx A + CP^2 \tag{4.39}$$

where A (in units of 1/s) models Shockley-Read-Hall (SRH) recombination and surface recombination, while C (in units of cm^6/s) is the Auger coefficient (where we have implicitly assumed that the LED is P-doped with a hole majority carrier concentration of $P \approx N_A$). Typically, the empirical Auger coefficient for InGaAs is in the range $C = 10^{-28}\text{cm}^6/\text{s} - 10^{-30}\text{cm}^6/\text{s}$. Assuming a worst case, $C = 10^{-28}\text{cm}^6/s$, and assuming a large P-doping concentration of $P = 2 \times 10^{19}\text{cm}^{-3}$ to maximize the spontaneous emission lifetime, then we find that the non-radiative lifetime due to Auger recombination is given by $1/CP^2 \approx 25\text{ps}$. Previously we found that the carrier lifetime of enhanced spontaneous emission in antenna-LEDs can be as fast as 6ps, and therefore Auger recombination does not significantly limit the internal quantum efficiency (unless we were to use extremely large doping concentrations $P > 2 \times 10^{19}\text{cm}^{-3}$).

The most critical contribution to non-radiative recombination in antenna-LEDs is surface recombination. Surface recombination is typically modeled using a surface recombination velocity term such that [22]:

$$A \approx \text{Surface to Volume Ratio} \times \text{Surface Recombination Velocity} \tag{4.40}$$

where the surface to volume ratio depends on the geometry of the LED ridge, and the surface recombination velocity is an empirical coefficient indicating the severity of surface recombination effects. In the cavity-backed slot antenna-LED from Fig. 4.4, the dimensions of the LED ridge are given by width $\times$ length $\times$ height $= (20\text{nm}) \times (160\text{nm}) \times (100\text{nm})$. Noting that width $\ll \{\text{height}, \text{length}\}$, the surface to volume ratio is given by:

$$\frac{\text{Surface Area}}{\text{Volume}} = \frac{2 \cdot \text{length} \cdot \text{height} + 2 \cdot \text{height} \cdot \text{width}}{\text{width} \cdot \text{length} \cdot \text{height}} \approx \frac{2}{\text{width}} = 10^6 \frac{1}{\text{cm}} \tag{4.41}$$

where we simplified the expression in the third equality by ignoring the much longer length and height dimensions that will have negligible contribution to the surface-to-volume ratio. Thus, Eq. 4.40 becomes:

$$A \approx \frac{2}{\text{width}} \cdot \text{Surface Recombination Velocity} \equiv \frac{2}{d}\text{SRV} \tag{4.42}$$

where d is the LED width, and SRV is the surface recombination velocity. In general, surface recombination velocity is highly contingent upon processing, and fabricating extremely narrow LED ridges can be quite harsh to the material. Without using any special techniques to treat the LED surface, typical surface recombination velocity values are in the 10^5cm/s range or worse. Therefore, the non-radiative carrier lifetime due to surface recombination can be as fast as $\tau_{SR} < 1/(10^6 \frac{1}{\text{cm}} \cdot 10^5 \frac{\text{cm}}{\text{s}}) = 10\text{ps}$ which is approaching or exceeding the carrier lifetime of enhanced spontaneous emission. This is a severe effect that will need to be addressed for antenna-LEDs to be viable in optical interconnects.

Fortunately, there has been recent progress in the surface passivation of nanoscale LEDs [36]. A detailed discussion of these processes is out-of-scope for this thesis. However, surface recombination velocity SRV $< 10^4$cm/s was achieved by Fortuna [36]. More recently, unpublished work by Andrade et al has claimed SRV < 100cm/s in InGaAsP/InP ridges. Applying these surface treatment processes to antenna-LED ridges with the dimensions of interest ($d \approx 20$nm) would represent a significant achievement, lending credence to the feasibility of these devices.

We may now summarize and conclude our discussion of internal quantum efficiency in the antenna-LED. After taking into account Auger recombination and surface recombination (Eq. 4.42) we may write Eq. 4.38 as:

$$\text{Internal Quantum Efficiency} = \frac{\dfrac{F_{\text{average}}}{\tau_{\text{sp}}}}{\dfrac{F_{\text{average}}}{\tau_{\text{sp}}} + \dfrac{2}{d}\text{SRV} + CP^2} \tag{4.43}$$

We note that in the previous chapter we found that the carrier lifetime of spontaneous emission in intrinsic bulk semiconductors saturates to $\tau_{\text{sp}} \to \tau_{\text{o}}$ in heavily-doped LEDs. But, in general τ_{sp} is a function of doping concentration, taking the empirical value $1/B_{\text{o}}P$ at small doping. Therefore, we may solve for the internal quantum efficiency (Eq. 4.43) as a function of doping concentration. The result of this rigorous calculation (using the active region model developed in Appendix B), is provided in Fig. 4.11. For this calculation we assumed $d = 20$nm, $C = 10^{-28}\text{cm}^6/\text{s}$, and $F_{\text{average}} = 166$ (which corresponds to

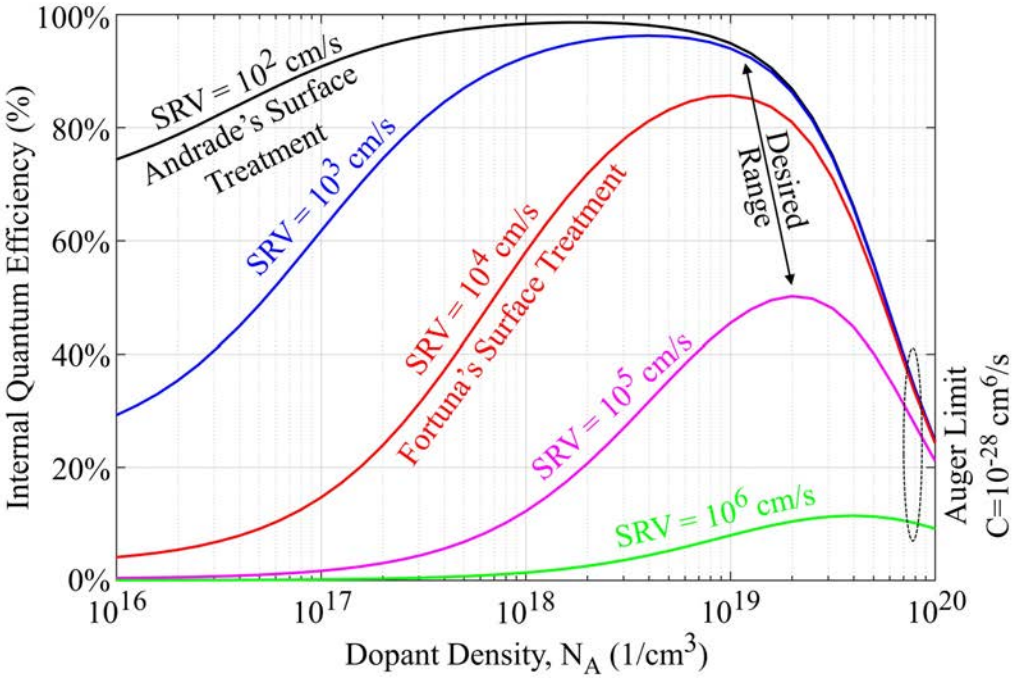

Figure 4.11: Internal quantum efficient of the antenna-LED as a function of dopant density and surface recombination velocity. High-speed and efficient emission can be obtained with heavy-doping and improved III-V surface treatment.

the average enhancement needed for a carrier lifetime due to spontaneous emission of 6ps). In addition, several curves are shown representing various possible values of the surface recombination velocity, SRV. In particular, Fortuna's surface treatment [36] is represented by SRV $= 10^4$cm/s and Andrade's surface treatment is represented by SRV $= 10^2$cm/s. We find that internal quantum efficiency exceeding 50% can be achieved with SRV $= 10^4$cm/s assuming reasonable doping concentrations $P > 10^{19}\text{cm}^{-3}$.

Antenna-LED continuous-wave optical power

Using the total antenna-LED quantum efficiency, we may now estimate the optical power transmitted in a photonic link. For now, let us consider the steady-state CW optical power with a constant current source:

$$P_{\text{optical}} = \text{Efficiency} \cdot \frac{\hbar\omega}{q} I \tag{4.44}$$

where I is the CW current and,

$$\text{Efficiency} = \text{Waveguide Coupling Eff.} \cdot \text{Antenna Eff.} \cdot \text{Internal Quantum Eff.} \tag{4.45}$$

$$\text{Efficiency} \approx 90\% \cdot 70\% \cdot 80\% = 50\% \tag{4.46}$$

where we inserted each of the relevant efficiency terms to find a total efficiency (sometimes called "wall-plug" efficiency or "external quantum efficiency") of 50%. We will justify the values of the waveguide coupling efficiency and antenna efficiency in the next chapter. Furthermore, we assumed an optimistic internal quantum efficiency of 80% assuming that III-V surface treatment will eventually mature for nanoscale ridges of this dimension.

What is a reasonable amount of current for an antenna-LED? In steady-state, we may write the current in terms of the recombination rate due to spontaneous emission, N/τ_{sp}^*, and the internal quantum efficiency, η_{IQE}:

$$\frac{N}{\tau_{sp}^*} = \eta_{IQE}\frac{I}{qV} \tag{4.47}$$

where N is the steady-state minority carrier concentration, $\tau_{sp}^* = 6\text{ps}$ is the enhanced spontaneous emission carrier lifetime assuming an average enhancement of 166 and heavy doping, and V is the antenna-LED active region volume. A reasonable upper bound to the minority carrier concentration of the antenna-LED is $N \approx 5\times10^{18}\text{cm}^{-3}$ where high-level injection effects are not severe. Thus, rearranging Eq. 4.47 for current, we have:

$$I = \frac{qV}{\eta_{IQE}}\frac{N}{\tau_{sp}^*} = \frac{q\cdot(160\text{nm}\cdot 50\text{nm}\cdot 20\text{nm})}{80\%}\frac{5\times10^{18}\text{cm}^{-3}}{6\text{ps}} = 27\mu\text{A} \tag{4.48}$$

and then plugging this current into Eq. 4.44 above,

$$P_{\text{optical}} = \text{Efficiency}\cdot\frac{\hbar\omega}{q}I = 50\%\cdot 0.8\frac{\text{W}}{\text{A}}\cdot 27\mu\text{A} \tag{4.49}$$

$$P_{\text{optical}} = 11\mu\text{W} \tag{4.50}$$

where we took $\hbar\omega = 0.8\text{eV}$. This is a relatively small amount of power compared to what could be possible with micro-lasers with much larger active volume. Unfortunately, this represents an approximate maximum on the optical power from antenna-LEDs. The most naively obvious way to increase the antenna-LED power would be to increase the active region volume, but as we found earlier on in the chapter this will result in a trade-off with the average enhancement factor thereby reducing the antenna-LED speed, internal quantum efficiency, as well as optical power.

Furthermore, pumping the device with more current is not likely a solution. The current that we found in Eq 4.48 is a large value for a device of this volume, pumping it any harder would lead to parasitic effects such as enhanced Auger recombination, velocity overshoot (which lowers the injection efficiency), increased Ohmic loss, and band-filling. Note that antenna-LED self-heating effects could potentially be severe at our assumed current value based on a consideration of the current density injected in the device. However, a simple heat-equation analysis suggests that self-heating is negligible because of the presence of the metallic antenna which acts as a large thermal reservoir for heat conduction [20].

[20] This analysis is not provided here for brevity, but it indicates approximately 1 kelvin temperature rise for 10μA of current in the steady-state.

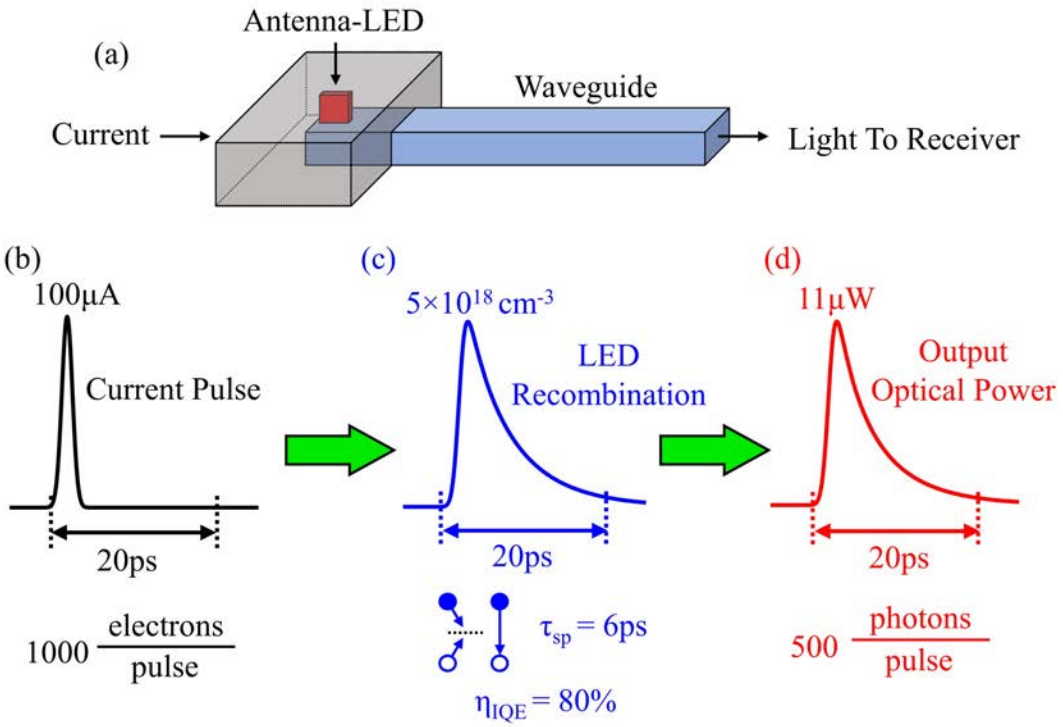

Figure 4.12: Full transient model of antenna-LED transmitter. The antenna-LED (a) is excited by a sharp and narrow Gaussian current pulse in (b), with a total charge of 1000 electrons. The current causes generation of minority carriers in the LED which recombine radiatively with an internal quantum efficiency of 80% (c). Light from the antenna-LED radiates with 70% efficiency and is mode-matched to a waveguide with 90% efficiency. The final optical pulse that is sent to the receiver is shown in (d). The optical pulse consists of about 500 photons with $<$ 20ps full-width.

Another option to increase the optical power would be to multiplex multiple antenna-LEDs in parallel. This is generally a difficult task, because parallel inputs must be differentiated in either wavelength or spatial mode profile. An example of the former was analyzed by the current author in Ref. [48], where wavelength division multiplexing was implemented by using optical antenna-LEDs with differentiated antenna resonance frequencies. A moderate increase of 2$\times$ optical power could be achieved using three antenna-LEDs in parallel.

Antenna-LED large-signal modulation bandwidth and photons per bit

While the peak optical power in steady-state may be limited, perhaps of more interest to the construction of optical links would be how many photons could be transmitted per optical bit in the time domain. Correspondingly, we simulated a full transient transmitter model in Fig. 4.12 using the detailed model of the antenna-LED physics provided in Appendix B and Appendix E. We assume $N_A = 2 \times 10^{19}\text{cm}^{-3}$

doping, $F_{\text{average}} = 166$, surface recombination velocity SRV $= 10^4$cm/s, and total efficiency defined above in Eq. 4.46. The optical transmitter structure (Fig. 4.12(a)) consists of a cavity-backed slot antenna-LED mode-matched to a waveguide. In Fig. 4.12(b) we pump the antenna-LED with an extremely short Gaussian current pulse with a peak current of 100μA and FWHM of about 2ps, which corresponds to 1000 total electrons. The current pulse causes generation of minority carriers within the antenna-LED according to the internal quantum efficiency in Fig. 4.12(c). For the assumed enhancement, SRV, device dimension ($d = 20$nm), and Auger constant ($C = 10^{28}$cm^6/s), the internal quantum efficiency for this conversion is about 80%. Note that we have pumped this device with a very high peak current pulse, but because of the high recombination rate and carrier dynamics the antenna-LED only reaches a peak carrier concentration of about 5×10^{18}1/cm^3. The antenna-LED has intrinsic absorption in the form of Ohmic loss, so it only radiates with $\eta_{\text{Antenna}} = 70\%$ efficiency. Then, the radiation is coupled into an optical waveguide with 90% efficiency. After all of these effects are taken into account, Fig 4.12(d) provides the optical power transient that is sent towards the receiver. The peak optical power is 11μW, corresponding to our steady-state calculation provided above with peak carrier concentration 5×10^{18}1/cm^3. There is then a long tail (turn-off) transient, correspondingly approximately to the enhanced spontaneous emission carrier lifetime $\tau_{\text{sp}}^* = 6$ps. Note that the turn-on transient was very short owing to the sharpness of the input current pulse.

The full optical power transient from the antenna-LED is less than 20ps in width. Therefore, we claim that the antenna-LED is capable of bandwidth exceeding $f > 1/20\text{ps} = 50$Gb/s. However, the total photons delivered in that pulse is about 500 photons/pulse (or photons/bit if desired). This is consistent with the total quantum efficiency of the system – in other words 50% of 1000 input electrons is 500 photons. Note that if we were to operate the transmitter at a slower modulation speed (longer pulses), the number of photons per bit could be increased. To summarize, we claim a modulation bandwidth and photons/bit of:

$$\text{Antenna-LED Large-Signal Modulation Bandwidth} > 50\text{Gb/s} \tag{4.51}$$

$$\text{Antenna-LED Photons/bit} = 500 \tag{4.52}$$

Is 500 photons/bit sufficient for adequate signal-to-noise ratio at the photonic receiver? Perhaps the ultimate limit to receiver scaling would be a shot noise limited detector. The shot noise limit for photodetectors can be estimated as follows. The electrical current generated by a photodetector with unity quantum efficiency is $I_{\text{signal}} = Pq/\hbar\omega$, where P is the received optical power. The shot noise current corresponding to this optical power is then given by $I_{\text{noise}} = \sqrt{2qI_{\text{signal}}\Delta f} = \sqrt{2q^2P\Delta f/\hbar\omega}$, where Δf is the modulation bandwidth. Thus, the shot noise limited signal-to-noise ratio is given by,

$$\text{SNR}|_{\text{shot noise limited}} = \frac{I_{\text{signal}}}{I_{\text{noise}}} = \frac{Pq/\hbar\omega}{\sqrt{2q^2P\Delta f/\hbar\omega}} = \sqrt{\frac{P}{2\hbar\omega\Delta f}} \tag{4.53}$$

Furthermore, the required photons per bit can be estimated as,

$$\frac{\text{photons}}{\text{bit}}| \approx \frac{P\tau}{\hbar\omega} \approx \frac{P}{2\hbar\omega\Delta f} \tag{4.54}$$

where $\tau = 1/2\Delta f$ is the bit period. But this is just the argument of the radical in Eq. 4.53. Thus, we may rewrite in favor in favor of photons/bit as,

$$\frac{\text{photons}}{\text{bit}}\Big|_{\text{shot noise limited}} \approx \text{SNR}^2\big|_{\text{shot noise limited}} \tag{4.55}$$

To achieve a bit error rate $< 10^{-9}$ for errorless operation, an SNR of at least 6 is recommended [70]. Therefore, the shot-noise limited photons per bit is photons/bit ≈ 36. This represents a massive improvement over current technology (which requires around 10,000 photons/bit), but indicates the feasibility of a very low-power optical transmitter like the antenna-LED.

More practical recent analyses anticipating low-capacitance on-chip photodiodes and next-generation CMOS receiver circuitry suggest huge improvements to the required photons/bit. Lalau-Keraly [70] performed a detailed analysis of a number of receiver circuit front-ends that could operate at high speed. A low-capacitance photodiode paired with a CMOS trans-impedance amplifier could potentially operate at a transistor-noise-limited <1000 photons/bit assuming optimistic but realistic technology improvements. Alternatively, a novel bipolar phototransistor with decoupled gain and absorption regions could offer unparalleled sensitivity and speed. Thus, while the antenna-LED power is low, there is hope that it will be viable with reasonable improvements to next-generation receivers.

4.5 Conclusion

In this chapter we demonstrated the physics of antenna-enhanced spontaneous emission using circuit theory. We went on to perform a detailed analysis of the anticipated average enhancement seen by a semiconductor LED coupled to a cavity-backed slot antenna. Ultimately, we found that the enhanced spontaneous emission carrier lifetime in antenna-LEDs could be as fast as the stimulated emission carrier lifetime in lasers, given approximately by $\tau = 6$ps for each case respectively. We argued that nanoscale optical sources in next-generation on-chip optical interconnects must be directly electrically modulated in a large-signal modulation format. Under this condition, both lasers and antenna-LEDs are limited by their respective carrier lifetimes. Therefore, antenna-LEDs can be as fast lasers. From an optical link purview, we went on to examine the total antenna-LED quantum efficiency and photons per transmitted bit. We showed that antenna-LEDs could reach excellent total efficiency of 50% if surface recombination velocity can be improved to SRV $= 10^4$cm/s or better. Moreover, we showed that the antenna-LED is capable of >50Gb/s large-signal modulation with 500 photons/bit. The low-optical power of the antenna-LED is fundamental, owing to the nanoscale LED active volume. Thus, high-speed receivers will require major sensitivity improvements in order to make antenna-LEDs feasible.

Chapter 5

Efficient Antenna-LED Waveguide Coupling and Metal-Dielectric Antennas

In the previous two chapters we gave a fundamental description of lasers and LEDs, showed the potential benefits of an optical antenna-LED for speed and quantum efficiency, and went on to analyze the anticipated benefits and remaining challenges of using an antenna-LED in a photonic link. In this chapter we will discuss two supplemental topics: (1) antenna-LED single-mode waveguide coupling, and (2) metal-dielectric antenna-LEDs for efficient spontaneous emission.

The former topic (1) is important for our discussion of the transmitter efficiency from the previous chapter where we assumed a simple value of 90% coupling efficiency to a single-mode waveguide. In Section 5.1 we will justify this value, and show how large single-mode waveguide coupling efficiency may be achieved with the cavity-backed slot antenna. This work is largely imported verbatim from the published manuscript [4], citation reproduced here:

> N. M. Andrade, S. Hooten, S. A. Fortuna, K. Han, E. Yablonovitch, and M. C. Wu, "Inverse design optimization for efficient coupling of an electrically injected optical antenna-LED to a single-mode waveguide," Opt. Express, vol. 27, no. 14, pp. 19802–19814, Jul. 2019.

To address the latter topic (2), Section 5.2 discusses how metal-dielectric antennas with sharp dielectric tips can overcome an Ohmic loss barrier to efficient spontaneous emission enhancement. It goes on to discuss the Purcell effect for antennas. This work is largely imported verbatim from the unpublished but accepted work with citation:

> S. Hooten, N. M. Andrade, M. C. Wu, and E. Yablonovitch, "Efficient spontaneous emission by metal-dielectric antennas; antenna Purcell factor explained," Optics Express (Accepted but Unpublished), 2021.

Please note that the appendices referred to in Section 5.1 and Section 5.2 are provided at the bottom of these respective sections and not in the main appendices of this thesis, except when specifically indicated otherwise.

5.1 Inverse design optimization for efficient coupling of an electrically injected optical antenna-LED to a single-mode waveguide

Abstract

Efficient high speed nanoscale optical sources are required for low power next generation data communication. Here we propose an integrated antenna-LED on a single-mode optical waveguide. By leveraging inverse design optimization, we achieved a waveguide coupling efficiency of 94% and an antenna efficiency of 64%, while maintaining a high average enhancement of 144 – potentially enabling >100GHz direct modulation.

Introduction

The development of high-density integrated optical interconnects is increasingly important to reduce on-chip energy consumption to less than 10fJ/bit [93]. Integrated optical interconnects require fast and efficient nanoscale light sources that are electrically injected and capable of being efficiently coupled to a photonic waveguide. While lasers are extensively used for efficient high speed optical communication, shrinking them down to the nanoscale poses significant problems due to metal loss [25]. LEDs are capable of scaling down to the nanoscale and can operate efficiently without a threshold, but they are limited in speed by their spontaneous emission rate to about 1GHz. However, by coupling the LED to an optical cavity, we can enhance the spontaneous emission rate [30, 43, 110, 31, 37], which would allow for ¿100GHz direct modulation. Only a few reports have demonstrated electrical injection [37, 51, 62], with the electrically injected cavity-backed slot antenna (Fig. 5.1) demonstrating ˜200x peak enhancement [37].

A cross-section of the cavity-backed slot antenna-LED is shown in Fig. 5.1[1]. As shown in Fig. 5.1(b), the radiation of the cavity-backed slot antenna is primarily directed towards the substrate, making it a non-trivial problem to couple to a photonic waveguide. Many methods have been used to couple nanoscale devices to waveguides, including coupling an optically pumped dipole antenna to a multimode waveguide using the waveguide height to cancel the electric field propagating toward the substrate [29], an electrically injected metal cavity LED and laser on a single-mode waveguide using the mode shape in the metal cavity [27, 63], and using anti-symmetric second-order resonance for a double nanogap plasmonic antenna [58]. Overall, efficient devices that are compatible with electrical injection and have high enhancement are still needed.

Electromagnetic inverse design[2] has been used to improve characteristics of a multitude of photonic devices [90, 121, 38, 14, 111, 71, 87, 76, 32]. For example, inverse design has been used to find high efficiency vertical grating couplers [90], to design a small footprint polarization beamsplitter [121], to op-

[1] Note that this differs from the cavity-backed slot antenna from the previous chapter in Fig 4.4 because we have assumed an InGaAs quantum well active region and the antenna has slightly different dimensions and enhancement properties that will be discussed below.

[2] This topic will be discussed in much greater detail in Chapter 6 of this thesis

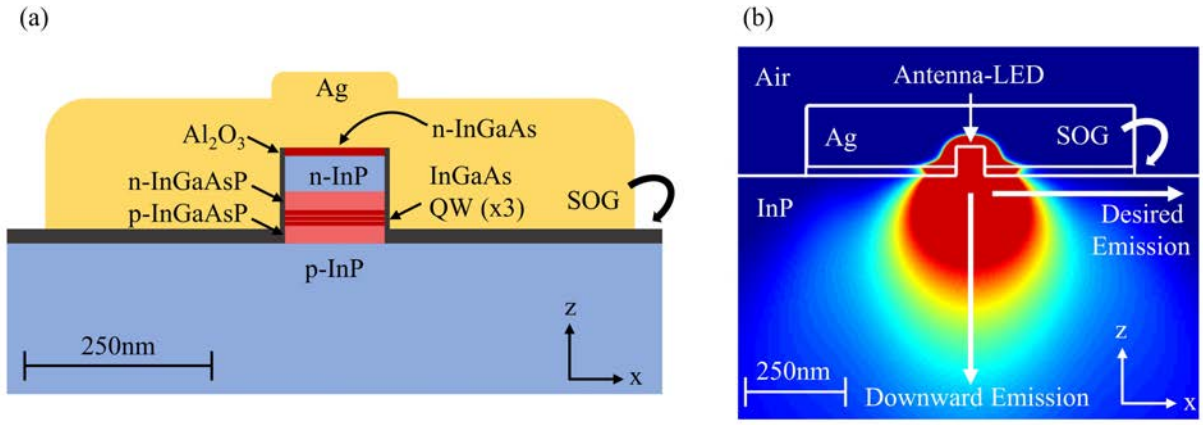

Figure 5.1: (a) Vertical cross section schematic and (b) power flow of optical antenna-LED on a bulk InP substrate. The XZ cross section depicts the LED length and height.

timize a broadband two-mode de-multiplexer [38], to increase the near-field enhancement of an optical antenna while minimizing temperature rise [14], and to optimize fabrication-constrained silicon photonic devices [111].

In this report, we designed and simulated various waveguide coupled antenna-LEDs. We then used inverse design optimization on our best hand-designed structure to generate two structures optimized at a single frequency and multiple frequencies, respectively. In our multi frequency design, we achieved a waveguide coupling efficiency of 94%, an antenna efficiency of 64%, and an average enhancement of 144. The proposed design is potentially compatible with electrical injection and top down fabrication.

Design background

Cavity-backed slot antenna on a bulk InP substrate

The cavity-backed slot antenna is a promising candidate as an optical source due to its high spontaneous emission enhancement and compatibility for top down fabrication and electrical injection [37]. As shown in Fig.5.1(a), the cavity-backed slot antenna is self-aligned to an InP/InGaAs/InP ridge (length: ≈130nm, width: 20nm, height: 140nm), where the height and length were chosen to tune the resonance frequency to best match the LED material spectrum, while maximizing the radiated power for the fundamental antenna mode. The antenna is electrically connected to the top of the ridge, where it is used as a contact to inject electrons into the n-InGaAs contact layer. The holes are injected into the p-InP layer, which is insulated from the antenna using a 40nm thick spin on glass (SOG). Finally, the InGaAs quantum well active region is electrically insulated from the antenna using a 1nm thick Al_2O_3 surrounding the ridge sidewalls. When an electron and hole recombine in the active region it acts as a dipole excitation of the antenna mode. In our 3D finite-difference time-domain (FDTD) simulations, we excited the antenna by placing an electric dipole source in the active region.

Figures of merit

The presence of the optical antenna causes the dipole to radiate more power than if it was in bulk InGaAs, the ratio of these powers provides the enhancement spectrum [30]. For a fair analysis we considered all the dipoles in the active region, accounting for polarization, position, and overlap with the material spectrum. This gives the average enhancement (F_{avg}), which is directly related to both the output power and the modulation rate[3].

In addition to the average enhancement, we considered the average antenna efficiency (η_{antenna}) and waveguide coupling efficiency to the fundamental mode (η_{WC}). The antenna efficiency is the fraction of total optical power which is not lost to metal (i.e. the power that reaches the far field). The waveguide coupling efficiency is the fraction of the far field power in the fundamental mode of the waveguide (i.e. it only accounts for the scattering loss). For explicit definitions and averaging factors used see *Appendix: Figures of merit*. In the remainder of the text the figures of merit discussed are these average quantities, unless otherwise noted.

Waveguide coupling design

In our previous work we proposed designs to couple light from the cavity-backed slot antenna to a single-mode waveguide [3, 2]. In this subsection we will describe some of the intuition behind these designs, and how they helped achieve efficient waveguide coupling. As shown in Fig. 5.2(a), we optimized the waveguide height and width in order to cancel the fields propagating towards the substrate (similar to [29]) and achieved a waveguide coupling efficiency of 24% in each direction with a waveguide height of 180nm and a width of 550nm. In Fig. 5.2(b), we truncated the waveguide and wrapped metal around the end of the facet to effectively act as a mirror. In addition to making the coupling unidirectional, the mirror created an image dipole 180° out of phase with the antenna-LED, which further suppressed fields propagating toward the substrate. By minimizing the separation between the antenna-LED and the back mirror, we achieved a waveguide coupling efficiency of 74% – note this was more than double the result from Fig. 5.2(a). Finally, in Fig. 5.2(c) we improved the coupling to the fundamental mode by tapering the waveguide near the antenna-LED and wrapping metal around the sidewall of the tapered section. Figure 5.3(a) shows the perspective view and Fig. 5.3(b) shows the enhancement, antenna efficiency, and waveguide coupling efficiency spectra. With this structure we were able to achieve an average enhancement of 162, a waveguide coupling efficiency of 90%, and an antenna efficiency of 49%.

Although our hand-optimized results are comparable to the best results in the literature, we were restricted to exploring only simple geometries of the waveguide coupler due to the immense computational resource requirements of simulating fine-meshed three-dimensional optical structures. In order to more completely explore the parameter space associated with this waveguide coupler, we applied computational inverse design techniques.

[3] Average enhancement was detailed in the previous chapter of this thesis, and also in Appendix G.

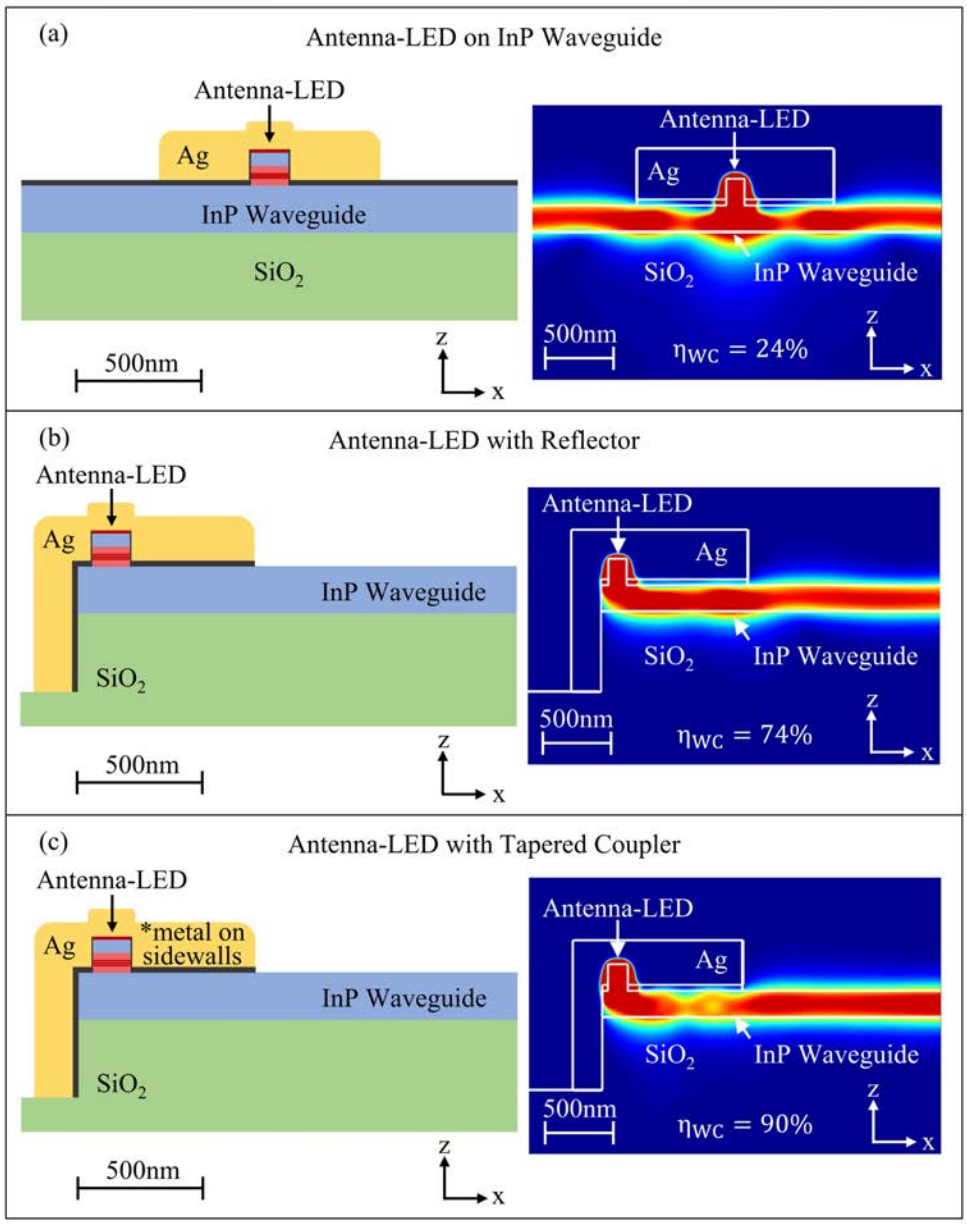

Figure 5.2: Cross section, power flow, and waveguide coupling efficiency to the fundamental mode (η_{WC}) for (a) antenna-LED on single-mode InP waveguide and SiO_2 ridge, (b) antenna-LED on single-mode InP waveguide with metal wrapped around waveguide facet, and (c) antenna-LED on single-mode InP tapered waveguide with metal wrapped around waveguide facet and sidewalls (see Fig. 5.3(a) for perspective view, Fig. 5.4(b) for top view cross section). See *Appendix: Field profiles* for the E_x and E_y field profiles of the mode in the InP waveguide.

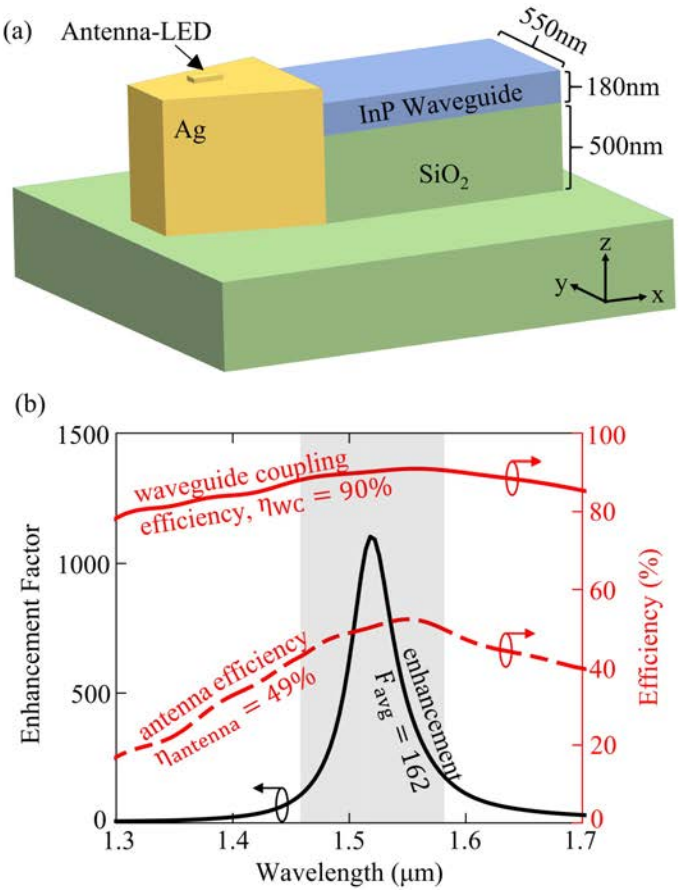

Figure 5.3: (a) Perspective view of tapered waveguide coupler with a waveguide height of 180nm and width of 550nm on a 500nm tall SiO_2 ridge, and (b) enhancement, antenna efficiency, and waveguide coupling efficiency spectra.

Inverse design

Gradient-descent based optimization using the adjoint method can be used to optimize almost any user-defined electromagnetic figure of merit over an arbitrarily large parameter space with minimal computational resource requirements [71, 87]. In the literature this optimization method and similar topology optimization methods are commonly referred to by the more general term inverse design, which we will adopt in order to help easily distinguish the various results in this report. For brevity we will not delve into the details of the method, but we recommend the reader review the works in [71, 87, 111, 32, 76] for more information See *Appendix: Inverse design* for specifics regarding our implementation of inverse design [4].

Inverse design was applied to the 2D cross section of the tapered coupler (Fig. 5.4(b)) to optimize

[4] This method is also detailed in Chapter 6 of this thesis.

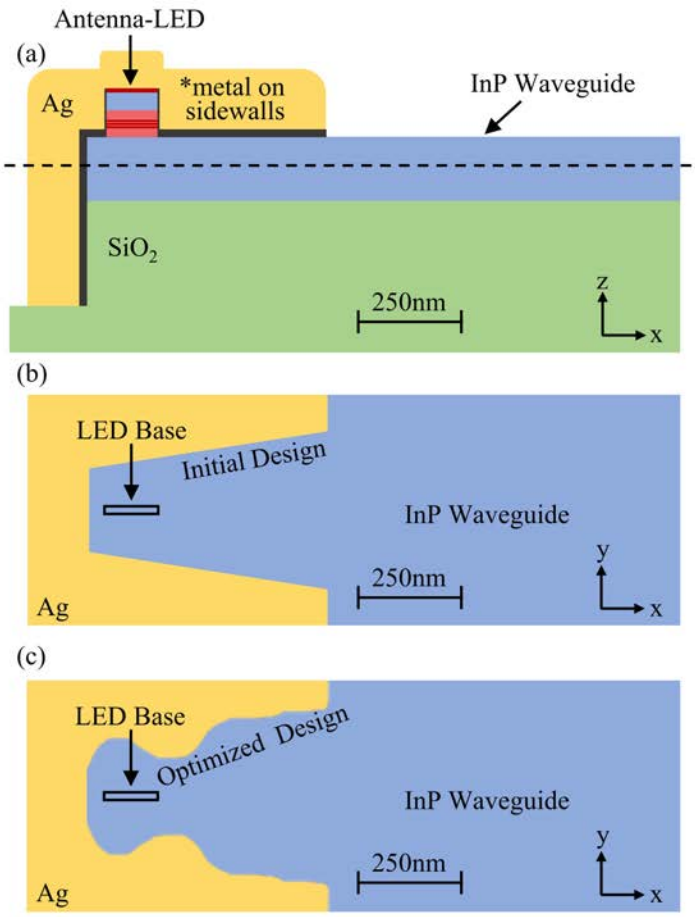

Figure 5.4: (a) Cross section schematic (XZ) of tapered waveguide coupler showing dashed cutline, and (b) top view XY cross section of waveguide along dashed cutline. (c) XY cross section of coupler after optimization, showing perturbations to Ag-InP boundary. Note (b) and (c) also show the projection of the LED base.

enhancement, antenna efficiency, and waveguide coupling efficiency by perturbing the interface between InP and Ag (Fig. 5.4(c)) – the spectra before optimization are shown in Fig. 5.3(b). Our initial inverse design cost function was the power transmitted through the waveguide at a single frequency (spectral product at the resonant frequency of enhancement, antenna efficiency, and waveguide coupling efficiency). This led to a slight improvement in the power transmitted at resonance compared to the tapered coupler – shown in Fig. 5.5(a) and Fig. 5.3(b), respectively. However, when calculating the average values, we noticed there was a large trade-off between average enhancement and antenna efficiency. When compared to the tapered coupler, even though the peak enhancement increased from 1034 to 1312, the average enhancement only increased from 162 to 164. However, the antenna efficiency dropped from 49% to 40%. Waveguide coupling increased slightly from 90% to 94%. When we combine these numbers, we see that the average power of the single frequency optimization was lower than the tapered coupler. This is not surprising since the cost function did not represent an average value.

In order to increase the average power transmitted, we changed the inverse design cost function to be the weighted sum of the optical power at three frequencies. We weighted the power transmitted at resonance ten times less than the power transmitted at $\pm 55\text{THz}$ ($\pm 40\text{nm}$) from resonance to encourage a broader enhancement spectrum. As shown in Fig. 5.5(b), we were able to create a broader enhancement spectrum with a greater antenna efficiency – ultimately achieving $F_{\text{avg}} = 144$ and $\eta_{\text{antenna}} = 64\%$.

Discussion

Our design methodology is contingent on the LED material spectrum, shown in the Appendix, Fig. 5.7. Given a narrower material spectrum, the single frequency design could be more desirable since the average enhancement would be much larger than the multi frequency design or tapered coupler. Even with our current material spectrum, the single frequency design will theoretically have the fastest direct modulation rate – however at a great expense to antenna efficiency. In contrast, the multi frequency design will have a slower direct modulation rate, but it maintains high enhancement while achieving the highest efficiency making it capable of delivering the most optical power to the waveguide. In fact, when we compare the product of F_{avg}, η_{antenna}, and η_{WC} from the multi frequency design with the cavity-backed slot antenna on a bulk InP substrate, we find that we could emit slightly more power in the fundamental mode of an InP waveguide than would be radiated in all directions for the bulk InP substrate case.

Close observation of the multi frequency design enhancement spectrum in Fig. 5.5(b) reveals two distinct peaks. This can be explained by thinking of the antenna-LED and coupler section (see inset Fig. 5.6(a)) as coupled resonators. When they have the same resonance frequency, it will lead to a frequency split that can be observed in the enhancement spectra. This was confirmed by sweeping the LED length in the multi frequency design, which resulted in an avoided crossing between the antenna-LED resonance and the coupler section resonance, as shown in Fig. 5.6. The dashed black line was generated by sweeping the length of the antenna-LED on a bulk InP substrate (Fig. 5.1(a)). The dashed green line was created by placing a dipole in the coupler section (see inset) and sweeping the length of an off-resonance antenna-LED. During the length sweep we found that the antenna efficiency always peaked at the coupler section resonance rather than at the antenna-LED resonance.

A similar observation was made in the single frequency design in Fig. 5.5(a), the antenna efficiency peak was associated with the coupler resonance. However, in contrast to the multi frequency design,

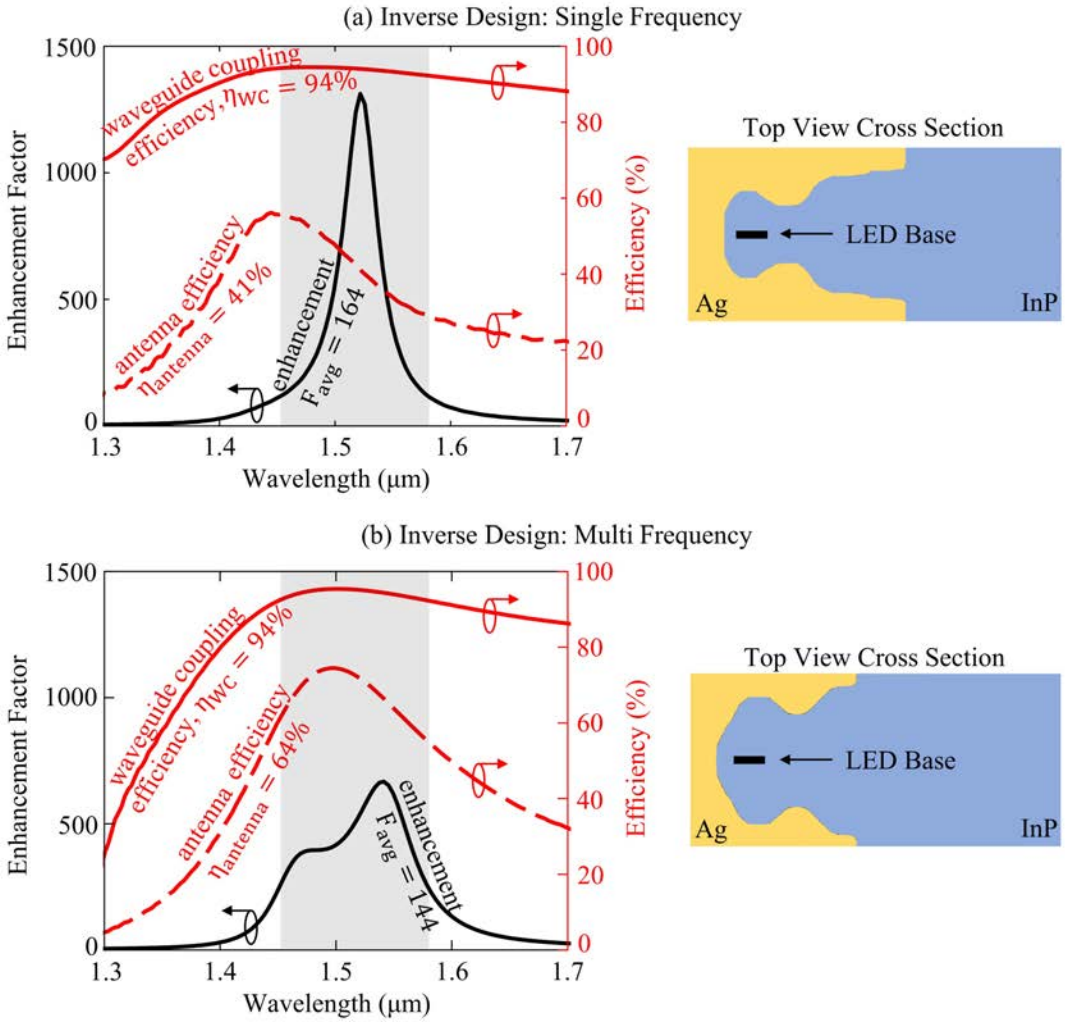

Figure 5.5: Enhancement, antenna efficiency, waveguide coupling efficiency spectra and top view XY cross sections for (a) single frequency optimization and (b) multi frequency optimization. For reference, the LED material spectrum $[L(\omega)]$ between its 50% power points is shown by the gray shaded region.

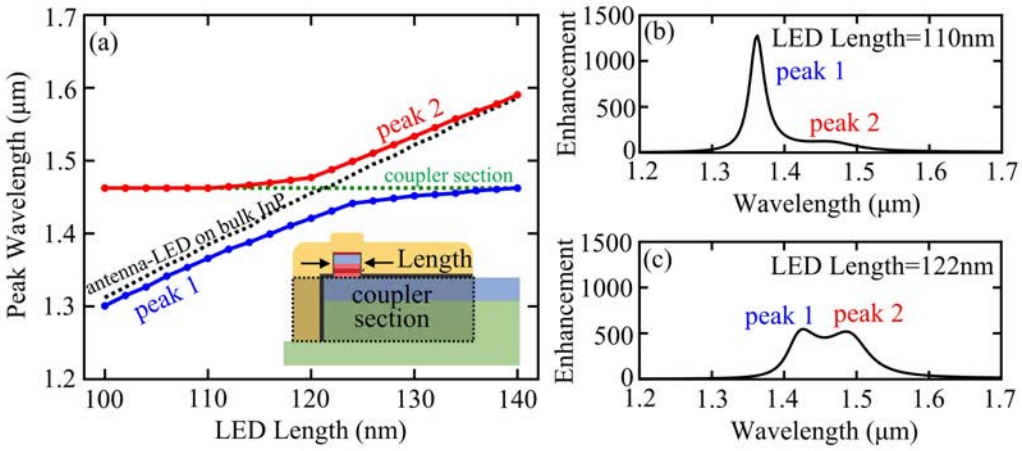

Figure 5.6: (a) Avoided crossing between the optical antenna resonance and the inverse design coupler resonance. For reference, dashed black and green lines show independent resonances of the antenna-LED on a bulk InP substrate and the coupler section as a function of LED length, respectively. Enhancement spectra for LED lengths of (b) 110nm and (c) 122nm.

the antenna-LED and coupler section resonances are detuned – evident by the offset between the peak enhancement and antenna efficiency wavelengths in Fig. 5.5(a).

To summarize, the spectra of the waveguide coupling designs can be explained by considering the antenna-LED and the coupler section as coupled resonators. When the resonances are tuned (multi frequency design), we have an impedance match and frequency splitting. Due to the impedance match, the optical power is able to quickly leave the lossy antenna-LED (lower Q factor) resulting in less metal loss (higher antenna efficiency). In contrast, when the resonances are detuned (single frequency design), we have an impedance mismatch which results in the optical power reflecting back to the lossy antenna region. This results in more metal loss (lower antenna efficiency) and higher enhancement. A similar conclusion was reached in [26], where detuned resonators were exploited to achieve higher peak enhancement. Note that regardless of how the coupler section resonance was tuned, both these designs yielded higher waveguide coupling efficiency than the tapered coupler.

Conclusion

We have demonstrated that the cavity-backed slot antenna-LED can be efficiently coupled to a single-mode waveguide, which was validated using relevant figures of merit in an optical interconnect. Then, using inverse design we further optimized the cavity-backed slot antenna coupling, ultimately achieving a waveguide coupling efficiency of 94%, antenna efficiency of 64%, while maintaining a high average enhancement of 144. We found that inverse design was able to achieve these results by tuning the optical resonance of the coupler section relative to the antenna-LED based on our cost function.

Due to its high efficiency, nanoscale size, compatibility with top-down fabrication, and speed the

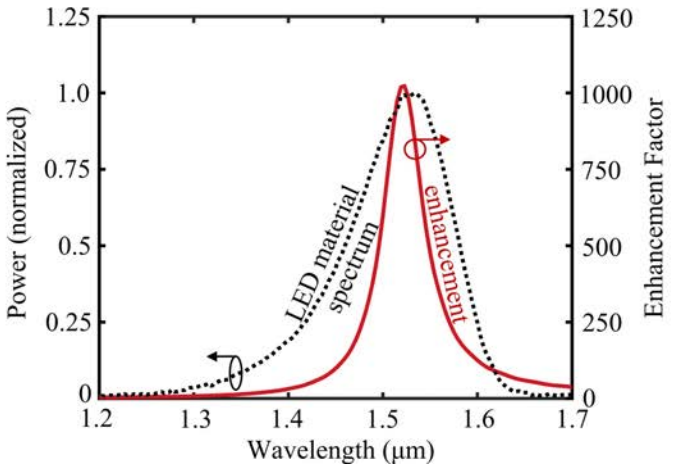

Figure 5.7: Dashed black and solid red lines show the experimental non-enhanced material spectrum $[L(\omega)]$ and the simulated enhancement spectrum $[F(\omega)]$ of the cavity-backed slot antenna on a bulk InP substrate, respectively.

cavity-backed slot antenna-LED is a very promising transmitter for an on-chip optical interconnect.

Appendix

Figures of merit

In Fig. 5.7, we show the FDTD simulation of the enhancement spectrum of a dipole source with the optimal polarization and position in the cavity-backed slot antenna on a bulk InP substrate from Fig. 5.1(a). Additionally, the experimental material spectrum from a large area LED with the same epitaxial layers as Fig. 5.1(a) is provided by the dashed black line in Fig. 5.7. The product of the material spectrum and the enhancement spectrum gives the output power spectrum from the dipole source.

In order to calculate the average increase in output power (i.e. the average enhancement) we need to account for all the dipoles in the active region. Therefore, we consider the dipole frequency response, polarization dependence, and position dependence. We defined the average enhancement as the following:

$$F_{\text{avg}} = \frac{1}{2} \times 0.79 \times \frac{\int F(\omega)L(\omega)d\omega}{\int L(\omega)d\omega} \tag{1}$$

where $\frac{1}{2}$ is the polarization average, 0.79 is the spatial average, $F(\omega)$ is the overall enhancement spectrum seen by a dipole with the optimal polarization and position, and $L(\omega)$ is the experimental material spectrum without an antenna present. Note that the final term in Eq. (1) is the spectral average. In principle the material spectrum $L(\omega)$ is dependent on the carrier concentration; however, in this report we fixed $L(\omega)$, and therefore the carrier concentration in order to simplify the analysis. The spatial average and

polarization average were found by sweeping dipole position and polarization in the quantum well active region in the cavity-backed slot antenna.

In addition to the average enhancement, we considered the antenna efficiency (η_{antenna}) and waveguide coupling efficiency to the fundamental mode (η_{WC}). The antenna efficiency only accounts for the metal loss, and the waveguide coupling efficiency only accounts for the scattering loss. The explicit definitions for the antenna efficiency and waveguide coupling efficiency spectra are shown below:

$$\eta_{\text{antenna}}(\omega) = \frac{P_{\text{total}}(\omega) - P_{\text{metal loss}}(\omega)}{P_{\text{total}}(\omega)} \tag{2}$$

$$\eta_{\text{WC}}(\omega) = \frac{1}{\eta_{\text{antenna}}(\omega)} \frac{P_{\text{fundamental mode}}(\omega)}{P_{\text{total}}(\omega)} \tag{3}$$

$$P_{\text{total}}(\omega) = P_{\text{fundamental mode}}(\omega) + P_{\text{scattering}}(\omega) + P_{\text{metal loss}}(\omega) \tag{4}$$

where $P_{\text{total}}(\omega)$ is the total optical power leaving the dipole source, $P_{\text{metal loss}}(\omega)$ is the power lost to metal, and $P_{\text{fundamental mode}}(\omega)$ is the power in the fundamental mode of the waveguide which was found by taking an overlap integral between the eigenmode solution and the simulated waveguide field profile. Note that the product of these efficiencies gives the fraction of the total optical power coupled to the fundamental waveguide mode. Additionally, we calculated the average antenna efficiency and waveguide coupling efficiency. Below are the explicit definitions for η_{antenna} and η_{WC}:

$$\eta_{\text{antenna}} = 0.96 \times \frac{\int \eta_{\text{antenna}}(\omega) F(\omega) L(\omega) d\omega}{\int F(\omega) L(\omega) d\omega} \tag{5}$$

$$\eta_{\text{WC}} = \frac{\int \eta_{\text{WC}}(\omega) \eta_{\text{antenna}}(\omega) F(\omega) L(\omega) d\omega}{\int \eta_{\text{antenna}}(\omega) F(\omega) L(\omega) d\omega} \tag{6}$$

where 0.96 is the spatial average for the antenna efficiency. Note that the polarization dependence was negligible for both average efficiencies, since a dipole oriented along the width of the LED sees much greater enhancement than a dipole oriented along the length. Additionally, the spatial dependence was negligible for the waveguide coupling efficiency.

These average values could now be used to calculate relevant device metrics since they represent the average response of a carrier in the device. Two important metrics are the power in the fundamental mode of the waveguide and the 3dB frequency, given in Eqs. (7) and (8), respectively.

$$P_{\text{fundamental mode}} = F_{\text{avg}} \eta_{\text{antenna}} \eta_{\text{WC}} \hbar\omega B_0 N^2 V \tag{7}$$

$$f_{\text{3dB}} = \frac{2 F_{\text{avg}} B_0 N}{2\pi} \tag{8}$$

where B_0 is the radiative recombination coefficient, N is the carrier concentration, V is the active region volume, and f_{3dB} is the 3dB modulation frequency assuming the radiative recombination rate is dominant. If we assume $F_{\text{avg}} = 164$, $B_0 = 10^{-10}\text{cm}^3\text{s}^{-1}$ [151], and $N = 2 \times 10^{19}\text{cm}^{-3}$ we could reach a 3dB frequency of 104GHz.

Inverse design

In this work we used the Berkeley Photonic Inverse Design package, originally described in [71]. The inverse design optimization problem that was solved can be written as the following:

$$\max_{\theta} \sum_{\omega} c_{\omega} T_{\omega}(x, r) : \text{Radius of Curvature} \geq 100\text{nm} \tag{9}$$

where θ denotes the optimization parameter space – which in this case is the interface between InP and Ag in the metal-optic waveguide coupler region, ω is an index that defines the frequency bandwidth of the optimization, T is the Poynting vector evaluated at positions r in the waveguide for electric and magnetic fields abbreviated by vector x, and c is a user-defined weight chosen for each frequency index. Finally, we included an optimization constraint on the radius of curvature to ensure fabricability. A brief discussion of the limitations of our inverse design implementation follow.

The objective function that was used in inverse design does not give individual control over our figures of merit, F_{avg}, η_{antenna}, and η_{WC}. Consequently, we included the weights, c, in the objective function to provide this control. An additional limitation comes in reference to Fig. 5.4(c) where the length of the metal along the coupler section sidewalls is not perturbed. Since it is undesirable to have metal along the sidewalls of the coupler section (XY plane) with a different length than the metal on top of the waveguide (XZ plane), the metal on top of the waveguide effectively constrained the designable region. Therefore, we used several metal lengths as initial conditions for inverse design optimization.

Lastly, one of the most important considerations for our choice of the waveguide coupler structure in Fig. 5.4 was its compatibility with top-down fabrication. In other words, since the entire ridge must share the same etch mask, it must also share the same 2D cross-sectional shape in the XY plane. Therefore, a geometrical constraint is required in the inverse design optimization to maintain the conformal nature of the ridge which is composed of several materials. Such a constraint was unavailable in our basic implementation of inverse design. We imposed this constraint ad hoc by updating the SOG-Ag and SiO_2-Ag interfaces every three iterations to match the changing InP-Ag interface, but no significant convergence issues were encountered.

Field profiles

In Figs. 5.8(a)-5.8(c) we plotted the E_{x} and E_{y} field profiles of the mode in the InP waveguide at 1550nm for the structures given in Figs. 5.2(a)-5.2(c), respectively. In each structure the electric field profiles have been self-normalized by the maximum electric field magnitude.

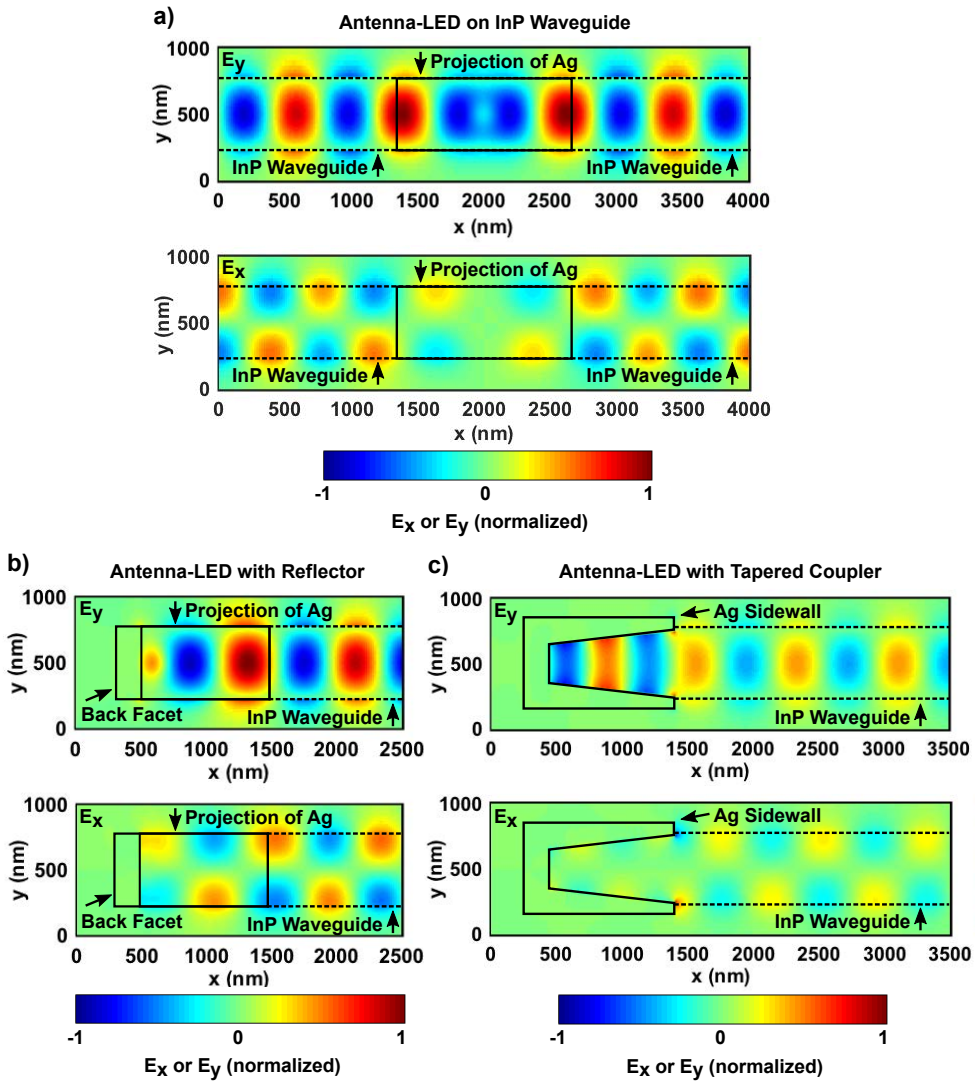

Figure 5.8: E_x and E_y field profiles for (a) antenna-LED on single-mode InP waveguide and SiO_2 ridge, (b) antenna-LED on single-mode InP waveguide with metal wrapped around waveguide facet, and (c) antenna-LED on single-mode InP tapered waveguide with metal wrapped around waveguide facet and sidewalls.

5.2 Efficient spontaneous emission by metal-dielectric antennas; antenna Purcell factor explained

Abstract

The rate of spontaneous emission from an optical emitter can be greatly enhanced using a metallic optical antenna at the penalty of efficiency. In this paper we propose a metal-dielectric antenna that eliminates the tradeoff between spontaneous emission enhancement and radiative efficiency by using nanoscopic dielectric structures at the antenna tips. This tradeoff occurs due to Ohmic loss and is further exacerbated by electron surface collisions. We find that our metal-dielectric antenna can enhance spontaneous emission by a factor 5×10^5 with efficiency = 70%, greatly exceeding the radiative efficiency of a purely metallic antenna with similar enhancement. Moreover, the metal-dielectric antenna design strategy is naturally amenable to short-distance optical communications applications. We go on to discuss the Purcell Effect within the context of antenna enhancement. Metallic optical antennas are best analyzed with conventional antenna circuit models, but if the Purcell Enhancement were to be employed, we provide the effective mode volume, $V_{\text{eff}} = (3/4\pi^2)^2 d^2 \lambda (\lambda/l)^5$, that would be needed.

Section 1: Introduction

Enhancing the rate of decay and spontaneous light emission from nanoscale optical sources using antennas has been the subject of considerable classical and contemporary research [35, 17, 98, 7, 99, 131, 41, 16], with potential applications in spectroscopy [102, 101, 56, 146], single-photon sources [110, 74, 33, 43], and efficient on-chip optical data communications [30, 126, 37, 29, 27, 4, 58]. Metallic optical antennas are well-suited for spontaneous emission enhancement because electromagnetic fields are naturally confined to sharp metallic tips, thereby boosting the radiative transition rate of excited molecules near the tips by the increased electric dipole interaction potential [21, 22]. However, one finds that large enhancement factor comes at the expense of inefficiency in metallic optical antennas [30]. In Section 2, we discuss this tradeoff of enhancement versus efficiency, which occurs because of Ohmic loss and is further exacerbated by nonlocal surface collision effects [30, 108, 28, 68]. To alleviate loss, metal-dielectric antennas have been proposed [107, 26, 61, 147, 127, 152, 128]. These antennas typically use lossless dielectrics to reduce Ohmic loss by pulling the highest field regions away from the lossy metal, which may come at the penalty of reduced antenna enhancement. In Section 3, we propose a novel metal-dielectric antenna that leverages dielectrics for extreme near-field light focusing – inspired by the purely dielectric cavities in [19, 49], but also benefitting from the presence of metal. The proposed antenna improves radiative efficiency and maintains the ultra-high spontaneous emission enhancement usually attained by purely metallic antennas. Section 4 demonstrates that the metal-dielectric antenna design principle is applicable to electrically-injected antenna-enhanced light-emitting diodes (antenna-LEDs), which can be used for on-chip optical communications. In Section 5, we derive a new antenna effective mode volume formula, which permits continued use of the Purcell effect for describing antenna enhancement. We compare the effective mode volume formula to a full electromagnetic numerical analysis in Section 6.

Section 2: Tradeoff of enhancement versus efficiency in metallic optical antennas

Optical antenna-enhanced spontaneous emission can be regarded as the increase in steady-state radiated power from an oscillating dipole when coupled to an optical antenna:

$$\text{Enhancement} = \frac{P_{\text{rad}}}{P_{\text{o}}} \tag{5.1}$$

where P_{o} is the nominal radiated power from the light source without the antenna present and P_{rad} is the radiated power with the antenna [5]. For consistency, the reference source power P_{o} is chosen to be a point dipole emitting into free space. Note that in Eq. 5.1, P_{rad} includes only the radiated power, not the power that goes into Ohmic heating. To account for these additional metal losses, the antenna efficiency is defined as:

$$\text{Efficiency} = \frac{P_{\text{rad}}}{P_{\text{rad}} + P_{\text{loss}}} \tag{5.2}$$

where P_{loss} is synonymous with Ohmic loss. Neither antenna directivity nor waveguide mode-matching efficiency will be considered here.

Consider the metallic dipole antenna in Fig. 5.9(a). The optical antenna consists of two cylindrical silver wires with 25nm radii. At the center feedgap the antenna includes sharp cone-shaped metallic tips that are adjacent to an optical point dipole source, which could represent a dye molecule or other atomically sized emitter. Importantly, the tips are separated by a vacuum gap of width d. In the limiting case where d=1nm the radius of curvature at the tips is 1nm, but the radius of curvature increases as d increases. Practically speaking, a 1nm tip is technologically difficult to achieve, with at least one recent report claiming experimentally fabricated metallic tips of this dimension [146] to the authors' knowledge. Nevertheless, in this report we will examine several antennas with very sharp nanoscale tips in order to investigate their limiting behavior.

Eggleston et al [30] demonstrated a circuit model for a metallic dipole antenna similar to that shown in Fig 5.9(a). A simplified illustration of the antenna circuit model is presented in Fig. 5.9(b) [6]. The point dipole source is modeled as a current source (J_{source}) in series with radiation resistance in the antenna arms ($R_{\text{radiation}}$, which accounts for radiated light) and a parasitic spreading resistance (R_{spread}, which accounts for most Ohmic loss). The enhancement predicted by the antenna circuit model is plotted in Fig. 5.9(c), which agrees with full 3D Finite-Difference Time-Domain (FDTD) Maxwell simulations (black curve and red points, respectively).

When d is small, very large antenna enhancement is accompanied by a severe drop in antenna efficiency, as revealed in Fig. 5.9(d). Antenna efficiency decreases dramatically because of spreading resistance [40, 124], which is inversely proportional with the vacuum gap width; $R_{\text{spread}} = 2 \cdot \text{resistivity}/d$. Both

[5] Note that we are not considering average enhancement of an optical antenna-LED in this section, only the peak enhancement seen by an idealized dipole point source in the antenna gap.

[6] The circuit model enhancement and efficiency are given in Appendix F, and the basic physics of the circuit model enhancement were provided in the previous chapter of this thesis.

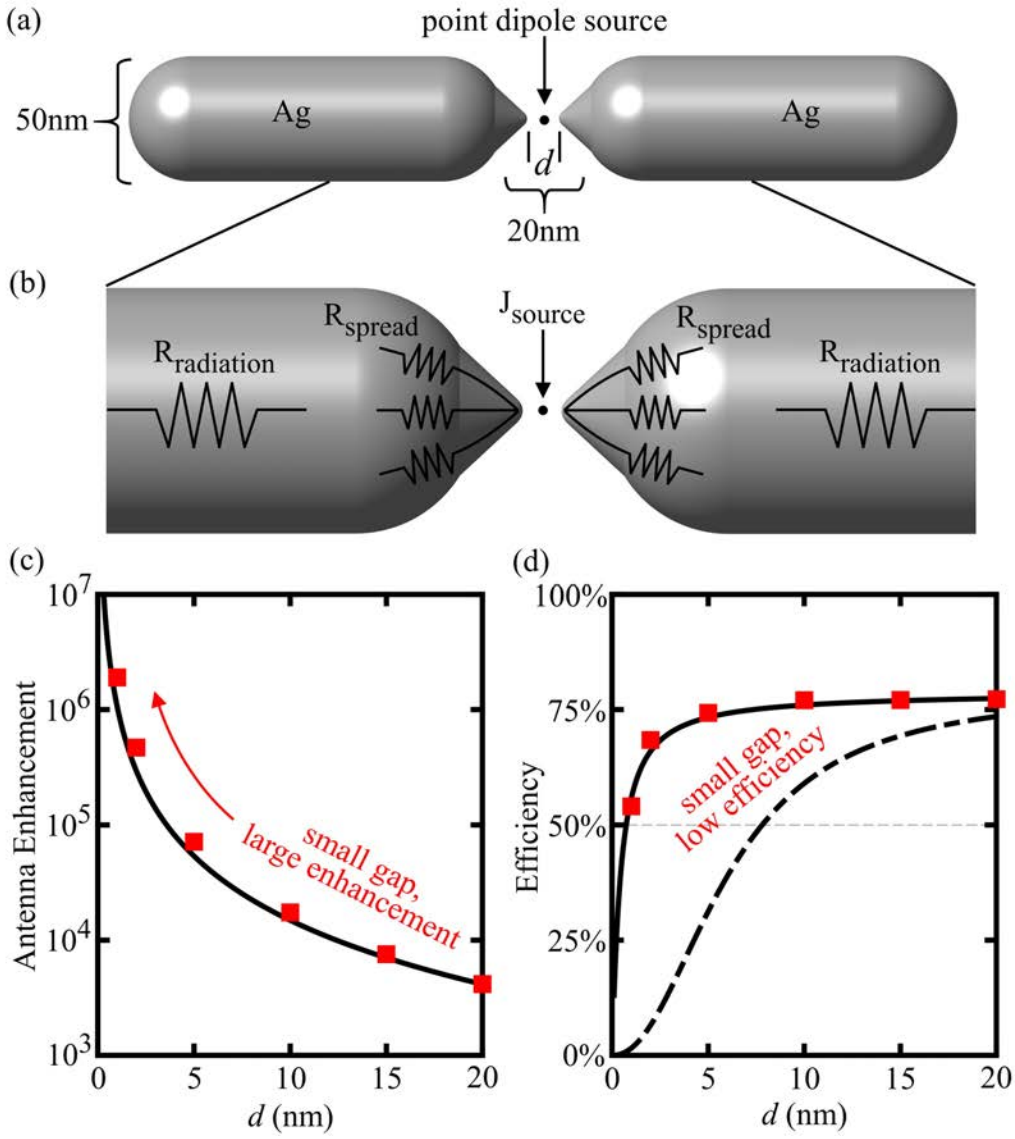

Figure 5.9: The efficiency of metallic antennas suffers due to spreading resistance and surface collisions. (a) Metallic dipole antenna. An optical point source resides in a vacuum gap of length d between sharp metallic tips (minimum radius of curvature = 1nm, cone angle = 90°). (b) The simplified circuit model of metallic optical antenna shows the antenna radiation resistance in series with a parasitic spreading resistance. (c) The spontaneous emission enhancement of the metallic antenna versus the vacuum gap d at a wavelength of λ=1550nm calculated using both a circuit model (black line) [30] and full 3D FDTD simulations (red squares). (d) The efficiency of the metallic antenna versus the vacuum gap d, calculated by circuit model (black line) and FDTD (red squares). For small d, the efficiency falls off dramatically due to spreading resistance. Also shown is the antenna efficiency that includes an estimate of the surface collision effect in the sharp tips (dashed line), which further exacerbates the spreading resistance effect.

the antenna circuit model and the full wave FDTD simulation correctly account for spreading resistance, as shown in the solid black curve and red points respectively.

However, there are additional losses associated with electron surface collisions that are not captured by Maxwell simulators, sometimes called the anomalous skin effect. This is a nonlocal effect that does not appear in optical material data handbooks due to its contingency on the specific geometrical structure of an optical material and motion of free electrons in the confined geometry. In fact, it is mentioned as one of the main sources of error for metals in Palik's Handbook [108]. Consequently, this effect is not included in Maxwell simulators and some prior investigators have been overoptimistic with regard to efficiency.

Eggleston et al [30] modeled electron surface collisions in the dipole antenna, which we have reproduced here in the dashed line of Fig. 5.9(d). Surface collisions increase the effective spreading resistance in the concentrated current region near the center antenna feedgap region by the factor $(1+l_\infty/\beta d)$, where l_∞ is the bulk electron mean free path in silver, d is the vacuum gap width (which also defines the radius of curvature at the antenna tips), and β is a numerical parameter that requires an intricate nonlocal electrodynamic calculation. Note that the factor $(l_\infty/\beta d)$ can be regarded as a term that corrects the mean free path of electrons from the nominal bulk mean free path l_∞, to a mean free path that is contingent upon the radius of curvature in the confined metallic tips [68, 28], $l_e \approx \beta d$. In Fig.5.9(d) we plotted an estimate of this surface collision effect with l_∞=50nm [68, 28] and β=0.5. With our chosen parameter β=0.5, the surface collision effect bounds the expected antenna efficiency to ≈50% for a practical antenna gap, d=10nm.

Section 3: Metal-dielectric antenna

In the previous section we demonstrated that purely metallic antennas suffer from poor efficiency at small gap width d due to Ohmic losses. In this section we will show that by including dielectrics in the antenna design we can greatly boost the antenna efficiency without significantly compromising the antenna enhancement at small d. This metal-dielectric antenna performs the best balance between all-metal and all-dielectric antenna designs when combining the two metrics of enhancement and efficiency.

Work from Vanderbilt and MIT demonstrated that the effective mode volume (i.e. the spatial light field concentration) of photonic crystal cavities is drastically improved by using sharp dielectric tips [19, 49]. This electromagnetic enhancement effect surpasses the anticipated enhancement associated with simple dielectric boundary conditions. Furthermore, dielectrics are effectively lossless compared to metals so this field concentration can be achieved with no series resistance limitation. We will demonstrate that metallic antennas augmented with dielectric tips can improve antenna efficiency while maintaining large enhancement.

Consider the metal-dielectric antenna in Fig. 5.10(a). This antenna is similar to the all-metal antenna in Fig. 5.9(a) except that the sharp metal tips have been replaced by sharp dielectric tips (refractive index n=3.4) covering hemispherical metal tips, indicated by the white dashed lines. The dielectric tips are separated by vacuum gap d, while the larger metal-to-metal distance at the metallic hemisphere tips is fixed to 20nm. In Fig. 5.10(b) and Fig. 5.10(c) we compare the enhancement (Eq. 5.1) and efficiency (Eq. 5.2) of the metal-dielectric antenna versus the all-metal antenna from Fig. 5.9. The enhancement factor, Eq. 5.1, was determined by direct FDTD computation, which is reliable for enhancement factor but not for efficiency. The efficiency was obtained by FDTD with a correction provided by the surface collision effect

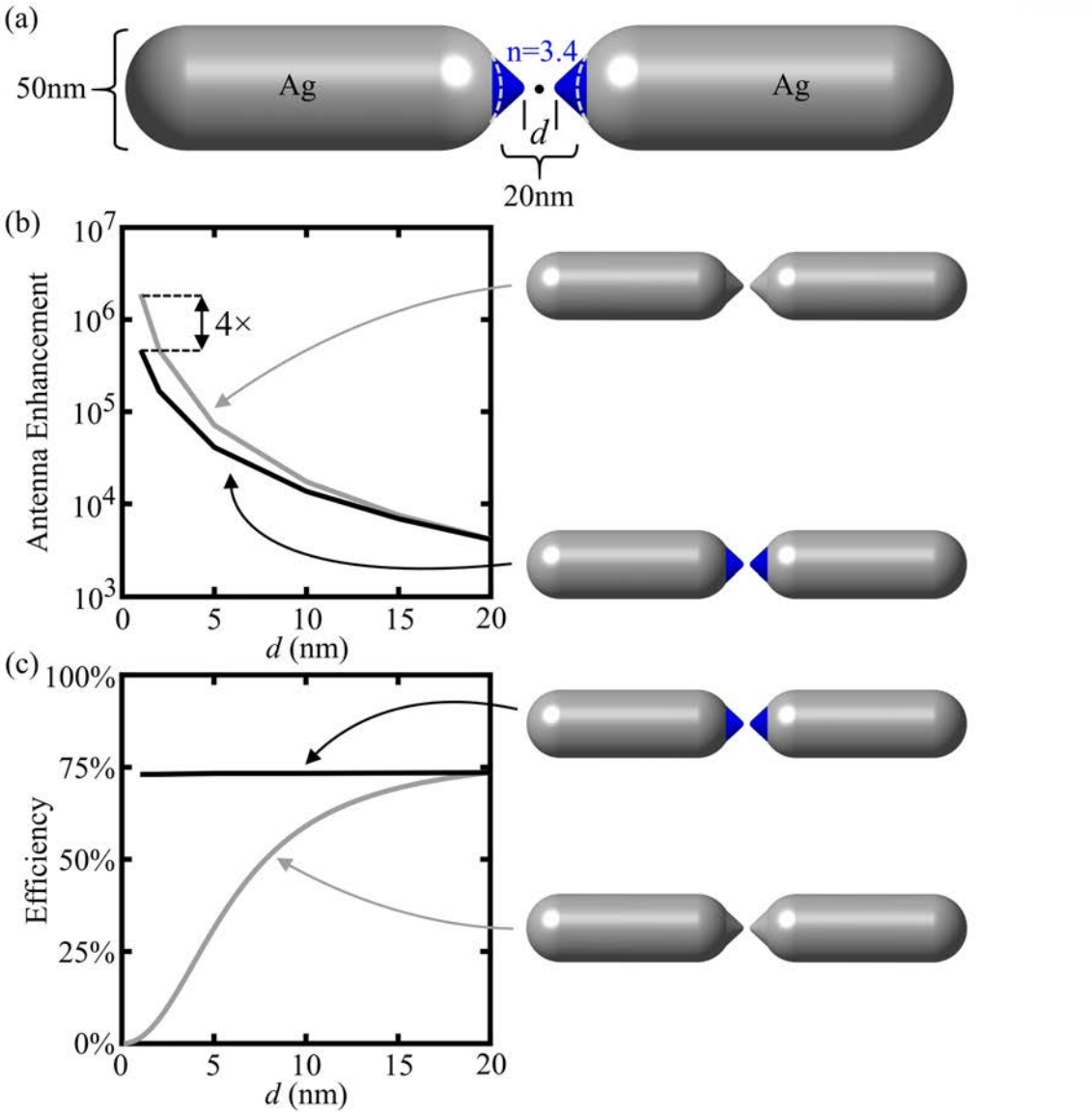

Figure 5.10: The metal-dielectric antenna uses sharp dielectric tips to maintain high efficiency with little compromise to the enhancement factor. (a) Metal-dielectric dipole antenna. This antenna is similar to the all-metal antenna in Fig. 5.9(a) except the sharp metal tips have been replaced with sharp dielectric tips of refractive index n=3.4 (minimum radius of curvature = 1nm, cone angle = 90°). (b) FDTD calculation of the enhancement of the metal-dielectric antenna (black line) compared to the all-metal antenna (silver line) as a function of d at a wavelength of λ=1550nm. (c) Efficiency of the metal-dielectric antenna compared to the all-metal antenna as function of d. The efficiency of the all-metal antenna was calculated using the circuit model including the surface collision effect (Fig. 5.9(c)). The efficiency of the metal-dielectric antenna was calculated in FDTD with a correction for surface collisions.

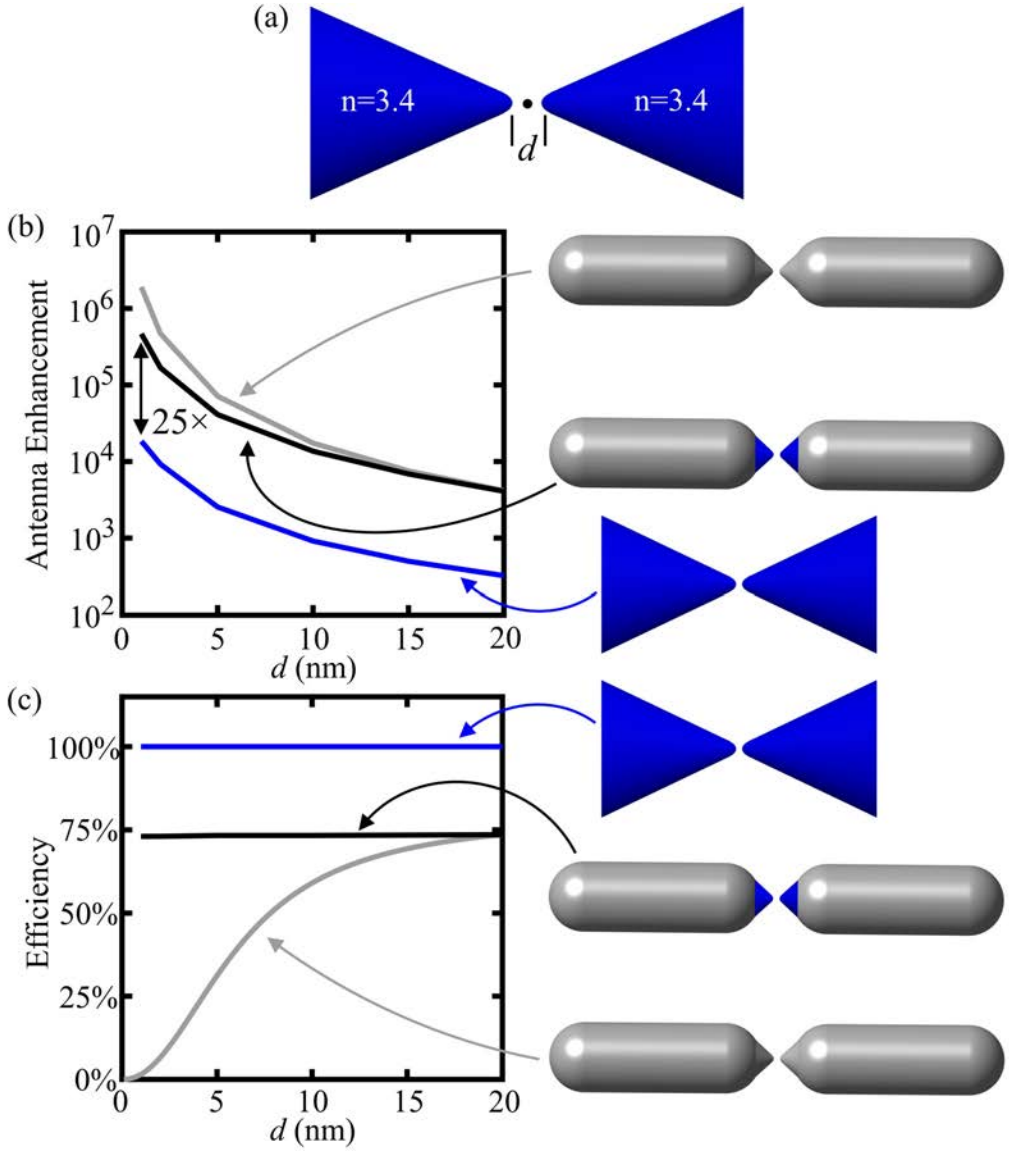

Figure 5.11: All-dielectric bowtie antenna provides insufficient enhancement compared to the all-metal and metal-dielectric variants. (a) All-dielectric bowtie antenna. The antenna consists of two opposing cones with a center vacuum gap of width d (minimum radius of curvature = 1nm, cone angle = 90°). (b) Comparison of the antenna enhancement provided by the all-dielectric bowtie (blue line) with the all-metal (silver line) and metal-dielectric (black line) antennas as a function of d. (c) Efficiency of the all-dielectric bowtie with comparison to the all-metal and metal-dielectric antennas as a function of d. The dielectric antenna is lossless.

circuit model given by [30] using numerical coefficient β=0.5 and 20nm metal-to-metal spacing. We find that the metal-dielectric antenna maintains ultra-high efficiency with little compromise to the peak enhancement at small dielectric tip spacing d. For d=1nm, the metal-dielectric antenna reaches a peak radiation enhancement of 5×10^5. Although 4 times less enhancement than the all-metal antenna with the same dimensions, the corresponding metal-dielectric antenna efficiency is 70% versus 2% for the all-metal antenna. Note that if the dielectric cones in Fig. 5.10(a) are removed (e.g. refractive index n=1), the efficiency of the resulting antenna is approximately unchanged, but the enhancement is reduced dramatically. Indeed, compared to the metal-dielectric antenna with d=1nm, the enhancement is reduced by over $100\times$ in simulation. Therefore, from an alternative perspective, the metal-dielectric antenna tips boost enhancement without changing the antenna efficiency.

Given the clear efficiency improvement provided by nanoscale dielectrics, to what degree is some metal required for optimal light concentration? To address this question, we investigated the enhancement offered by an all-dielectric antenna in comparison to the all-metal and metal-dielectric antennas in Fig. 5.9(a) and Fig. 5.10(a). Consider the all-dielectric bowtie antenna in Fig. 5.11(a). This antenna consists of two opposing dielectric cones of refractive index n=3.4 (cone angle = 90°) with a small vacuum gap of width d at the center. The length of the bowtie is chosen to be 680nm in order to tune the fundamental resonance wavelength to 1550nm. The antenna enhancement (Eq. 5.1) and efficiency (Eq. 5.2) are shown in Fig. 5.11(b) and Fig. 5.11(c) respectively. As depicted here, the enhancement of the dielectric antenna increases with decreasing vacuum gap d, similar to the all-metal and metal-dielectric antennas. For d=1nm, the antenna enhancement peaks at 1.8×10^4, which is 25 times smaller than the metal-dielectric antenna enhancement. The corresponding all-dielectric bowtie's efficiency improves to 100% versus 70% for the metal-dielectric antenna. From this analysis, we can conclude that although the all-dielectric antenna can provide high efficiency, some metal in the antenna design drastically improves the enhancement and is therefore beneficial. Conversely, the all-metal antenna provides high enhancement, but at very poor efficiency.

Up to this point we have demonstrated that (1) purely metallic antennas suffer from an efficiency versus enhancement tradeoff due to Ohmic losses, which are worse than typically predicted because of the anomalous skin effect; (2) this tradeoff can be mitigated by including dielectrics in the antenna design, thus enabling high efficiency and high antenna enhancement simultaneously; and (3) a purely dielectric antenna design is efficient, but does not offer comparable enhancement to the metal-dielectric and all-metal designs. Going forward, we will show how the metal-dielectric antenna design strategy can be applied to semiconductor spontaneous emission enhancement.

Section 4: Metal-dielectric antenna-LED

The antennas discussed in the previous sections use a light source in vacuum, but communications applications require an electrically-injected light source such as a semiconductor. The optical antenna-enhanced light emitting diode (antenna-LED) emits from a semiconductor [37, 4, 36]. Consider the metal-dielectric antenna-LED depicted in Fig. 5.12(a). The structure is similar to the metal-dielectric antenna given in Fig. 5.10(a) except now the two inner tips are connected by a semiconductor bridge of width b at the center. The bridge is composed of a material with refractive index n=3.4, which is similar to the refractive index of many III-V semiconductors. By contrast, we have also investigated the antenna-

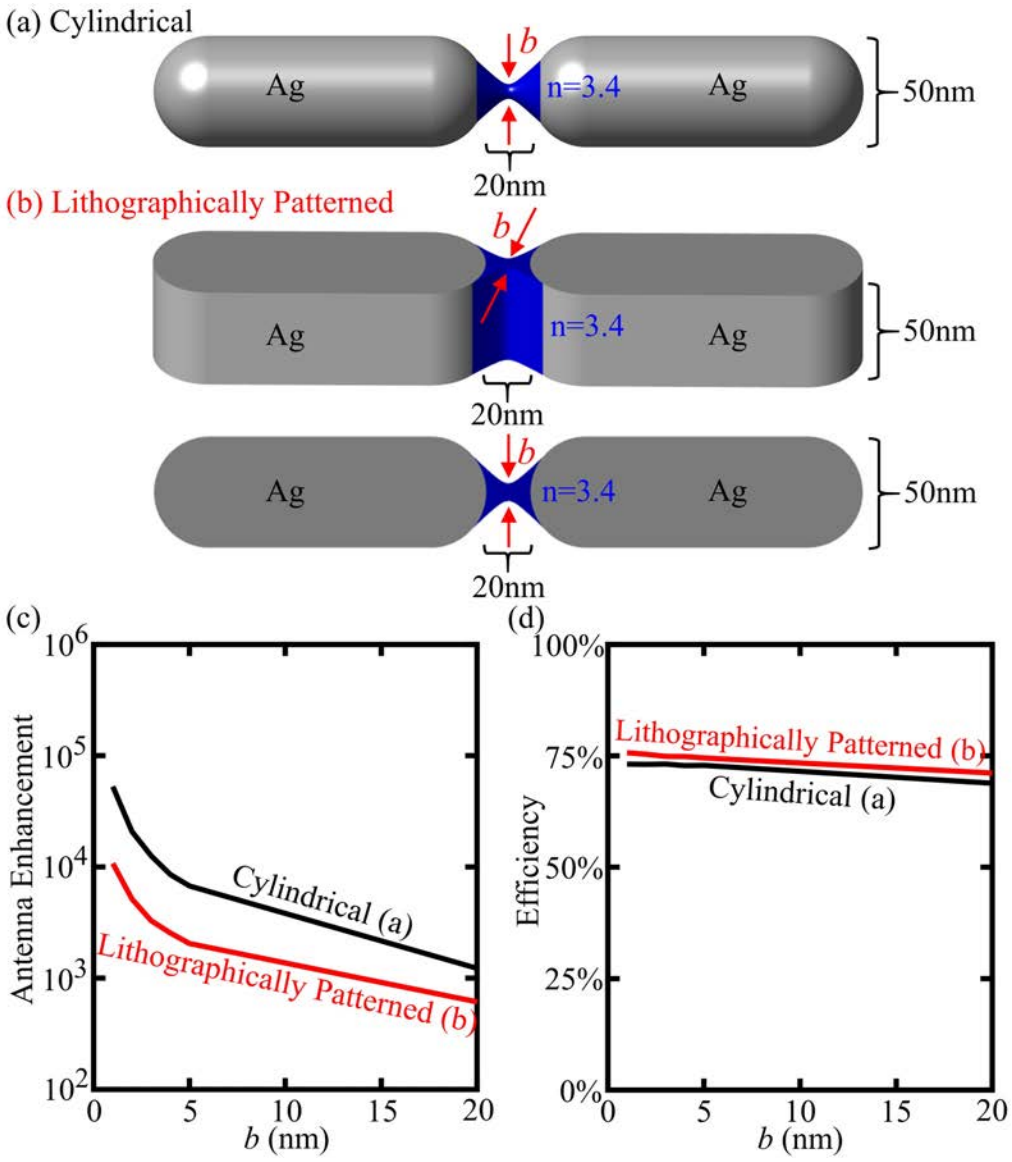

Figure 5.12: Continuous semiconductor bridge antenna provides efficient enhancement for electrically-injected semiconductor devices. (a) Metal-dielectric antenna-LED with cylindrical symmetry. The structure is similar to that in Fig. 5.10(a) except the sharp dielectric tips have been connected by a bridge of width b. Perspective view ((b), upper graphic) and top view ((b), lower graphic) of a metal-dielectric antenna-LED that is compatible with top-down semiconductor fabrication. This antenna has the same cross-section as the antenna in (a), but the cross-section is extruded 50nm in depth. (c) Peak spontaneous emission enhancement as a function of bridge width b calculated in FDTD. (d) Efficiency of the antennas as a function of bridge width b calculated in FDTD and corrected for the surface collision effect.

LED in Fig. 5.12, denoted "lithographically patterned" in the sense that this structure could be fabricated using two-dimensional top-down processing. Both antennas assume the optical point source is located at the center of the bridge.

As shown in Fig. 5.12(c)-(d), there is a marked increase in antenna enhancement as the b dimension is reduced below 10nm without compromise to the antenna efficiency; this is yet another version of dielectric focusing, similar to that shown before in Fig. 5.10 and Fig. 5.11. Note that the enhancement reference according to Eq. 5.1 is a point dipole source radiating into vacuum. To obtain the enhancement of a reference dipole source in a homogeneous semiconductor, divide the value in Fig. 5.12(c) by n=3.4. Furthermore, the antenna enhancement and efficiency were calculated at a resonance wavelength of λ=1550nm for all values of b. For small b, real semiconductor optical emitters may undergo a blueshift from their nominal bandgap energy. The antenna resonance can be tuned to accommodate this effect by decreasing the total antenna length.

In addition to improved light concentration, a small semiconducting bridge can provide favorable polarization selection rules. Electron-to-heavy hole (C-HH) radiative transitions are preferentially stimulated by an electric field polarized along the long direction of a semiconducting quantum wire, perpendicular to the confinement direction [21, 22]. Thus, the polarization of band-edge light emission from an unstrained semiconductor bridge is in favorable alignment with the metal-dielectric antenna-LED mode polarization [36]. The combination of all these benefits indicates that antenna-enhanced, efficient, electrically-injected antenna LED devices are feasible. Such a device could serve as a nanoscale optical source for ultra-fast, low-power, on chip data communications.

Section 5: Antenna enhancement versus Purcell enhancement

In Sections 2-4 we showed metal-dielectric antennas that can provide high efficiency and high antenna enhancement of both atoms and semiconductors. Up to this point we have only employed the antenna enhancement metric defined in Eq. 5.1 without invoking the Purcell Effect [112], which has been emphasized by many prior investigators. While we advocate that the Purcell Effect is not needed to describe antenna properties, we show how it may be employed in the context of antenna enhancement.

An antenna can concentrate zero-point electromagnetic energy into a sub-wavelength volume, thus enhancing the spontaneous emission rate over the normal vacuum emission. This fact is reflected in the Purcell enhancement factor, defined as,

$$\text{Purcell Factor} = \frac{1/\tau_{\text{enhanced}}}{1/\tau_{\text{o}}} = \frac{3}{4\pi^2} Q \frac{\lambda^3}{V_{\text{eff}}} \tag{5.3}$$

where τ_{o} is the radiation lifetime of a free-space dipole, τ_{enhanced} is the lifetime of a dipole radiating into an optical cavity or antenna mode (not necessarily into free space, which is critical when considering lossy antennas), λ is the wavelength, Q is the quality factor, and V_{eff} is the effective mode volume. Customarily V_{eff} is defined by [64],

$$V_{\text{eff}} \equiv \frac{\int_0^{r'} \text{Re}\left[\frac{\partial(\varepsilon\omega)}{\partial\omega}\right] |E|^2 d^3r}{\varepsilon |E|^2_{\text{peak}}} \tag{5.4}$$

where the integral in the numerator represents the total energy in the antenna mode (corrected for potential material dispersion [55, 72]), and the denominator is the peak energy density in the antenna mode. Eq. 5.4 requires a full antenna electromagnetic analysis, but in that case the electromagnetic analysis can provide all antenna properties and the Purcell enhancement factor is not needed (as demonstrated in the previous sections and, for example, in [30, 67]). While the effective mode volume can be estimated for dielectric cavities, it is unclear how to obtain a suitable estimate for antennas. In this section we will derive the appropriate antenna effective mode volume, V_{eff}, to insert into the Purcell factor by comparing the enhancement predicted by antenna theory versus the Purcell effect.

To estimate antenna enhancement, an engineer would use the circuit representation of a dipole antenna [30]; a simple version of the metallic dipole antenna was shown in Fig. 5.9(a)-(b). The antenna enhancement factor $(P_{\text{rad}}/P_{\text{o}})$ can be written as [7],

$$F = \left(\frac{l}{d}\right)^2 \tag{5.5}$$

A consequence of Eq. 5.5 is that the maximum antenna power is attained for the longest single-mode resonant antenna (namely, the half-wave antenna). Note that Eq. 5.5 applies generally to one-dimensional metallic antennas with high conductivity. Antennas with arbitrary geometrical configurations, including the metal-dielectric antenna discussed in Fig. 5.10, require a more detailed treatment.

In contrast with Eq. 5.5, the Purcell Factor, Eq. 5.3, is repeated here:

$$F = \frac{3}{4\pi^2} Q \frac{\lambda^3}{V_{\text{eff}}} \tag{5.6}$$

By equating Eq. 5.5 & Eq. 5.6 in Eq.5.7, we may obtain the effective mode volume:

$$\left(\frac{l}{d}\right)^2 = \frac{3}{4\pi^2} Q \frac{\lambda^3}{V_{\text{eff}}} \tag{5.7}$$

Rearranging to solve for V_{eff} and combining terms, we find:

$$V_{\text{eff}} = \frac{3}{4\pi^2} Q d^2 \lambda \left(\frac{\lambda}{l}\right)^2 \tag{5.8}$$

where the only remaining unknown is the quality factor, Q. If the antenna loss is limited primarily by radiation and not resistance, the Q may be obtained from the well-established Wheeler-Chu Limit [140, 20],

$$Q \geq \frac{3}{4\pi^2} \left(\frac{\lambda}{l}\right)^3 \tag{5.9}$$

[7] This formula was provided in the previous chapter from circuit theory, Eq.4.11.

where l is the size of the longest antenna dimension (in this case, the antenna length) and λ is the wavelength. Notably, the quality factor increases rapidly when the antenna length is very small, but in antennas we want a low Q representing efficient radiation. Combining Eq. 5.8 with the lower bound of Eq. 5.9 we may obtain the effective mode volume of a Wheeler-Chu limited antenna:

$$V_{\text{eff}} = \left(\frac{3}{4\pi^2}\right)^2 d^2\lambda \left(\frac{\lambda}{l}\right)^5 \tag{5.10}$$

which has no unknowns. An interesting special case of Eq. 5.10 is the half-wave dipole antenna; plugging in $l \to \lambda/2$:

$$\text{Halfwave Dipole } V_{\text{eff}} \text{ Limit} = 0.185 \cdot d^2\lambda \tag{5.11}$$

which represents a bound on the single-mode antenna effective mode volume. In principle, the vacuum gap width d can be as small as 1nm, and therefore the effective mode volume of antennas may be extremely small. Note that based on our derivation, Eq. 5.10 and Eq. 5.11 apply to one-dimensional purely metallic antennas, such as that depicted in Fig. 5.9(a). Antennas of arbitrary geometry may require a full numerical electromagnetic analysis. In the next section we will check the half-wave dipole antenna effective mode volume formula, Eq. 5.11, against full numerical calculations using the customary formula (Eq. 5.4).

Section 6: Electromagnetic numerical calculations of antenna effective mode volume

The effective mode volume that is used for electromagnetic numerical calculation was given above in Eq. 5.4 and is reproduced here:

$$V_{\text{eff}} \equiv \frac{\int_0^{r'} \text{Re}\left[\dfrac{\partial(\varepsilon\omega)}{\partial\omega}\right] |E|^2 d^3r}{\varepsilon|E|^2_{\text{peak}}} \tag{5.12}$$

A detailed discussion of the full-wave calculation using Eq. 5.12 may be found in *Appendix: Effective Mode Volume*.

We considered three antennas for numerical calculation, depicted in Fig. 5.13. The all-metal, metal-dielectric, and all-dielectric antennas refer to the antennas from Fig. 5.9(a), Fig. 5.10(a) and Fig. 5.11(a) respectively with vacuum gap widths of d=1nm between respective metallic or dielectric tips and minimum radius of curvature of 1nm. The antenna effective mode volume (Fig. 5.13, x-axis), is normalized by wavelength and inverted (λ^3/V_{eff}) so that it may easily be plugged into the Purcell factor (Eq. 5.3). Antenna efficiency (Fig. 5.13, y-axis) was obtained previously in Sections 2-4. The electromagnetically calculated effective mode volume values of the all-dielectric, metal-dielectric, and all-metal antennas were $5.6 \times 10^{-6}\lambda^3$, $7.8 \times 10^{-7}\lambda^3$, and $1.5 \times 10^{-7}\lambda^3$ at λ=1550nm respectively. Note that if we were to consider a larger or more practical tip parameter d in our calculation, the efficiency of the metallic antenna and the effective mode volume of all three antennas would increase. For example, if we used d=2nm, the effective mode volume for the three antennas would increase by approximately $4\times$.

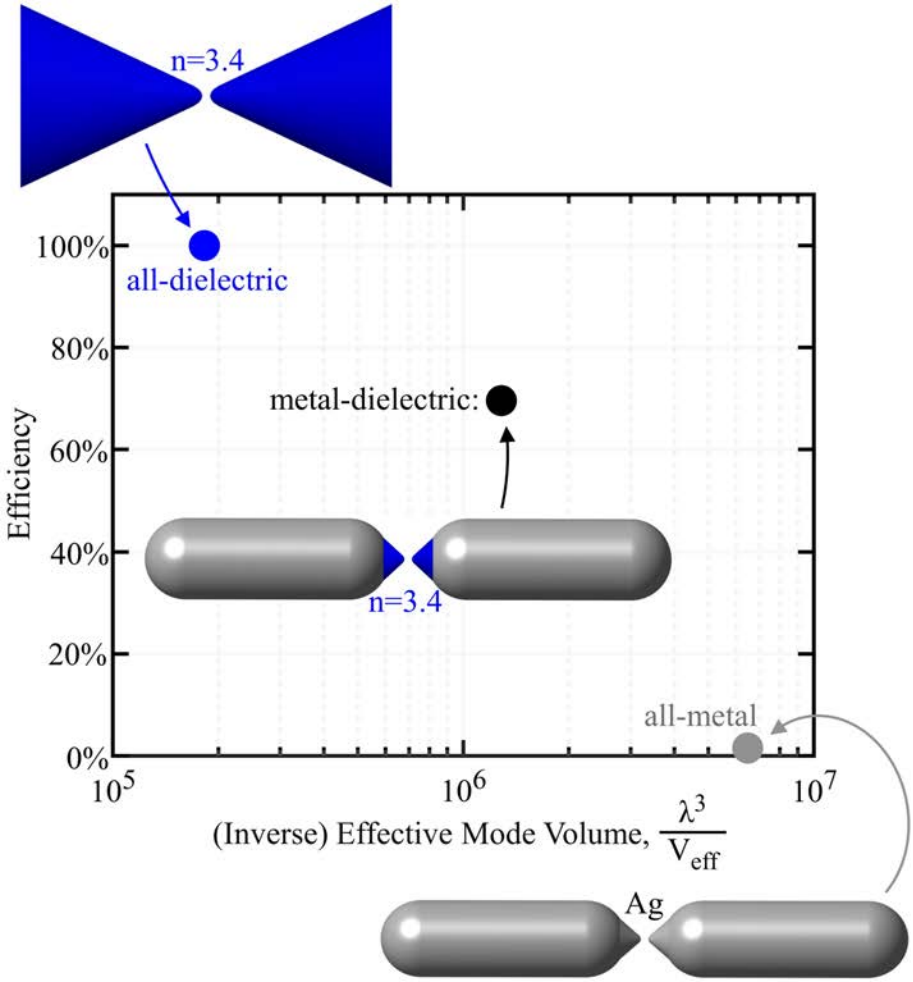

Figure 5.13: The efficiency and effective mode volume of three antennas from Fig. 5.9(a), Fig. 5.10(a), and Fig. 5.11(a) with vacuum gap widths of d=1nm and radius of curvature = 1nm are plotted.

The numerically calculated effective mode volume of the all-metal antenna ($1.5 \times 10^{-7}\lambda^3$) agrees within a factor 2 to the predicted value of the half-wave antenna effective mode volume from Eq. 5.11 using d=1nm: $V_{\text{eff}} = 0.185 \cdot d^2\lambda = 7.7 \times 10^{-8}\lambda^3$. The small disagreement between the two values may be attributed to inaccuracy in the numerical V_{eff} calculation (discussed in *Appendix: Effective Mode Volume*) or to our choice of the canonical lower bound of the Wheeler-Chu limited quality factor from Eq. 5.9 corresponding to Q=0.6 for a half-wave antenna, whereas a more realistic antenna quality factor is Q=1. Furthermore, compare the effective mode volume of the all-dielectric antenna ($5.6 \times 10^{-6}\lambda^3$) to MIT's photonic crystal cavity from [19] with $V_{\text{eff}} = 7.0 \times 10^{-5}\lambda^3$ at λ=1550nm, which reportedly uses a similar dielectric tip geometry chosen here except their structure was optimized for both high quality factor and low effective mode volume. These results indicate that antennas are capable of both extreme concentration of electromagnetic energy and good radiation. Notably, the metal-dielectric antenna provides the best balance of efficient radiation (70%) and effective mode volume ($7.8 \times 10^{-7}\lambda^3$), complimentary to our results from prior sections.

Section 7: Conclusion

In this work we have introduced a metal-dielectric antenna that eliminates the tradeoff of enhancement versus efficiency present in purely metallic optical antennas. We proposed a feasible structure for a metal-dielectric antenna-enhanced light-emitting diode that could lead to improvements in the speed and efficiency of nanoscale light sources for optical communications. By comparing antenna enhancement and the Purcell effect, we have introduced a new formula for antenna effective volume that permits continued use of the Purcell enhancement factor for metallic antennas.

Appendix: Effective mode volume

There are several methods in the literature to numerically calculate the effective mode volume of optical antennas using forms similar to Eq. 5.12 [64, 116, 69, 54, 66]. For this paper we used a variation on the method proposed originally in [64]. It should be remarked that calculating the effective mode volume of antennas is more nuanced than doing so for high-Q dielectric resonators, and our calculation involves heuristics that may be less accurate than strongly-mathematically motivated solutions in the literature [116, 69]. Nevertheless, we will briefly outline the challenges of performing this calculation, and how they were addressed in this work. All simulations of effective mode volume were performed in Lumerical FDTD with 0.25nm resolution in the antenna feedgap, 1nm resolution of the rest of the antenna, and 10nm resolution of the rest of the simulation domain which was $(2\mu\text{m})^3$ in total.

There are four points to address in calculation of the numerator of Eq. 5.12:

1. Antennas are leaky, or in other words, it is difficult to distinguish the mode that is bound to the antenna and radiation in the far-field.

2. The energy of the simulation source (e.g. a plane wave) must also be distinguished from the antenna mode energy.

3. The excited antenna may consist of both dark (nonradiative) and bright (radiative) modes.

4. Metals are dispersive.

Point (1) was addressed originally in [64] and is out of the scope of this appendix. Essentially, one may find a transition between the local antenna mode and the far-field radiation by self-consistently choosing the radius of integration, r', in Eq. 5.12. Since far-field antenna radiation intensity falls off as $1/r^2$, there is a well-defined transition.

Point (2) was addressed in two ways: (a) A total-field scattered-field source (TFSF) source was used, which minimizes the total footprint of the incident plane wave source in simulation. (b) Time-apodization in the Fourier transform of the FDTD data was employed. Time-apodization is essentially a long-pass filter for time, which allows one to effectively filter out an incident broadband source pulse. Thus, the simulation will only capture electric field data from the resonant antenna mode that continues to oscillate after the source has died out. Because antennas have very low quality factor, this process is not perfect and it is difficult to completely decouple the source from the antenna response. Nevertheless, testing with different apodization cutoffs tended to provide convergence of the energy integral for sufficiently long cutoff after the source pulse.

Point (3) is believed to be addressed by the same procedures as Point (2) above. Because an incident plane wave (TFSF) source was used, dark modes should not have been excited in simulation. This was confirmed by measuring the antenna scattering efficiency, which was much greater than the radiative efficiency of the antennas as excited by point dipole sources. Furthermore, dark modes most likely have smaller Q than the antenna radiative mode, and therefore time-apodization filtered them out. Since we have removed dark modes from simulation, the effective mode volume values reported in Section 6 correspond only to the antenna radiative mode, and therefore may be properly compared to the antenna effective mode volume formula derived in Eq. 5.10 & Eq.5.11.

Point (4) is addressed by using the dispersion correction term, Re $[\partial(\varepsilon\omega)/\partial\omega]$, in Eq. 5.12 [55, 72]. Palik's data for silver [108] was interpolated to obtain this term.

The peak energy density term in the denominator of Eq. 5.12 was calculated at the center of the antenna feedgap, after time-apodization. This is the correct choice because we were interested in the enhanced radiation of dipole point sources from this location. This differs slightly from the "true" location of peak energy density, which is very close to the metal boundary.

5.3 Conclusion

This chapter provided two published papers on optical antenna-enhanced spontaneous emission. In Section 5.1 we designed a cavity-backed slot antenna-LED that was capable of 94% coupling efficiency to a single-mode waveguide, potentially enabling efficient integrated optical interconnects. This was achieved using inverse design optimization, which will be described in the next chapter. In Section 5.2 we showed how enhancement can be boosted in optical antennas without compromising efficiency by using sharp dielectric tips in the antenna gap. By overcoming this efficiency barrier to enhancement, metal-dielectric antenna-LEDs could one day demonstrate extremely fast spontaneous emission carrier lifetime – perhaps testing the ultimate limits of efficient direct modulation.

Chapter 6

Inverse Electromagnetic Design via the Adjoint Method

In the previous chapters we have discussed the exciting prospects of on-chip optical interconnects using optical antenna-LEDs. From this chapter forward we switch topics and will discuss inverse electromagnetic design. There are several adept methods to solve for electromagnetic device characteristics and properties, to varying degrees of accuracy or approximation. For example, for complex nano-photonic devices, one may solve Maxwell's Equations directly using simulation tools like finite-difference time-domain (FDTD) or the finite-element method (FEM). Or, perhaps one is only interested in solving for a particular optical response from a periodic structure using rigorous-coupled wave analysis (RCWA). Maybe near-field electromagnetic characteristics are unnecessary, and one only desires to solve the Fresnel diffraction equation. These are each examples of methods that fall under the so-called "forward problem" in electromagnetics, where we computationally solve for a response or electromagnetic property that results from a given electromagnetic structure. By contrast, inverse design refers to solving the "inverse problem", which indicates that we solve for an electromagnetic structure given a desired response or electromagnetic property [1]. The inverse problem is generally difficult to solve for complex tasks, but by employing a special technique called the adjoint method we can greatly improve our design capabilities and speed of computation. In the next three sections we will describe the motivation and theory behind inverse design via the adjoint method. In Section 6.4 we will show how inverse design can be applied to fabrication-friendly vertical grating couplers. This work is derived nearly verbatim from the published work in [47], with citation reproduced here:

> S. Hooten, T. V. Vaerenbergh, P. Sun, S. Mathai, Z. Huang, and R. G. Beausoleil, "Adjoint Optimization of Efficient CMOS-Compatible Si-SiN Vertical Grating Couplers for DWDM Applications," Journal of Lightwave Technology, vol. 38, no. 13, pp. 3422–3430, Jul. 2020.

Please note that the appendices referred to in Section 6.4 are provided within the section and not in the main appendices of this thesis.

[1] In other words, inverse design is technically just synonymous with computational design of electromagnetic structures. However, it recently has become a buzzword referring to special design techniques

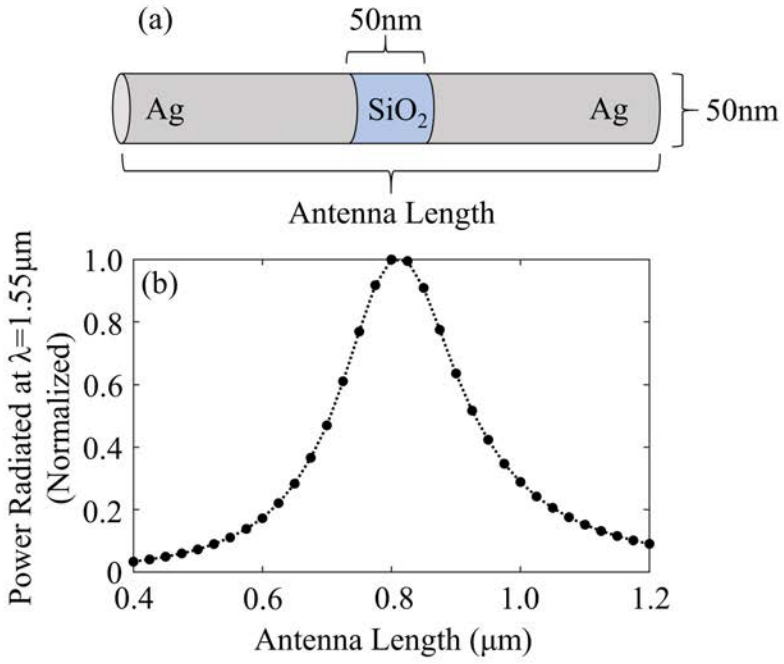

Figure 6.1: Parameter sweep, or exhaustive search, is the conventional method to design an electromagnetic device. Here we show a simple example of sweeping the length of an antenna (a) to maximize the radiated power at $\lambda = 1550$nm. Each point in (b) corresponds to an individual simulation using FDTD.

6.1 The inefficiency of exhaustive search

The most ubiquitous and conventional method to design electromagnetic structures is by parameter sweep. In other words, an engineer may use their physical intuition to design a structure that performs a desired function then tune a few parameters by sweeping them in simulation to optimize that structure's properties. A very simple example of this is shown in Fig 6.1. Here we sweep the length of a cylindrical antenna, with a 50nm diameter and 50nm glass gap where a point source is placed, in order to maximize the radiated power at a wavelength of 1550nm. Each point in Fig 6.1 corresponds to an individual simulation. As one might expect, we find that the radiated power is maximized when the antenna length is approximately equal to half a wavelength $L \approx \lambda/2$, corresponding to the fundamental antenna mode.

In this example, no special design techniques were necessary. However, one can conceive of much more difficult problems with many design parameters (potentially hundreds or thousands), where the underlying physics are much more difficult to ascertain. For example, suppose we were to co-optimize the antenna length, diameter, gap dimension, and gap material in Fig 6.1 simultaneously. The best combination of all these parameters is much harder to directly intuit. And furthermore, as the number of parameters increases, the number of simulations that are required to perform an (exhaustive) parameter

sweep increases exponentially by the following relation:

$$\begin{aligned} \text{Exhaustive Search Parameter Sweep: } S &= m^n \\ S &\equiv \text{Number of simulations} \\ m &\equiv \text{Number of sweeped values per parameter} \\ n &\equiv \text{Number of design parameters} \end{aligned} \tag{6.1}$$

Indeed, if we were to sweep 10 values of the 4 parameters mentioned above for our simple antenna example, this already amounts to 10,000 simulations. In some cases individual Maxwell simulations of complex electromagnetic structures may take minutes or hours even on research-grade high-performance servers. Thus, parameter sweeps are infeasible for complex tasks and a better method is needed. We will address this problem using inverse electromagnetic design via the adjoint method.

6.2 Formal description of electromagnetic design and optimization

The description of the adjoint method requires some mathematical rigor in the fields of optimization and linear algebra. In this section we will provide a brief primer on the required knowledge and notation. In the first step, we will show how Maxwell's Equations may be rewritten as a matrix equation. Then, we will describe formal optimization notation and a description of a generalized electromagnetic design problem. Finally, we will describe an optimization technique called gradient descent.

Maxwell's equations in matrix notation

For electromagnetic design problems, we will consider Ampere's Law and Faraday's Law in Eq. 6.2 and Eq. 6.3 respectively:

$$\nabla \times \mathbf{H} = \mathbf{J_e} + \frac{\partial \mathbf{D}}{\partial t} \tag{6.2}$$

$$\nabla \times \mathbf{E} = \mathbf{J_m} - \frac{\partial \mathbf{B}}{\partial t} \tag{6.3}$$

where $\mathbf{H}$ is the magnetizing field, $\mathbf{B}$ is the magnetic field, $\mathbf{E}$ is the electric field, and $\mathbf{D}$ is the displacement field. $\mathbf{J_e}$ and $\mathbf{J_m}$ are current source terms referring to electric current and "magnetic current" respectively. Magnetic current is unknown to exist in nature, but can be occasionally useful in electromagnetic simulation. Note that we need not consider Gauss's Laws because in most cases they are automatically satisfied by Ampere's Law and Faraday's Law.

The formulation of the adjoint method that we will be considering will require two simplifying assumptions: (1) time-harmonic fields and (2) linear materials [2]. Assumption (1) allows us to write the field

[2] It is possible to formulate the adjoint method without these assumptions [32, 52]. Nevertheless, many design problems fall under this umbrella, and the simplicity of the adjoint method is most clearly seen this way.

and source quantities as complex phasors with explicit exponential time dependence:

$$\begin{aligned} \mathbf{E} &= \widetilde{\mathbf{E}} e^{-j\omega t} & \mathbf{H} &= \widetilde{\mathbf{H}} e^{-j\omega t} \\ \mathbf{D} &= \widetilde{\mathbf{D}} e^{-j\omega t} & \mathbf{B} &= \widetilde{\mathbf{B}} e^{-j\omega t} \\ \mathbf{J_e} &= \widetilde{\mathbf{J}}_\mathbf{e} e^{-j\omega t} & \mathbf{J_m} &= \widetilde{\mathbf{J}}_\mathbf{m} e^{-j\omega t} \end{aligned} \tag{6.4}$$

where the $\widetilde{\cdot}$ quantities are complex phasors defined at frequency ω that include spatial amplitude and phase information. Assumption (2) allows us to simplify the magnetic and displacement field terms:

$$\begin{aligned} \widetilde{\mathbf{B}} &= \mu \widetilde{\mathbf{H}} \\ \widetilde{\mathbf{D}} &= \varepsilon \widetilde{\mathbf{E}} \end{aligned} \tag{6.5}$$

where μ and ε are the (complex, spatially-distributed) permeability and permittivity at frequency ω. Note that these may be tensor quantities, but for simplicity we will assume isotropic materials. Using Eq. 6.4 and Eq. 6.5 we may simplify Maxwell's Equations from Eqs. 6.2-6.3 above:

$$\nabla \times \widetilde{\mathbf{H}} = \widetilde{\mathbf{J}}_\mathbf{e} - j\omega\varepsilon \widetilde{\mathbf{E}} \tag{6.6}$$

$$\nabla \times \widetilde{\mathbf{E}} = \widetilde{\mathbf{J}}_\mathbf{m} + j\omega\mu \widetilde{\mathbf{H}} \tag{6.7}$$

After some rearrangement we find:

$$j\omega\varepsilon \widetilde{\mathbf{E}} + \nabla \times \widetilde{\mathbf{H}} = \widetilde{\mathbf{J}}_\mathbf{e} \tag{6.8}$$

$$\nabla \times \widetilde{\mathbf{E}} - j\omega\mu \widetilde{\mathbf{H}} = \widetilde{\mathbf{J}}_\mathbf{m} \tag{6.9}$$

Observe that this may be rewritten as a matrix equation by factoring out $\widetilde{\mathbf{E}}$ and $\widetilde{\mathbf{H}}$:

$$\begin{bmatrix} j\omega\varepsilon & \nabla\times \\ \nabla\times & -j\omega\mu \end{bmatrix} \begin{bmatrix} \widetilde{\mathbf{E}} \\ \widetilde{\mathbf{H}} \end{bmatrix} = \begin{bmatrix} \widetilde{\mathbf{J}}_\mathbf{e} \\ \widetilde{\mathbf{J}}_\mathbf{m} \end{bmatrix} \tag{6.10}$$

To which we may define the following quantities:

$$\mathbf{A} \equiv \begin{bmatrix} j\omega\varepsilon & \nabla\times \\ \nabla\times & -j\omega\mu \end{bmatrix}, \quad \mathbf{x} \equiv \begin{bmatrix} \widetilde{\mathbf{E}} \\ \widetilde{\mathbf{H}} \end{bmatrix}, \quad \mathbf{b} \equiv \begin{bmatrix} \widetilde{\mathbf{J}}_\mathbf{e} \\ \widetilde{\mathbf{J}}_\mathbf{m} \end{bmatrix} \tag{6.11}$$

allowing us to rewrite Eq. 6.10 in a convenient linear algebra notation:

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{6.12}$$

where $\mathbf{x}$ is a vector representing the electric and magnetic fields, $\mathbf{b}$ represents the electric and magnetic current sources, and $\mathbf{A}$ is a Maxwell operator that defines the physics of Maxwell's equations subject to the simulation materials and geometry (represented by ε and μ). Eq. 6.12 allows us to easily describe what

a Maxwell simulation's function is. In particular, a solution to Maxwell's Equations, or the "forward problem" amounts to:

$$\text{Forward Problem: Solve } \mathbf{Ax} = \mathbf{b} \text{ for } \mathbf{x} \text{ given } \mathbf{A} \text{ and } \mathbf{b} \tag{6.13}$$

in other words, we solve for the electric and magnetic fields that result from provided sources and materials[3].

Electromagnetic design as an optimization problem

Now that we have defined a simple shorthand for Maxwell's equations in Eq. 6.12, we may write out a formal description of a design problem. In inverse electromagnetic design, we are typically interested in solving for a merit function that is explicitly a function of the electric and magnetic fields. In other words, we'd like to optimize some merit function $f(\mathbf{x})$, where $\mathbf{x}$ is the electric and magnetic fields and f is some (scalar) function of those fields, $f : \mathcal{C}^m \rightarrow \mathcal{R}$ where m is the dimension of vector $\mathbf{x}$. f could represent any arbitrary figure of merit including electric field intensity, optical absorption, waveguide coupling efficiency, and more. In our antenna length example from Fig. 6.1, the merit function was the power radiated from the antenna, which can be expressed as the time-averaged Poynting vector integrated along a surface enclosing the antenna. Moreover, a general design problem consists of n parameters, or design elements, that we would like to tune (e.g. length, width, material value, etc.). Let,

$$\mathbf{p} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_{n-1} \\ p_n \end{bmatrix} \tag{6.14}$$

be a vector of such parameters [4].

We are interested in either minimizing or maximizing our merit function f. Since maximizing f is equivalent to minimizing the negative of f, we may without loss of generality express the inverse problem as a minimization problem with:

$$\text{Inverse Problem: } \min_{\mathbf{p}} f(\mathbf{x}), \text{ s.t. } \mathbf{Ax} = \mathbf{b} \tag{6.15}$$

where this equation reads as "minimize with respect to parameters $\mathbf{p}$, the electromagnetic merit function $f(\mathbf{x})$ that is subject to Maxwell's Equations $\mathbf{Ax} = \mathbf{b}$". In other words, $\mathbf{p}$ is an optimization variable that we can vary, $f(\mathbf{x})$ is the function we'd like to minimize, and $\mathbf{x}$ is required to satisfy Maxwell's equations (in order to be physically meaningful).

[3] The forward problem can also be thought of as taking the inverse of $\mathbf{A}$, e.g. $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$, but this isn't typically how linear equations of this type are solved. In fact, Eq. 6.13 is just a representation of the forward problem and most electromagnetic solvers (e.g. finite-difference time-domain, FDTD) do not explicitly solve this equation.

[4] Note, furthermore, that we will assume the spatially-distributed permittivity (ε) and permeability (μ) are smooth functions of our parameter vector, $\mathbf{p}$. This allows us to easily update the Maxwell operator, $\mathbf{A}$.

Eq. 6.15 formalizes our description of the electromagnetic design problem, and will allow us to formulate the Adjoint Method in the next section. However, we will first need to describe an optimization technique called gradient descent.

Optimization by gradient descent

In general, the Inverse Problem in Eq. 6.15 is very hard to solve. We showed previously that one technique to solve Eq. 6.15 was an exhaustive parameter sweep, but that technique rapidly becomes intractable as we increase the resolution of our search or the number of design variables. Fortunately, mathematical optimization theory offers alternative techniques that are vastly more efficient. Perhaps the most ubiquitous technique is called *gradient descent*. As the name implies, this technique uses information about a function's gradient to guide the optimization to an optimum. The gradient is simply the multivariate derivative of a function with respect to its dependent variables (e.g., vector $\mathbf{p}$). In our case, the gradient of f with respect to $\mathbf{p}$ is a vector of partial derivatives:

$$\frac{\partial f}{\partial \mathbf{p}} = \begin{bmatrix} \dfrac{\partial f}{\partial p_1} \\ \vdots \\ \dfrac{\partial f}{\partial p_n} \end{bmatrix} \tag{6.16}$$

Intuitively, the gradient gives information about the slope of a function. When the gradient is very large, it means that we can greatly improve the merit function by adjusting our parameter vector in the direction of the gradient. By contrast, when the gradient is very small or identically zero, we have found a (local) optimum. This constitutes a convergence criterion for gradient descent. Thus, in order to solve the Inverse Problem from Eq. 6.15 (in other words, the minimization of merit function f) we need to follow the negative of the gradient, which points in the direction of steepest descent.

We intend to iteratively update the parameter vector $\mathbf{p}$ that represents our electromagnetic structure. Accordingly, let vector $\mathbf{p}_0$ represent the electromagnetic structure before optimization, and let $\mathbf{p}_i$ represent the i-th optimization step where $i = 1, 2, 3, \ldots$. Using the negative gradient, we may update the parameter vector iteratively in the following way:

$$\text{Gradient Descent Algorithm: } \mathbf{p}_{i+1} = \mathbf{p}_i - \varepsilon \frac{\partial f}{\partial \mathbf{p}_i} \tag{6.17}$$

where ε is called the step-size, which is usually chosen by trial-and-error. After iterating many times with Eq. 6.17, the function will settle to a local optimum where $\partial f / \partial \mathbf{p} \approx 0$. This constitutes the convergence criterion – the condition for ending the optimization. With proper choice of design parameters $\mathbf{p}$, Eq. 6.17 can be very effective at producing good solutions to the inverse problem [5].

[5] In practice we typically use some variation on regular gradient descent (Eq. 6.17), such as the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm. Nevertheless, the idea is conceptually similar to what we have shown here, and gradient calculations will be required.

How may we obtain the gradient of our electromagnetic merit function, f, with respect to the design parameters, $\mathbf{p}$? Perhaps the simplest way to get $\frac{\partial f}{\partial \mathbf{p}}$ is by finite-difference derivative approximation. This involves an approximation of each individual partial derivative from the definition of the gradient in Eq. 6.16. For example, consider the partial derivative with respect to design element p_i: $\frac{\partial f}{\partial p_i}$. The finite-difference partial derivative approximation of this value is given by:

$$\frac{\partial f}{\partial p_i} \approx \frac{f(p_1, ..., p_i + \delta p_i, ..., p_n) - f(p_1, ..., p_i, ...p_n)}{\delta p_i} \tag{6.18}$$

where δp_i is a small perturbation to the i-th design element. However, observe that in order to the the full gradient with n total partial derivatives, we need to perform $n + 1$ simulations per iteration; 1 simulation for the nominal value of f defined at $\mathbf{p} = [p_1, ..., p_n]$, and n simulations for n perturbations to the design elements. Therefore, to perform a full gradient descent optimization, this scheme will require the following number of simuations:

$$\begin{aligned} \text{Finite-difference gradient descent: } S &= c \times (n + 1) \\ S &\equiv \text{Number of simulations} \\ c &\equiv \text{Number of gradient descent iterations} \\ n &\equiv \text{Number of design parameters} \end{aligned} \tag{6.19}$$

This is a massive improvement over the exhaustive search parameter sweep (reduced from an exponential dependence to a simple polynomial dependence on the number of design parameters). However, for potentially 1000's of design parameters and 100's of gradient descent iterations, even this scheme remains intractable for electromagnetic optimization. Can we do better?

6.3 The adjoint method

In the previous section, we introduced a formal mathematical notation to describe an electromagnetic design problem, called the inverse problem. We went on to show that the inverse problem could be solved using gradient descent optimization. However, up to now, the only way to obtain the gradient of a function with respect to its design parameters is by finite-difference. While this method greatly outperforms exhaustive search methods like parameter sweeps, finding the gradient of a function by finite-difference still requires a large number of simulations. In this section, we will demonstrate a technique called the Adjoint Method, which allows one to find the gradient of an electromagnetic merit function using just two simulations, regardless of the number of design elements. This derivation follows largely from [87].

The adjoint method is also a gradient descent method, so we will be interested in calculating the gradient of f: $\frac{\partial f}{\partial \mathbf{p}}$. However, in this case we will make use of the chain rule, because we assumed that our merit function is explicitly a function of $\mathbf{x}$. We will assume that the gradient with respect to $\mathbf{x}$, $\frac{\partial f}{\partial \mathbf{x}}$, is known [6]. Then, by the chain rule, the partial derivative of f with respect to element $p_i \in [p_1, ...p_n]$ is

[6] In practice, this is a good assumption. For example, if one is interested in optimizing the electric field intensity, $f(E) = |\mathbf{E}|^2 = \mathbf{E} \cdot \overline{\mathbf{E}}$, then $\frac{\partial f}{\partial \mathbf{E}} = \overline{\mathbf{E}}$, the complex conjugate of the electric field. This can easily be computed from the forward solution which provided us with $\mathbf{E}$.

given by,

$$\frac{\partial f}{\partial p_i} = \sum_j \left(\frac{\partial f}{\partial x_j}\frac{\partial x_j}{\partial p_i} + \frac{\partial f}{\partial \overline{x}_j}\frac{\partial \overline{x}_j}{\partial p_i} \right) = \frac{\partial f}{\partial \mathbf{x}} \cdot \frac{\partial \mathbf{x}}{\partial p_i} + \frac{\partial f}{\partial \overline{\mathbf{x}}} \cdot \frac{\partial \overline{\mathbf{x}}}{\partial p_i} \tag{6.20}$$

where in the second equality we simplified the sum of chain rule derivatives by writing them as a dot product between corresponding vectors. Notice that because $\mathbf{x}$ is complex-valued in general, we also needed to take partial derivatives with respect to its complex conjugate, $\overline{\mathbf{x}}$. Because f is a real-valued function, Eq. 6.20 can be rewritten more conveniently in terms of the unconjugated derivatives involving $\mathbf{x}$:

$$\frac{\partial f}{\partial p_i} = 2\text{Re}\left(\frac{\partial f}{\partial \mathbf{x}} \cdot \frac{\partial \mathbf{x}}{\partial p_i} \right) = 2\text{Re}\left(\left[\frac{\partial f}{\partial \mathbf{x}}\right]^T \frac{\partial \mathbf{x}}{\partial p_i} \right) \tag{6.21}$$

where "Re" denotes the real part. In the second equality, we rewrote the dot product as a matrix product by taking the matrix transpose of $\frac{\partial f}{\partial \mathbf{x}}$ which transforms it into a column vector.

The right-hand side of Eq. 6.21 consists of a known quantity ($\frac{\partial f}{\partial \mathbf{x}}$, which can be calculated analytically from knowledge of the merit function), and an unknown quantity $\frac{\partial \mathbf{x}}{\partial p_i}$ (which describes changes in the electric and magnetic fields with changes in design element p_i). One may think that we haven't made any progress, but we still have an additional equation to work with, namely Maxwell's equations:

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{6.22}$$

We may take the partial derivative of Eq. 6.22 with respect to design element p_i on both sides:

$$\frac{\partial}{\partial p_i}(\mathbf{A}\mathbf{x}) = \frac{\partial \mathbf{b}}{\partial p_i} \tag{6.23}$$

$$\mathbf{A}\frac{\partial \mathbf{x}}{\partial p_i} + \frac{\partial \mathbf{A}}{\partial p_i}\mathbf{x} = \frac{\partial \mathbf{b}}{\partial p_i} \tag{6.24}$$

where in Eq. 6.24 we expanded the partial derivative using the product rule. At this point we note that the derivative $\frac{\partial \mathbf{b}}{\partial p_i} = 0$ in most situations. This is because we are generally not interested in optimizing simulation regions where there are sources present. For instance, in many silicon photonic devices we would like to maintain the same waveguide mode at the input(s) even as the device structure changes. We will maintain this assumption for brevity; it will not affect our conclusions if we were to include an explicit dependence on $\frac{\partial \mathbf{b}}{\partial p_i}$. Thus, Eq. 6.24 may be written:

$$\mathbf{A}\frac{\partial \mathbf{x}}{\partial p_i} = -\frac{\partial \mathbf{A}}{\partial p_i}\mathbf{x} \tag{6.25}$$

we will keep this equation in mind.

The crux of the adjoint method lies in the next step. Let $\mathbf{z}$ be a vector with the same dimension as $\mathbf{x}$. Suppose $\mathbf{z}$ satisfies the following equation:

$$\mathbf{A}^T\mathbf{z} = \frac{\partial f}{\partial \mathbf{x}} \tag{6.26}$$

where $\mathbf{A}^T$ is the matrix transpose of the Maxwell operator[7] and $\frac{\partial f}{\partial \mathbf{x}}$ is the gradient of f with respect to $\mathbf{x}$, a quantity we have already assumed is known. Thus we see that Eq. 6.26 is very similar to Maxwell's equations except we have a new "adjoint operator" $\mathbf{A}^T$ and the original source term $\mathbf{b}$ has been replaced by $\frac{\partial f}{\partial \mathbf{x}}$. We will denote solving Eq. 6.26 as the "Adjoint Problem", analogous to the "Forward Problem" from Eq. 6.13:

$$\text{Adjoint Problem: Solve } \mathbf{A}^T\mathbf{z} = \frac{\partial f}{\partial \mathbf{x}} \text{ for } \mathbf{z} \text{ given } \mathbf{A}^T \text{ and } \frac{\partial f}{\partial \mathbf{x}} \tag{6.27}$$

Intuitively, the adjoint problem can be thought of as injecting the desired solution (represented by $\frac{\partial f}{\partial \mathbf{x}}$) into the simulation domain. By reciprocity, the Maxwell operator is symmetric, and therefore the adjoint problem can be solved using the same computational method as the forward problem [8].

In summary, we have so far applied the chain rule gradient of the merit function with respect to design parameter p_i, obtaining:

$$\frac{\partial f}{\partial p_i} = 2\text{Re}\left(\left[\frac{\partial f}{\partial \mathbf{x}}\right]^T \frac{\partial \mathbf{x}}{\partial p_i}\right) \tag{6.28}$$

Then, we took the gradient of Maxwell's equations $\mathbf{A}\mathbf{x} = \mathbf{b}$ with respect to design element p_i, finding:

$$\mathbf{A}\frac{\partial \mathbf{x}}{\partial p_i} = -\frac{\partial \mathbf{A}}{\partial p_i}\mathbf{x} \tag{6.29}$$

Finally, we proposed the adjoint problem in Eq. 6.27, which can be equivalently written:

$$\mathbf{z}^T\mathbf{A} = \left[\frac{\partial f}{\partial \mathbf{x}}\right]^T \tag{6.30}$$

where we simply took the transpose of both sides. We may now plug Eq. 6.30 into Eq. 6.28:

$$\frac{\partial f}{\partial p_i} = 2\text{Re}\left(\mathbf{z}^T\mathbf{A}\frac{\partial \mathbf{x}}{\partial p_i}\right) \tag{6.31}$$

Then we can insert Eq. 6.29 into Eq. 6.31, obtaining our final equation:

$$\frac{\partial f}{\partial p_i} = -2\text{Re}\left(\mathbf{z}^T\frac{\partial \mathbf{A}}{\partial p_i}\mathbf{x}\right) \tag{6.32}$$

[7] In the literature, it is common to state that for complex-valued $\mathbf{A}$, we should take the complex-conjugate transpose: $\mathbf{A}^\dagger$, also called the adjoint (hence the name of the method). However, as we will find, the regular transpose satisfies the mathematics of the method, confirmed by simulation. Perhaps this disagreement lies in the definition of inner products in quantum mechanics, where complex conjugates are taken implicitly. By contrast, in this derivation we are using the regular inner products defined in linear algebra.

[8] This is not always true if we are using magnetic current sources in simulation. Furthermore, when solving Maxwell's equations, one usually provides a discretized grid. In practice this can break the symmetry between $\mathbf{A}$ and $\mathbf{A}^T$, but in general the geometry of the problem can be represented nearly equivalently, so this is a minor point.

Amazingly, we find that only one quantity on the right hand side of Eq. 6.32 depends on changes in p_i, namely $\frac{\partial \mathbf{A}}{\partial p_i}$.

What is the meaning of $\frac{\partial \mathbf{A}}{\partial p_i}$? Recall that $\mathbf{A}$ defines the materials and physics of Maxwell's equations:

$$\mathbf{A} = \begin{bmatrix} j\omega\varepsilon & \nabla\times \\ \nabla\times & -j\omega\mu \end{bmatrix} \tag{6.33}$$

Therefore, the derivative of $\mathbf{A}$ can be regarded as a "shape derivative" which determines how the permittivity and permeability are updated with respect to our design parameters:

$$\frac{\partial \mathbf{A}}{\partial p_i} = \begin{bmatrix} j\omega\frac{\partial \varepsilon}{\partial p_i} & 0 \\ 0 & -j\omega\frac{\partial \mu}{\partial p_i} \end{bmatrix} \tag{6.34}$$

where the curl terms cancel out as they do not depend on the geometric parameters. In general, the changes in the (spatially-distributed) materials throughout the design region as a function of the design parameters $\mathbf{p}$ is well-known (since it is user-defined). Therefore, Eq. 6.34 is typically easy to calculate for all design elements[9] $p_i \in [p_1, ..., p_n]$.

These observations have an incredible consequence: Eq. 6.32 implies that the partial derivatives of f with respect to every design parameter $p_1, ..., p_n$ can be obtained with just two Maxwell simulations regardless of the number of design parameters (n)! In particular, we require one solution to the forward problem ($\mathbf{x}$ from Eq. 6.13) and one solution to the adjoint problem ($\mathbf{z}$ from Eq. 6.27). Meanwhile, the shape derivatives can be obtained very easily in comparison using Eq. 6.34. Therefore, we have drastically reduced our simulation requirements to perform gradient descent:

$$\begin{aligned} \text{Adjoint method gradient descent: } S &= c \times 2 \\ S &\equiv \text{Number of simulations} \\ c &\equiv \text{Number of gradient descent iterations} \end{aligned} \tag{6.35}$$

In other words, the number of simulations now has no dependence on the number of geometric parameters. In practice, $c \sim 100$, enabling full electromagnetic optimizations with 200 total simulations or less. Consequently, it is now feasible to design incredibly complex electromagnetic structures using computational optimization. In the next section we will show how the adjoint method can be applied to the design of fabrication-compatible vertical grating couplers, devices that are difficult if not impossible to design by conventional means.

[9] In large simulations with hundreds to thousands of parameters, this can still be quite a challenging calculation. Nevertheless, it is far better than alternative means, such as finite-difference.

6.4 Adjoint optimization of efficient CMOS-compatible Si-SiN vertical grating couplers for DWDM applications

Abstract

Data communication in silicon photonic interconnects requires efficient and broadband on/off-chip coupling components. Recently, perfectly vertically-emitting grating couplers have been proposed to increase spatial I/O density of the optical link and potentially improve manufacturing costs and ease of optical beam characterization. In this work, adjoint optimization was leveraged in the design of low-loss single (silicon) and dual layer (silicon + silicon nitride) perfectly-vertical grating couplers that are compatible with a scalable silicon-on-insulator (SOI) platform for the 65nm CMOS technology node. In simulation, the best design peaks at -0.52dB insertion loss with a 1dB-bandwidth of 24nm at the 1310nm datacom wavelength.

Section I: Introduction

The development of high-efficiency silicon photonic links is being pursued to reduce power consumption in data communication on a server- and chip-level scale [11, 92]. A critical challenge to make silicon photonics commercially viable is cost-efficient packaging and on/off-chip coupling [39, 100, 123]. To address part of this problem, Hewlett Packard Labs is developing a modular and wear-tolerant re-pluggable optical connector that leverages the use of on-chip grating couplers for high alignment tolerance [82], which is similar to other grating coupler aided strategies [79, 96, 85, 117]. A schematic of an example connector design is provided in Fig. 6.2(a).

The performance of this type of optical connector relies upon the coupling efficiency of grating couplers, which are micro-scale passive devices that can efficiently convert a waveguide mode to a largely unidirectional output beam of a desired mode shape via the constructive interference of periodically etched scatterers. Due to fundamental physical limitations, it is very difficult to scatter light perfectly vertically (i.e., perpendicular to the chip) using conventional single-etch grating designs because of large back-reflection to the input [90]. Consequently, light is typically scattered at a slight angle to vertical. From a cost-benefit analysis perspective, compensating for this off-vertical scattering is undesirable because it requires more sag in the transceiver micro-lens, which increases the manufacturing challenge and cost [82]. Furthermore, if sag can be reduced, more lenses and gratings can be packed together to increase the spatial I/O density of the outgoing fiber array. A similar conclusion was reached in Ref. [138]. Lastly, characterization and assembly of the optical connector link can potentially be eased by using vertical grating couplers because perpendicular output beams are easier to align and less spherical aberration will be imparted by the micro-lens.

Recently, many flavors of grating coupler designs, both vertical and off-axis, have been proposed to provide low-loss off-chip coupling by, for example, exploiting multiple etch depths [138, 13, 91], multiple patterned layers of various materials [90, 135, 77, 78, 129, 125, 10], sub-wavelength features for anti-reflection [13, 137, 18], back-side metal mirrors or Bragg reflectors [79, 129, 150], or more exotic schemes such as angled-etch "blazed" designs [125]. However, many of these designs are impractical from the perspective of scalable silicon-on-insulator (SOI) fabrication techniques, and previously proposed designs

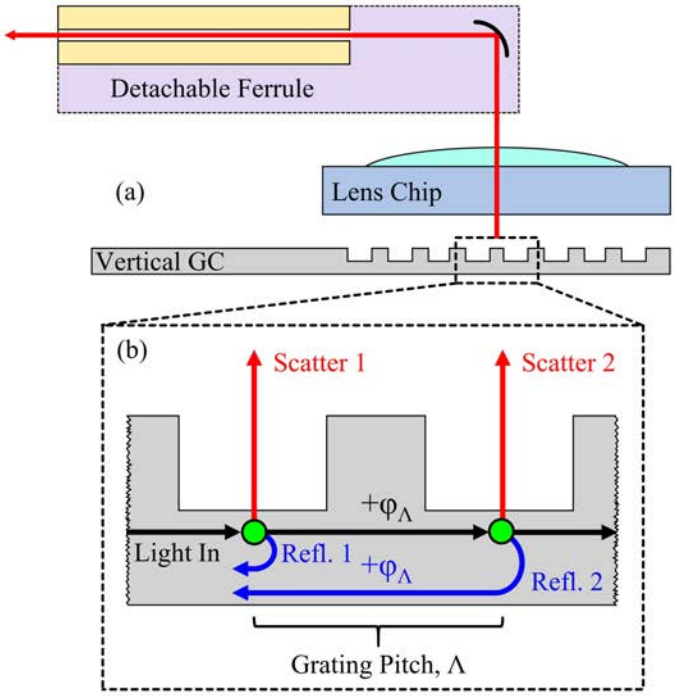

Figure 6.2: (a) Example optical connector schematic consisting of a vertical grating coupler, a microlens chip for focusing, and a detachable optical ferrule. Note that these elements are not to scale. (b) A two-trench slice of a partially-etched grating coupler illustrating the vertical scattering and back-reflection constructive interference phase conditions, which turn out to be equivalent for uniformly etched gratings. ϕ_Λ is the phase collected by a wave that has propagated a single grating period Λ, which is dependent on both the etch duty cycle and waveguide effective index (see *Appendix: Expanded Analysis Using the Grating Equation*).

that provide perfectly-vertical coupling may suffer from lower yield due to fabrication sensitivity, such as mask misalignment in the patterning of multiple layers or etches.

Overall, new design methodologies for perfectly-vertical grating couplers that are low-loss, broadband, have tolerable reflection, are compatible with the transceiver specifications in Datacom or High-Performance Computing (HPC) applications, and either consist of only a single etch or fabrication robust design are still of interest. In this article, an advanced optimization technique known as the adjoint method is employed to design grating couplers that meet these requirements. Ultimately, this work presents the design of the first CMOS-compatible dual-layer silicon nitride-on-silicon (Si-SiN) grating coupler that achieves an industry-competitive simulated -0.52dB insertion loss and 24 nm 1dB-bandwidth at 0° incidence.

Section II: Grating coupler optimization

A: The vertical coupling problem

Fundamentally, uniform single etch grating couplers that scatter light vertically will also have strong back-reflections [90, 91]. This is illustrated in Fig. 6.2(b) where a two-trench slice of a partially-etched grating coupler is depicted. After input light scatters from the first trench, light that is not scattered will continue to propagate, collect a spatial phase ϕ_Λ, and then partially scatter at the next trench. If the vertical scattering Bragg condition is satisfied,

$$\phi_\Lambda = 2\pi n \tag{6.36}$$

for integer n, then scattered light at each trench will interfere constructively in the vertical direction. However, if back-reflections are also taken into account at each of the scattering centers, then a similar constructive interference condition is found that occurs when $2\phi_\Lambda = 2\pi m$ for integer m. Trivially, if the vertical scattering condition is satisfied, then the back-reflection condition is also satisfied with $m = 2n$. Thus, the only way to avoid this problem is to increase the complexity of the design by, for example, adding more degrees of freedom in the form of multiple etch depths, patterned layers, or sub-wavelength features; or modifying the pitch and duty cycle of each trench in a non-trivial manner to reduce back-reflections using advanced computational methods [90, 13, 138, 91, 125, 80]. In this work, the latter method will initially be employed to design low-loss single-etch gratings in Section II-C, then an additional patterned SiN layer will be leveraged to gain higher efficiency and bandwidth in Section II-D.

B: Inverse design via the adjoint method

Design problems in the engineering of electromagnetic devices frequently require the optimization of the shapes and spatial distribution of material components to satisfy some well-defined electromagnetic figure of merit. In the case of the grating coupler in this work, the size and spacing of the grating trenches should be optimized to maximize the scattered power that is mode-matched to a vertically-oriented optical fiber. More explicitly this optimization can be written in the form,

$$F^* = \max_x \eta(E, H) \tag{6.37}$$

where x is a vector of optimization parameters that describe the grating coupler geometry and $\eta(E, H)$ is the coupling efficiency to a Gaussian optical fiber mode (beam diameter = 9.2μm at $\lambda = 1310\,\text{nm}$; see *Appendix: Adjoint Optimization Details and Workflow*), which is defined explicitly in terms of the electric and magnetic fields E and H but is only implicitly related to x via Maxwell's equations [10].

Furthermore, there are often practical constraints that are imposed on the fabrication of electromagnetic devices, e.g., feature sizes or radius of curvature. This can be included in Eq. 6.37 in the form of a penalty function, i.e.,

$$F_p^* = \max_x \left[\eta(E, H) - p(x)\right] \tag{6.38}$$

[10] Note that we are using different notation to describe the parameters and field quantities in this section, compared to the first few sections of Chapter 6. Namely, we have replaced $\mathbf{p}$ with x, and we will use p to describe a penalty function on the design parameters.

where p penalizes undesirable geometric features and is explicitly a function of the design parameters x.

The most common approach to solving Eq. 6.38 is to physically intuit relevant optimization parameters x, and sweep them in simulations. The problem with this approach, especially applied to grating couplers, is that the number of simulations required for parameter sweeps increases exponentially with the number of parameters [71].

To alleviate this problem a gradient-based optimization algorithm that leverages the adjoint method was used in this work. The adjoint method allows for efficient calculations of the gradient of an electromagnetic figure of merit (such as Eq. 6.38) with minimal simulation overhead. The method has its limitations since electromagnetic design problems are generally non-convex and hard to solve by gradient-descent, but it allows an engineer to optimize over large parameter spaces very efficiently when combined with physical intuition and a hierarchical design approach [89]. For brevity the adjoint method will not be rigourously described, but more information can be found in [87, 76, 97, 32, 44] where the adjoint method and other similar topology optimization methods are discussed in detail. Furthermore, *Appendix: Adjoint Optimization Details and Workflow* gives additional details and specifications about the optimization workflow [11]. The optimization and FDFD simulation package used in this work is EMopt [88]; insertion loss and reflection were calculated using Lumerical FDTD.

C: Multi-wavelength optimization of single-etch Si grating

This article is intended to provide a feasible design for a fabrication-friendly vertical grating coupler. Hence, all design choices were motivated by realistic foundry specifications. In particular, the following layer dimensions were chosen in this work: the height of the waveguide is 304 nm, the depth of the grating trenches are 159 nm, and the thickness of the buried oxide (BOX) layer is 2 μm – all of which are representative of a $1310\,\mathrm{nm}$-wavelength $300\,\mathrm{mm}$ wafer SOI platform available at pilot foundries, such as CEA-LETI [77, 78, 141]. Furthermore, for compatibility with immersion deep-UV photolithography patterning specifications [90, 141], a realistic critical feature size of 65nm was imposed on the silicon grating.

The simplest implementation of the adjoint method for the design of grating couplers is to apply Eq. 6.38 to a single-etch silicon grating coupler at the target wavelength of $1310\,\mathrm{nm}$. For off-axis coupling this figure of merit choice leads to state-of-the-art designs. For perfectly-vertical coupling, however, this optimization tends to produce good peak performance results (-0.55dB insertion loss) but narrow bandwidth (<3 nm 1dB-bandwidth) which is incompatible with dense wavelength division multiplexing (DWDM) applications where a $\sim 20\,\mathrm{nm}$ 1dB-bandwidth is desirable [75].

To incorporate the importance of bandwidth into the design, a multi-wavelength merit function was implemented by modifying Eq. 6.38 to the following:

$$F_p^* = \max_x \left[\sum_\lambda c_\lambda \eta_\lambda(E_\lambda, H_\lambda) - p(x) \right] \tag{6.39}$$

where the optimization argument is now a weighted average of the coupling efficiency across multiple discrete wavelengths with user-defined weights, c_λ, and coupling efficiency, $\eta_\lambda(E_\lambda, H_\lambda)$, that are wave-

[11] A derivation of the adjoint method is provided in Sections 6.1-6.3 of this thesis

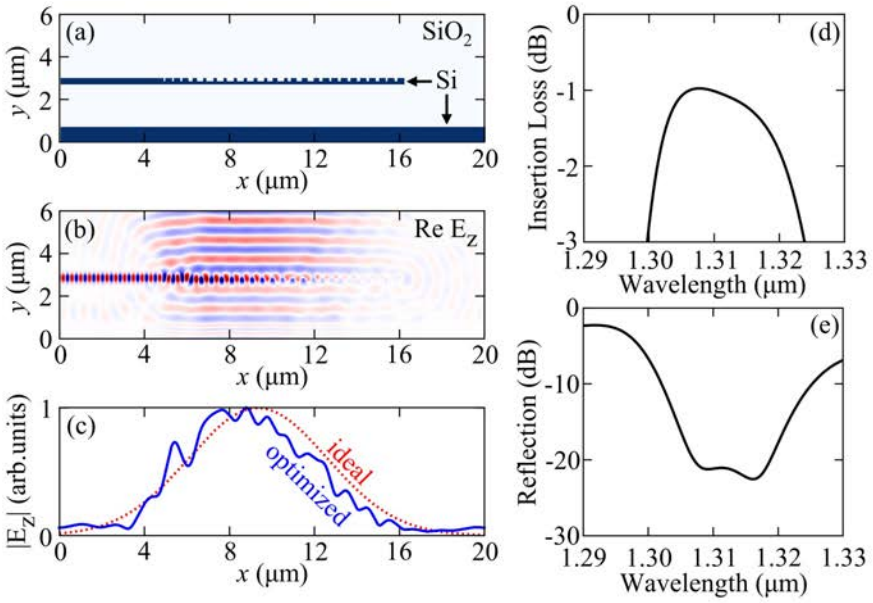

Figure 6.3: Multi-wavelength optimization result for a single-etch vertical grating coupler with Si layer thickness of 304nm, etch depth of 159nm, and minimum feature size of 65nm. (a)-(c) Structure, E_Z field profile, and $|E_Z|$ mode-match field slice at $\lambda = 1310$ nm. (d) Insertion loss and (e) reflection spectra.

length dependent. Note that in Refs. [125, 115, 4], conceptually similar broadband inverse design figures of merit have been proposed.

The merit function in Eq. 6.39 was applied to a single-etch silicon grating coupler (see *Appendix: Adjoint Optimization Details and Workflow* for implementation details, including the parameterization of the grating coupler geometry). The result of the optimization is shown in Fig. 6.3 where Fig. 6.3(a)-(c) gives the grating structure, frequency-domain electric field, and mode-match profile at a wavelength $\lambda = 1310$ nm while Fig. 6.3(d)-(e) give the insertion loss and reflection spectra. The grating reaches a peak insertion loss of -0.98dB (-1.03 dB at $\lambda = 1310$ nm) with a 1dB-bandwidth of 19nm – a 6x improvement over the single-wavelength optimized device discussed previously. The back-reflection of the grating coupler limits the 1dB-bandwidth and is severe at the edge of the 1dB-bandwidth (exceeding -10dB), but tolerable around peak.

D: Dual layer silicon nitride-on-silicon design

While the single layer Si grating coupler has many desirable characteristics, such as its fabrication-compatible critical feature size and its requirement for only a single etch, every improvement in insertion loss and bandwidth is important for the photonic link budget. Thus, a patterned dual layer silicon nitride-on-silicon (Si-SiN) design that is available in pilot foundries was pursued. Other Si-SiN grating couplers have been suggested in Refs. [78, 114], but in this work light is injected through the Si waveguide

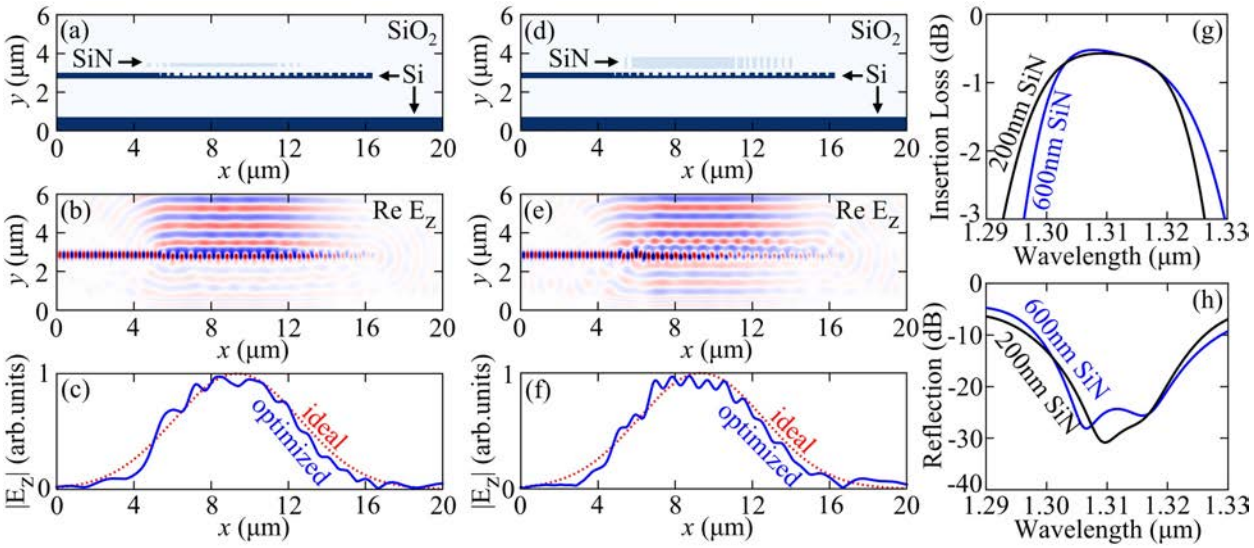

Figure 6.4: Multi-wavelength optimizations of dual layer Si-SiN vertical grating couplers consisting of two varying SiN layers on a Si layer with thickness of 304nm and etch depth of 159nm. (a)-(c) and (d)-(f) give the structure, E_Z field profile, and $|E_Z|$ mode-match field slice at $\lambda = 1310$ nm of the respective Si-SiN gratings. (a)-(c) 200nm SiN thickness and 300nm interlayer oxide spacing. (d)-(f) 600nm SiN thickness and 200nm oxide interlayer spacing. Each grating coupler uses Si and SiN critical feature sizes of 65nm and 100nm respectively. (g) Insertion loss spectra and (h) reflection spectra of the two designs.

and emitted perfectly-vertically from the device. Furthermore, the SiN acts primarily as an anti-reflection layer and does not significantly interact with the near-field of the Si grating.

Anti-reflection layers can improve the upward directionality of grating couplers by reducing the impedance mismatch between farfield radiated light and guided waveguide modes – similar but inverse to tuning the thickness of the buried oxide to prevent radiation to the substrate [91]. For what can be considered an additional functionality, anti-reflection layers can help cancel the back-reflection of a vertical grating coupler. Even further improvements to back-reflection and mode overlap can be made by patterning the anti-reflection layer to give more degrees of freedom to the grating coupler optimization and to facilitate angle correction of the output beam from the Si grating, acting almost like a conventional diffraction grating. Once again, the adjoint method was employed to design such a patterned layer.

Two Si-SiN layer geometries were pursued in this work. Inspired by the layer stack presented in Ref. [141], one of the dual layer designs consists of a 304nm silicon grating coupler layer and a 600nm SiN anti-reflection layer separated by 200nm of oxide. The alternate geometry consists of a 304nm silicon grating coupler layer and a 200nm SiN layer separated by 300nm of oxide. The directionality improvement offered by the anti-reflection layer is periodic in the SiN thickness with some dependence on the thickness of the interlayer oxide. In this case the SiN and interlayer oxide thicknesses were chosen to be compatible with fabrication restrictions and meanwhile maximize the grating coupler directionality.

The multi-wavelength optimization function given in Eq. 6.39 was applied to the dual layer design

with a similar implementation to that of the single-etch Si design (see *Appendix: Adjoint Optimization Details and Workflow*). To be compatible with fabrication constraints, the Si grating and SiN layer minimum feature size were constrained to 65nm and 100nm respectively (assuming a lower resolution mask would be chosen for the SiN layer, and a high aspect ratio etch can be achieved for the 600nm design [15]). The two results are shown in Fig. 6.4 with a similar format to that of the single layer Si design from Fig. 6.3. The optimized Si-200nm SiN and Si-600nm SiN designs reached peak insertion losses of $-0.57\,\mathrm{dB}$ and $-0.52\,\mathrm{dB}$, respectively, with 1dB-bandwidths of $25\,\mathrm{nm}$ and $24\,\mathrm{nm}$. Thus, compared to the single-etch Si design, the dual-layer Si-SiN designs offer approximately +0.5dB insertion loss and +5nm 1dB-bandwidth enhancement.

Section III: Physical analysis of designs

A: Silicon grating

The adjoint method produced a nontrivial Si grating coupler design in order to scatter light vertically while maximizing mode overlap and minimizing back-reflection. Some clues for how the optimization achieved this functionality are evident in the the plot of pitch and duty cycle at each grating period before and after minimum feature constraints were applied in the optimization, shown in Fig. 6.5. As can be seen in Fig. 6.5(a), the grating pitch switches abruptly from 0.38 μm to 0.47 μm in the first few periods. Using the expanded version of the grating equation from Eq. 6.36 indicates that a change in grating pitch corresponds to a change in scattering angle (see derivation in *Appendix: Expanded Analysis Using the Grating Equation*). This implies that different sections of the grating coupler scatter at two angles slightly off-vertical resulting in the cancellation of lateral field components and net vertical scattering. This can be qualitatively verified in the field profiles from Fig. 6.3(b), Fig. 6.4(b), and Fig. 6.4(e). Additionally, two distinct minima are visible in the reflection spectra of Fig. 6.3(e) and Fig. 6.4(h), which implies that the Bragg condition is met at two slightly detuned wavelengths, or equivalently scattering angles. Since the grating coupler scatters off-vertical in two sections, the back-reflection condition is alleviated.

B: Silicon nitride layer

As mentioned previously, the SiN layer acts primarily as an anti-reflection layer, thereby enhancing the directionality and bandwidth of the grating coupler while also cancelling back-reflections to the input. Consequently by allowing design freedom in the patterning of the SiN layer, the optimizer is capable of balancing the competing effects of mode overlap, back-reflections, and directionality. In particular, the patterned SiN layer helps improve the coupling efficiency of the grating coupler in two ways: (1) it provides additional degrees of freedom to the Si grating optimization to maintain low back-reflection while improving mode overlap, and (2) it provides angle correction for off-vertical scattering occurring at the ends of the Si grating. The latter effect is evident in the field profile of Fig. 6.4(b), especially on the right side of the grating. The action of the SiN layer can be thought of as similar to a conventional diffraction grating with a chirped pitch. Alternatively, it can be interpreted as an optical phase mask or binarized lens which slows vertical light emanating from the center of the grating while focusing light that is diffracting off-vertically at the edges. The middle portion of the SiN remained unpatterned after optimization in

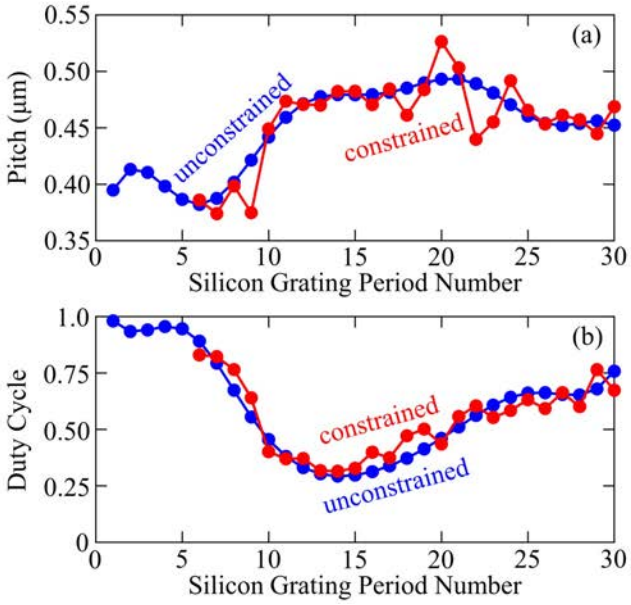

Figure 6.5: (a) Grating pitch and (b) duty cycle plotted versus the Si grating period number from the Si–600nm SiN dual layer design (Fig. 6.4(d)). The corresponding plots for the single-layer Si grating and Si–200nm SiN grating are qualitatively similar. In both plots, the pitch and duty cycle are plotted before (blue) and after (red) the set of constrained optimizations were performed. The full set of data is available in *Appendix: Grating Coupler Data*.

order to maintain directionality enhancement. At the expense of needing higher aspect ratio trenches in the SiN layer, the Si-600nm SiN grating coupler was able to achieve slightly better insertion loss than the 200nm SiN design because of the additional phase control offered by thicker SiN patterns, which in turn helped provide better mode overlap.

Section IV: Fabrication sensitivity

To more fully characterize the performance of the dual layer (Si-SiN) grating couplers in this work, the sensitivity of the Si-600nm SiN device to typical errors in fabrication and processing was simulated.

Perhaps one of the most promising characteristics of the dual layer design is depicted in Fig. 6.6(a), where the extreme tolerance of the design to the misalignment of the layers (i.e. mask misalignment) is shown. The device maintains an insertion loss >-1.0dB at $\lambda = 1310$nm for misalignment as large as $\pm 1\,\mu$m, which is well within the expected $3\sigma = 15\,$nm optical alignment tolerance for immersion lithography. This high tolerance can be attributed to the unique physics of the dual layer design, where the patterned SiN acts primarily as an anti-reflection and focusing layer and hence can handle a misalignment on the scale of the freely-propagating output wave of the grating.

Fig. 6.6(b) depicts the sensitivity of the design to changes in the SiN and interlayer oxide thicknesses.

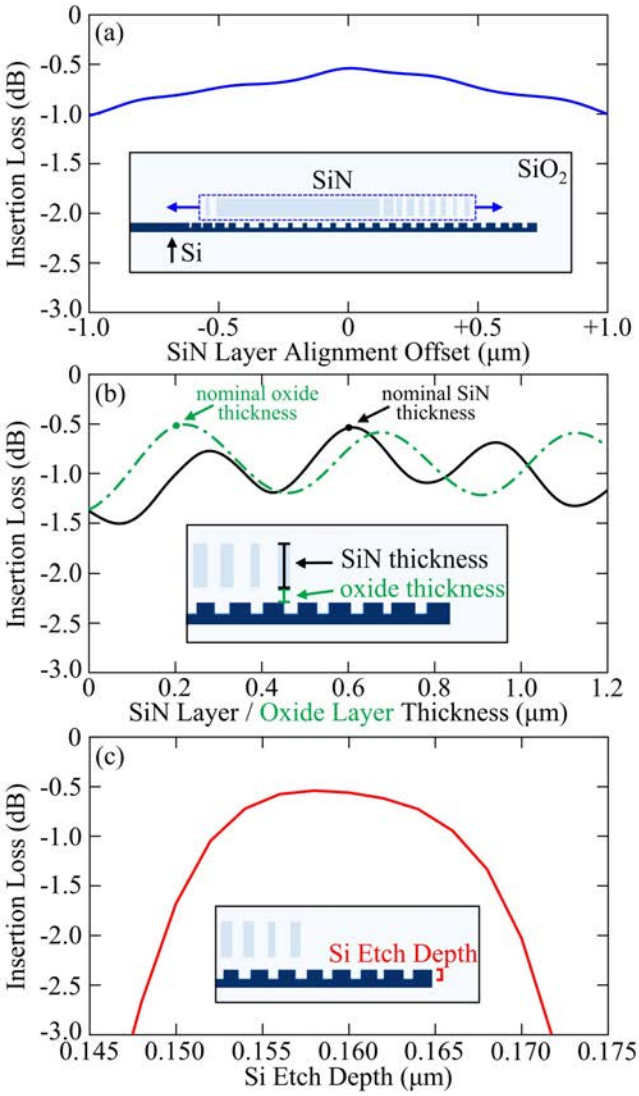

Figure 6.6: Fabrication sensitivity plots for the dual layer (Si-SiN) design with 600nm SiN thickness and 200nm interlayer oxide thickness as a function of (a) SiN layer misalignment, (b) SiN layer / interlayer oxide thickness, and (c) Si etch depth. Insets depict the property of the design that is being changed.

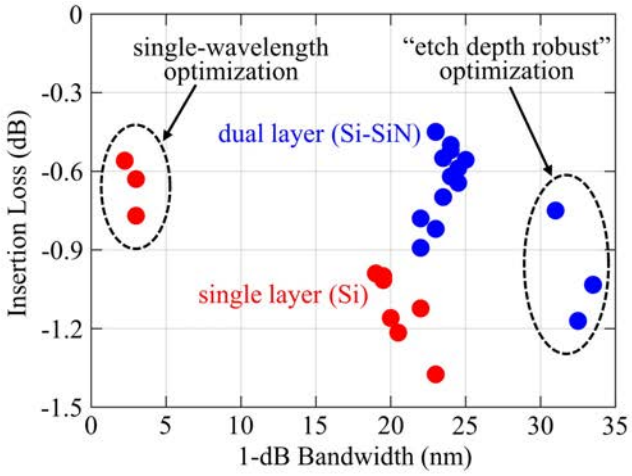

Figure 6.7: Insertion loss vs. 1dB-bandwidth for the various designs that were optimized in this work. The results from additional single-wavelength optimizations and etch depth robust optimizations are highlighted. The spread in the data can be attributed to optimizations performed at varied etch depth, constrained minimum feature size, and SiN thickness.

The anti-reflection property is very apparent in this plot, clearly illustrating the expected $\lambda/2n$ periodicity between maxima in the modulations of the insertion loss. Note that in the solid black curve the SiN thickness was varied from its nominal value of 600nm while the interlayer oxide spacing was fixed to its nominal value of 200nm. Similarly, the SiN thickness was fixed and the interlayer oxide thickness varied from nominal in the dashed green curve.

Finally, in Fig. 6.6(c) the sensitivity of the design to variations in the etch depth of the Si grating is shown. The design is much less robust to Si etch depth than other typical fabrication variations, tolerating at most ± 10 nm. This is not entirely surprising, because even small changes to the thickness of the etched silicon can dramatically alter the effective index, fundamentally shifting the phase accumulated by traveling waves along each period of the grating. Perfectly-vertical gratings are particularly susceptible to these changes, due to intense back-reflections that can result. Though not included here, temperature variations will have a similar effect of shifting the effective index of the grating. The effect appears to be tolerable within the typical operating temperature range of the grating coupler based on an analysis using the thermo-optic coefficient of Si [65].

Optimizations of the same methodology reported in this paper but for different Si etch depths were performed, but were not included for brevity. Generally it was found that the 159nm etch depth performed the best for the particular SOI material stack in this work, but similar properties and insertion loss could be achieved for other etch depths. "Etch depth robust" optimizations were also performed where etch depth was explicitly included in the optimization figure of merit to extend the width of the sensitivity plot in Fig. 6.6(c) at an expense to peak insertion loss. Data from these optimization results, among others, are reported in Fig. 6.7.

Section V: Discussion and conclusion

In this work it was demonstrated that adjoint optimization is a powerful tool for designing high-performance vertical grating couplers. This is shown with clarity in Fig. 6.7, where results from all of the optimizations performed using the methodologies presented in this article are aggregated. The bandwidth benefit resulting from the use of a multi-wavelength optimization protocol over the single-wavelength optimization is shown with clarity, where minimal concession was made to peak insertion loss even for the single-etch Si grating. In principle, using the optimization methods presented in this work, one can obtain the desired coupling efficiency vs. bandwidth trade-off needed for a particular photonic link. Moreover, the inclusion of a patterned SiN layer that provides anti-reflection and light focusing gives a clear advantage over the single-etch design. Due to the unique physics of the dual layer grating coupler, the design is very tolerant to fabrication errors and minimum feature constraints.

As an additional reference, the best results from this paper are compared to various other results from the literature in Table 6.1. In simulation, the single-etch Si grating matches, to the authors' knowledge, the other best result in the literature for vertical coupling [125], but with a more fabrication-compatible minimum feature size. Of the multi-layer and multi-etch perfectly-vertical designs, only two have reported better results than the multi-layer design in this work [90, 125], but these designs require the patterning of two (tightly aligned) silicon layers which is more challenging from a fabrication perspective.

Looking forward, the remaining back-reflection in this work's designs may potentially be reduced with the aid of more complex strategies such as the use of sub-wavelength structures [13, 18]. Adaptation of novel and global optimization methods will likely lead to further improvements [57, 86, 24]. Moreover, the design strategy from this work could be applicable to the the development of polarization-splitting grating couplers [137, 12, 149, 134], since orthogonal polarizations will be inherently decoupled for a perfectly-vertical angle of incidence. Lastly, future modular optical connector designs (Fig. 6.2(a)) may allow one to optimize vertical grating couplers for different output mode shapes, thereby providing more flexibility in optimization and possibly better performing designs. Overall, the vertical grating couplers presented in this work are stride towards developing high-efficiency silicon photonic links.

Table 6.1: Grating coupler literature comparison.

Cite	θ	λ (nm)	# Layers or # Etches	Sim. I.L. (dB)	Meas. I.L. (dB)	Sim. Refl. (dB)	Sim. 1dB-BW (nm)	Critical Dimension (nm)
*	0°	1310	1 (Si)	-1.0	–	-21	19	65
*	0°	1310	2 (Si-SiN)	-0.52	–	-27	24	65 (Si) 100 (SiN)
*	0°	1310	2 (Si-SiN)	-0.57	–	-31	25	65 (Si) 100 (SiN)
[125]	0°	1310	1 (Si)	-1.0	–	–	28	50
[90]	0°	1550	2 (Si-Si)	-0.137	–	-41	24	65
[125]	0°	1550	2 (Si-Si)	-0.25	–	–	22	65
[138]	0°	1550	2 etch (Si)	-0.60	-1.5	<-20	–	30
[91]	8°	1550	2 etch (Si)	-0.70	–	-49	22	100
[85]	8°	1550	1 (Si)	–	-1.25	<-17	>25	>100
[18]	10°	1310	2 etch (Si), SWG	-0.70	-1.9	–	30	100
[13]	–	1550	2 etch (Si), SWG	-1.1	-1.3	<-17	30	100
[80]	14.5°	1550	1 (Si)	-0.8	-0.9	<-17	38	60
[114]	21°	1550	2 (Si-SiN)	-1.0	-1.29	–	80	>200
[78]	29°	1310	2 (Si-SiN)	-1.37	-2.0	-22	73	120

* This work
SWG = Sub-Wavelength Grating

Appendix: Adjoint optimization details and workflow

Merit function definitions

In each of the adjoint method merit functions suggested in this work (Eq. (6.37)-(6.39)), a mode-match coupling efficiency to a Gaussian optical fiber mode, $\eta(E, H)$, was indicated. The explicit definition of the coupling efficiency is given by,

$$\eta(\mathbf{E}, \mathbf{H}) = \frac{1}{4P_m P_{src}} \left| \iint_A \mathbf{dA} \cdot \mathbf{E} \times \mathbf{H}^*_{\mathbf{m}} \right|^2 \tag{6.40}$$

where $\mathbf{A}$ is the area of integration at the location of the desired output mode (with vector direction given by the normal), P_m is the power in the desired mode, P_{src} is the input source power to the simulation, $\mathbf{E}$ is the simulated incident electric field, and $\mathbf{H_m}$ is the incident magnetic field of the desired mode which in this case is a vertically-propagating Gaussian beam. This definition was originally derived in Ref. [87] and equivalent definitions are used elsewhere [138].

Moreover, the multi-wavelength coupling efficiency merit function that was used for all of the optimizations showcased in the main text was given in Eq. 6.39. As a consequence of using an FDFD method to calculate the electric and magnetic fields at multiple wavelengths, each adjoint method calculation of the gradient requires an increase in the number of simulations proportional to the number of discrete wavelengths. Thus, when using many wavelengths the optimization time can become infeasible even on a high-performance server. Consequently, optimizations were limited to 3 discrete wavelengths where $\lambda = \{1305\text{nm}, 1310\text{nm}, 1315\text{nm}\}$ with corresponding weights $c_\lambda = \{c_{1305\text{nm}}, c_{1310\text{nm}}, c_{1315\text{nm}}\}$. The degree of wavelength spacing was not exhaustively experimented with, but it did not seem to greatly influence the end result. Changing the weights allows for some control over the shape of the insertion loss spectrum, and was varied based on whether the optimization was unconstrained or constrained.

The penalty function $p(x)$ in Eq. 6.39 is a smoothed rectangular function that penalizes any features with size $0 \lesssim x \lesssim x_0$ where x_0 is the minimum imposed feature size. The stringency of this penalty function can be tuned by varying the slope and amplitude of this function. A similar function was used in Refs. [90, 91].

Implementation and parameterization

For each of the designs presented in this work (namely the single-etch Si grating from Fig. 6.3 and the two Si-SiN grating couplers from Fig. 6.4), unconstrained and constrained optimizations were performed sequentially in the form of a hierarchical design protocol [89]. In the unconstrained optimization ($p(x) = 0$), the trench width and pitch of the Si grating (and SiN layer) were each parameterized using a Fourier series to allow for a smooth functional evolution of the device geometry,

$$\Lambda(i) = a_0 + \sum_{m=1}^{M} \left[a_m \sin\left(m \frac{\pi}{2} \frac{i}{N} \right) + b_m \cos\left(m \frac{\pi}{2} \frac{i}{N} \right) \right] \tag{6.41}$$

where i is the period number of the corresponding grating trench, N is the total number of trenches and periods, M is the number of Fourier series terms, and the a_m and b_m coefficients represent the optimization variables, x, from Eq. 6.39. This parameterization was originally proposed and used in the dual layer

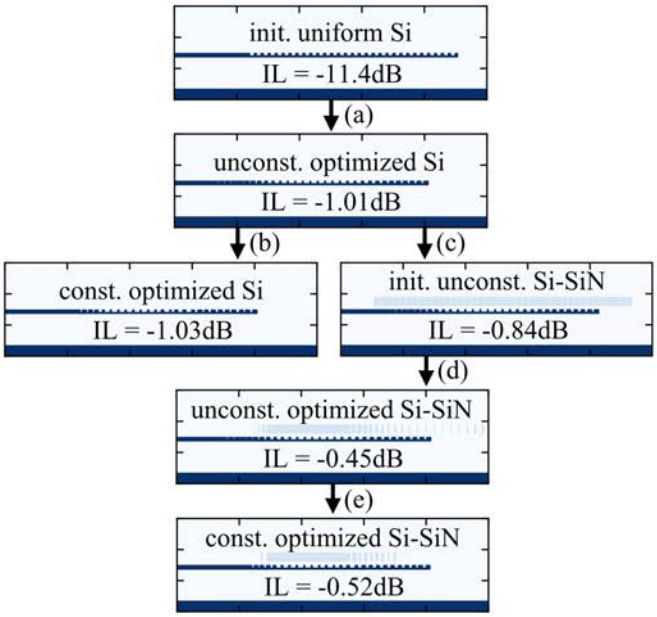

Figure 6.8: Optimization workflow, which shows the initial conditions and final results for each optimization as well as the corresponding insertion losses along each step. (a) Initial unconstrained optimization of a uniformly-etched Si grating coupler. (b) Constrained optimization of the single-etch Si grating. (c) Initial condition for the Si-SiN grating optimization, which uses the previous unconstrained Si grating optimization result and a uniformly etched SiN layer with large duty cycle (0.95). (d) Unconstrained optimization of the Si-SiN design. (e) Constrained optimization of the Si-SiN design.

Si grating optimization from Ref. [90], but in this work $M = 10$ Fourier terms were used as opposed to $M = 5$ in the cited work. Note that $N = 30$ and $N = 40$ in the Si grating and patterned SiN layer respectively. The c_λ wavelength weight coefficients in Eq. 6.39 were $c_\lambda = \{0.4, 0.2, 0.4\}$ during unconstrained optimization.

After the unconstrained optimization completed, the design was reparameterized to allow the pitches and trench widths to evolve independently of any functional form, but with added constraints on feature sizes (i.e. $p(x) \neq 0$ and takes the form of the smoothed rectangular function mentioned above). This causes the sharp deviations from the smooth form of the pitch and duty cycle in Fig. 6.5. As the constrained optimization was performed, the penalty function stringency was increased in 3 sequential optimization steps to completely enforce the minimum feature constraints. The c_λ wavelength weight coefficients in Eq. 6.39 were $c_\lambda = \{0.33, 0.33, 0.33\}$ during constrained optimization.

Workflow

The optimization workflow for the generation of grating coupler designs in this work is given in Fig. 6.8. Each arrow represents either an unconstrained or constrained optimization with corresponding initial

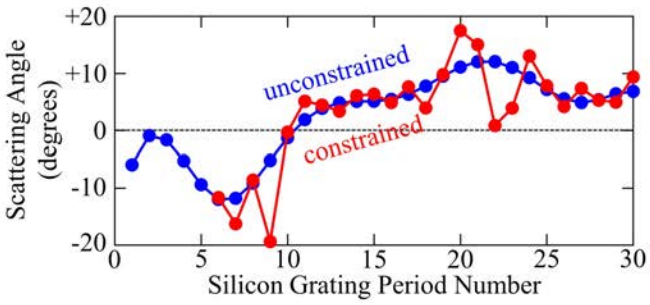

Figure 6.9: The scattering angle (in degrees) at each grating trench found using Eq. 6.43 applied to the pitch and duty cycle from Fig. 6.5.

conditions and final optimization results. The initial single-etch Si grating was uniformly-etched with a pitch that satisfied the vertical scattering condition (Eq. 6.36 with a duty cycle of 0.5). High back-reflection limited the insertion loss of this device. Note that many other initial conditions were tried in the context of this work, but this duty cycle tended to produce the best (and most constraint-friendly) results. After unconstrained optimization, the insertion loss dramatically improved, and this unconstrained design was used as an initial condition for the constrained single-etch Si grating optimization (which gave the result in Fig. 6.3) as well as an unconstrained optimization of the Si-SiN grating couplers. Finally, a constrained optimization was performed on the Si-SiN design which produced the designs in Fig. 6.4.

Appendix: Expanded analysis using the grating equation

In Eq. 6.36 the vertical coupling phase condition was provided. The more general equation, known as the grating equation, is defined here:

$$\phi_\Lambda - k_0 \Lambda \sin\theta = 2\pi m \tag{6.42}$$

where ϕ_Λ is the phase collected across one period Λ of the grating, k_0 the wavenumber of the cladding medium, θ is the angle of the output beam relative to zenith, and m is an integer. To obtain more information about the physics of an apodized grating coupler, it can be useful to solve for the scattering angle as a function of the grating period number. To do so, a duty cycle d_i and pitch Λ_i at each index i is assumed. Furthermore, the effective refractive index of the unetched and etched portions of the grating is taken to be n_1 and n_2, respectively. Using this, the grating equation can be rewritten in terms of scattering angle per grating period as,

$$\sin\theta_i = \frac{n_1}{n_0} d_i + \frac{n_2}{n_0}(1 - d_i) - \frac{m\lambda}{n_0 \Lambda_i} \tag{6.43}$$

where λ is the free-space wavelength and n_0 is the refractive index of the cladding medium.

Using the pitch and duty cycle information from Fig. 6.5, the scattering angle per grating period number of the Si–600nm SiN grating is plotted in Fig. 6.9.

It can be observed in Fig. 6.9 that the grating coupler rarely scatters perfectly-vertically for any given grating period in this approximation. Instead, the grating scatters light at a negative angle towards the beginning and at a positive angle towards the end of the grating. These two beams interfere to produce net vertical scattering.

Appendix: Grating coupler data

For reproducibility, the data for the two Si-SiN grating couplers showcased in this work is presented in Table 6.2; all data is given in units of μm. The trenches and lines occur sequentially in order. Note that the first SiN etched trench (T.) row for each design gives the offset of the first SiN line from the first Si etched trench. The refractive index of Si, SiN, and SiO_2 at $\lambda = 1310$nm was taken to be 3.5003, 1.956, and 1.447 respectively in this work. The mode-match was measured 2μm above the Si layer.

6.5 Conclusion

In this chapter we have introduced inverse design via the adjoint method, ultimately demonstrating that it can provide vastly improved computational electromagnetic design capabilities over conventional methods. Then, we showed how the adjoint method could be applied to the design of fabrication-friendly perfectly vertical grating couplers. This resulted in the design of an 65nm-CMOS compatible vertical grating coupler with an industry-competitive -0.52dB simulated insertion loss.

Table 6.2: Optimized Si-SiN grating coupler data.

Si – 600nm SiN GC				Si – 200nm SiN GC			
Si T.	Si L.	SiN T.	SiN L.	Si T.	Si L.	SiN T.	SiN L.
0.0655	0.3204	0.5310	0.1376	0.0664	0.3003	-0.6805	0.2367
0.0663	0.3075	0.2188	5.3666	0.1478	0.2114	0.2744	0.1016
0.0936	0.3047	0.1248	0.3320	0.3020	0.1444	0.1948	0.2505
0.1350	0.2395	0.1027	0.1877	0.3994	0.0769	0.1523	5.5697
0.2690	0.1800	0.1460	0.2578	0.3427	0.1544	0.1769	0.2401
0.2983	0.1751	0.1475	0.1667	0.3113	0.1783	0.2137	0.1708
0.2960	0.1745	0.1660	0.1979	0.3041	0.1655	0.2297	0.1435
0.3210	0.1490	0.1712	0.1823	0.2925	0.1816		
0.3304	0.1517	0.2377	0.1309	0.2930	0.1879		
0.3243	0.1579	0.2479	0.1667	0.2836	0.1680		
0.2829	0.1875			0.2974	0.1955		
0.3030	0.1812			0.3020	0.1573		
0.2437	0.2173			0.2523	0.2043		
0.2413	0.2424			0.3330	0.2170		
0.2973	0.2289			0.2180	0.2492		
0.2229	0.2802			0.1535	0.3193		
0.1742	0.2655			0.1883	0.2755		
0.2036	0.2514			0.1315	0.3296		
0.2046	0.2870			0.1506	0.2787		
0.1715	0.2939			0.1556	0.3141		
0.1849	0.2684			0.1319	0.3193		
0.1554	0.3056			0.1668	0.2758		
0.1825	0.2746			0.1041	0.3455		
0.1046	0.3399			0.1336	0.3173		
0.1529	0.3157						

T. = Etched Trench [μm]; L. = Unetched Line [μm]

Chapter 7

Novel Inverse Design Topics

In the previous chapter we discussed inverse design via the adjoint method in detail, then applied it to fabrication-friendly grating couplers. In this chapter we will present a brief look at novel, emerging topics in inverse design. In the literature, inverse design tends to refer to applications of the adjoint method, but in this case we are using the term loosely to describe general electromagnetic design techniques. In Section 7.1 we will present a novel semi-analytical transfer-matrix based optimization method for the design of thin-film interference filter devices (such as distributed Bragg reflectors). Section 7.2 demonstrates that this method can be used to obtain mirrors with $> 99\%$ reflectivity over an extremely broad frequency range. In Section 7.3 we will briefly an exciting new class of inverse design optimization techniques that leverage machine learning, namely reinforcement learning (RL). These methods use neural networks as an additional tool to help explore large design spaces in search of global optima.

7.1 Transfer-matrix inverse design method for 1D thin-film interference filters

In this section we will present an inverse design method developed by the current author and Omair to design thin-film devices [105, 46]. This method relies upon the semi-analytical transfer-matrix method to efficiently simulate the Fresnel coefficients of a one-dimensional stack of layers with arbitrary refractive index (including complex refractive index for lossy layers). In the next section we will apply this method to the design of an extremely broadband distributed Bragg reflector.

Introduction to the transfer-matrix method

Before proceeding we will describe the semi-analytical transfer matrix method. This will be integral to our derivation of the inverse design method. Let r and t denote the Fresnel coefficients of a one-dimensional layered device. Note that r and t are defined for a specific frequency, polarization, and incident angle of light implicitly. We will include these dependencies later on. The transfer-matrix method [22] posits that

r and t satisfy the following:

$$r = \frac{M_{21}}{M_{11}} \tag{7.1}$$

$$t = \frac{1}{M_{11}} \tag{7.2}$$

where M_{11} and M_{21} correspond to the matrix elements of a 2×2 transfer-matrix, $\mathbf{M}$:

$$\mathbf{M} = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \tag{7.3}$$

The transfer-matrix, $\mathbf{M}$, can be constructed by multiplying several building-block matrices, which will be described below.

Suppose that the thin-film device we wish to simulate consists of m layers with (complex) refractive indices, $n_1, ..., n_m$, and thicknesses $d_1, ..., d_m$. Let n_i and d_i define the refractive index and thickness of the i-th layer. Furthermore, let θ_0 be the angle of incident light on the device (relative to normal). Then, by Snell's Law we can define the angle of light propagation within each layer of the one-dimensional stack using:

$$\theta_{i+1} \equiv \sin^{-1}\left(\frac{n_i}{n_{i+1}} \sin\theta_i\right) \tag{7.4}$$

where θ_i is the angle of light in the i-th layer, where $i = 0, 1, 2, ..., m$. For convenience, we define the following quantities:

$$\beta_i \equiv \frac{2\pi n_i}{\lambda} \cos\theta_i \tag{7.5}$$

$$L_i \equiv \frac{\beta_{i+1}}{\beta_i} \tag{7.6}$$

where β_i can be thought of as an effective propagation wavevector in the i-th layer (λ is the free-space wavelength).

The transfer-matrix $\mathbf{M}$ can now be defined using these quantities. Note that the form of the $\mathbf{M}$ building-block matrices is different depending on the incident polarization of light. Thus, we will consider both s- (transverse-electric) and p- (transverse-magnetic) polarized incident light. $\mathbf{M}^{\{s,p\}}$ is given by the product of $m + 1$ alternating transmission and propagation matrices:

$$\mathbf{M}^{\{s,p\}} = \prod_{i=0}^{m} \mathbf{T}_{i,i+1}^{\{s,p\}} \mathbf{P}_{i+1} \tag{7.7}$$

where these matrices can be defined per layer of the device as:

$$\mathbf{T}_{i,i+1}^{s} = \frac{1}{2} \begin{bmatrix} 1 + L_i & 1 - L_i \\ 1 - L_i & 1 + L_i \end{bmatrix} \tag{7.8}$$

$$\mathbf{T}^p_{i,i+1} = \frac{1}{2}\begin{bmatrix} \frac{n_i}{n_{i+1}}L_i + \frac{n_{i+1}}{n_i} & \frac{n_i}{n_{i+1}}L_i - \frac{n_{i+1}}{n_i} \\ \frac{n_i}{n_{i+1}}L_i - \frac{n_{i+1}}{n_i} & \frac{n_i}{n_{i+1}}L_i + \frac{n_{i+1}}{n_i} \end{bmatrix} \tag{7.9}$$

$$\mathbf{P}_i = \begin{bmatrix} e^{-j\beta_i d_i} & 0 \\ 0 & e^{j\beta_i d_i} \end{bmatrix} \tag{7.10}$$

Notice that all m matrices can be defined with simple knowledge of the refractive indices and thicknesses of the layers. Most importantly, only $\mathbf{P}_i$ in Eq. 7.10 is dependent on the layer thicknesses, d_i. Therefore, using Eq. 7.7 we may solve for the Fresnel coefficients above, Eq. 7.1 and Eq. 7.2.

Inverse design formulation

Consider a merit function, f, that is explicitly a function of the Fresnel coefficients:

$$f\left(r^s, r^p, t^s, t^p\right) : \mathcal{C}^4 \to \mathcal{R} \tag{7.11}$$

where r^s, r^p, t^s, and t^p are the (complex) Fresnel coefficients of a thin-film device in the transverse-electric (TE) and transverse-magnetic (TM) polarizations. Note that the Fresnel coefficients are defined at a specific photon energy, $\hbar\omega$, and incident angle of light, θ. For brevity we will assume that $\hbar\omega$ and θ are fixed. We will show how generalizing this method to a range of energies and angles can be done in the next section.

We assume that the parameters of interest in our thin-film device are the m layer thicknesses, which will be denoted by thickness vector, $\mathbf{d}$ with:

$$\mathbf{d} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_{m-1} \\ d_m \end{bmatrix} \tag{7.12}$$

where $d_1, ..., d_m$ are the thicknesses of each layer. This derivation will not consider optimization of the layer refractive indices.

We are interested in optimizing f with respect to the layer thicknesses. Therefore our optimization problem is given by:

$$\min_{\mathbf{d}} f\left(r^s, r^p, t^s, t^p\right) \text{ s.t. Eq. 7.1 and Eq. 7.2} \tag{7.13}$$

where this design problem reads "minimize f with respect to the layer thicknesses $\mathbf{d}$ subject to the Fresnel coefficients satisfying the transfer-matrix equations".

Eq. 7.13 is very similar to the form of the inverse problem defined in the previous chapter for general electromagnetics problems, except we have exchanged Maxwell's Equations with the transfer-matrix solution. Thus, following in those footsteps, we will formulate a gradient descent algorithm to solve 7.13

by taking the chain rule gradient of f. Consider the partial derivative of f with respect to layer thickness d_i:

$$\frac{\partial f}{\partial d_i} = 2\text{Re}\left(\frac{\partial f}{\partial r^s}\frac{\partial r^s}{\partial d_i}\right) + 2\text{Re}\left(\frac{\partial f}{\partial r^p}\frac{\partial r^p}{\partial d_i}\right) + 2\text{Re}\left(\frac{\partial f}{\partial t^s}\frac{\partial t^s}{\partial d_i}\right) + 2\text{Re}\left(\frac{\partial f}{\partial t^p}\frac{\partial t^p}{\partial d_i}\right) \tag{7.14}$$

where the 2Re terms account for the complex conjugate partial derivatives. By assumption above, f is defined in terms of the Fresnel coefficients. Therefore, the partial derivative terms with respect to the Fresnel coefficients ($\frac{\partial f}{\partial r^s}$ and similar) are known. Thus, it remains to calculate the $\frac{\partial r^s}{\partial d_i}$ terms and similar. The derivatives of reflection Fresnel coefficient in terms of the transfer-matrix elements can be found using Eq. 7.1:

$$\frac{\partial r}{\partial d_i} = \frac{\partial}{\partial d_i}\left(\frac{M_{21}}{M_{11}}\right) = \frac{M_{11}\frac{\partial M_{21}}{\partial d_i} - M_{21}\frac{\partial M_{11}}{\partial d_i}}{M_{11}^2} \tag{7.15}$$

$$\frac{\partial r}{\partial d_i} = t\frac{\partial M_{21}}{\partial d_i} - rt\frac{\partial M_{11}}{\partial d_i} \tag{7.16}$$

where we applied the quotient rule and rearranged. Similarly for t we may use Eq. 7.2:

$$\frac{\partial t}{\partial d_i} = \frac{\partial}{\partial d_i}\left(\frac{1}{M_{11}}\right) = \frac{-\frac{\partial M_{11}}{\partial d_i}}{M_{11}^2} \tag{7.17}$$

$$\frac{\partial t}{\partial d_i} = -t^2\frac{\partial M_{11}}{\partial d_i} \tag{7.18}$$

where we applied the quotient rule and rearranged again. The Fresnel coefficients r and t may be obtained by solving for the transfer-matrix in the forward direction. Therefore, to solve for these derivatives, we only need solve for the partial derivatives of the matrix elements, M_{11} and M_{21}. These partial derivatives can be obtained by taking the derivative of the full transfer-matrix $\frac{\partial \mathbf{M}}{\partial d_i}$.

Returning to the definition of the transfer-matrix [1]:

$$\mathbf{M} = \prod_{i=0}^{m} \mathbf{T}_{i,i+1}\mathbf{P}_{i+1} \tag{7.19}$$

we found previously that only the $\mathbf{P}_i$ matrices depended on the layer thicknesses in Eq. 7.10. Therefore, let us isolate the i-th $\mathbf{P}$ matrix by defining the following matrices:

$$\mathbf{M} = \mathbf{X}_i\mathbf{P}_i\mathbf{Y}_i \tag{7.20}$$

where,

$$\mathbf{X}_i \equiv \left(\mathbf{T}_{01}\prod_{j<i}\mathbf{P}_j\mathbf{T}_{j,j+1}\right) \tag{7.21}$$

[1] Where we have dropped the polarization dependence, because it will not matter hereafter.

$$\mathbf{Y}_i \equiv \left(\mathbf{T}_{i,i+1} \prod_{k>i} \mathbf{P}_k \mathbf{T}_{k,k+1} \right) \tag{7.22}$$

Therefore, when we take the i-th partial derivative of $\mathbf{M}$, we may conveniently write it as follows:

$$\frac{\partial \mathbf{M}}{\partial d_i} = \mathbf{X}_i \frac{\partial \mathbf{P}_i}{\partial d_i} \mathbf{Y}_i \tag{7.23}$$

Finally, we consider the partial derivative $\frac{\partial \mathbf{P}_i}{\partial d_i}$. Using Eq. 7.10 we find:

$$\frac{\partial \mathbf{P}_i}{\partial d_i} = \frac{\partial}{\partial d_i} \begin{bmatrix} e^{-j\beta_i d_i} & 0 \\ 0 & e^{j\beta_i d_i} \end{bmatrix} \tag{7.24}$$

$$= \begin{bmatrix} -j\beta_i e^{-j\beta_i d_i} & 0 \\ 0 & j\beta_i e^{j\beta_i d_i} \end{bmatrix} \tag{7.25}$$

$$= j\beta_i \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} e^{-j\beta_i d_i} & 0 \\ 0 & e^{j\beta_i d_i} \end{bmatrix} \tag{7.26}$$

$$\frac{\partial \mathbf{P}_i}{\partial d_i} = j\beta_i \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{P}_i \tag{7.27}$$

Inserting this in the full-derivative of $\mathbf{M}$ in Eq. 7.23 we have:

$$\frac{\partial \mathbf{M}}{\partial d_i} = j\beta_i \mathbf{X}_i \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{P}_i \mathbf{Y}_i \tag{7.28}$$

However, notice that $\mathbf{P}_i \mathbf{Y}_i$ corresponds to the right hand side of Eq. 7.20:

$$\mathbf{P}_i \mathbf{Y}_i = \mathbf{X}_i^{-1} \mathbf{M} \tag{7.29}$$

We may then replace this above to find our final derivative:

$$\frac{\partial \mathbf{M}}{\partial d_i} = j\beta_i \mathbf{X}_i \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{X}_i^{-1} \mathbf{M} \tag{7.30}$$

where the final derivative of $\mathbf{M}$ is a simple rearranging of quantities that have already been calculated during the forward simulation (namely, $\mathbf{X}_i$ and M). This concludes the derivation of the transfer-matrix inverse design method.

Summary and algorithm

To summarize, we introduced the transfer-matrix method to solve for the Fresnel coefficients of a one-dimensional thin-film device. We then presented a general inverse design problem using this method in Eq. 7.13. To solve this problem, we proposed using gradient descent. We showed the chain rule partial derivatives of the merit function in Eq. 7.14, which relied upon partial derivatives of the Fresnel coefficients in Eq. 7.16 and Eq. 7.18. These, in turn, required a partial derivative of the transfer-matrix which we found in Eq. 7.30. The algorithm for obtaining the full gradient of the merit function is then provided as follows:

1. Solve for the transfer-matrix of the system using Eq. 7.7. During this solution, save the $\mathbf{X}_i$ matrices defined in Eq. 7.21.
2. Use the resulting Fresnel coefficients to solve for the merit function (Eq. 7.13).
3. Use the $\mathbf{X}_i$ matrices and $\mathbf{M}$ to solve for the partial derivative of $\mathbf{M}$ with respect to the design parameters (Eq. 7.30).
4. Solve for the partial derivatives of r and t with respect to the design parameters in Eq. 7.16 and Eq. 7.18.
5. Solve for the full chain-rule derivative of the merit function in Eq. 7.14.

We will apply this algorithm in the next section.

7.2 Inverse design of an extremely broadband distributed Bragg reflector for thermophotovoltaics applications

Thermophotovoltaics is a burgeoning field of research for energy production, harvesting, and storage. As the name implies, a photovoltaic cell is used to generate electricity from a hot thermal source (i.e. a blackbody). This could include the hot engine of an aircraft, a radioactive source, or a thermally-isolated supply of heated graphite for energy storage. In our example, we will assume a perfect blackbody source emitting the Planck spectrum at $T = 1200°\text{C}$, where the Planck spectrum is given by:

$$b(\hbar\omega, T) = \frac{(\hbar\omega)^2}{4\pi^2 c^2 \hbar^3} \frac{1}{\exp\left(\frac{\hbar\omega}{kT} - 1\right)} \tag{7.31}$$

The central problem facing thermophotovoltaic applications is that photovoltaic cells can only absorb above-bandgap photons with $\hbar\omega \geq E_g$, where $E_g = 0.75\text{eV}$ for a typical InGaAs photvoltaic cell. However, at the relatively low temperature that we are considering (compared to the Sun, for example), a large amount of thermal radiation is emitted at below bandgap energies ($\hbar\omega < E_g$). One method to resolve this problem is to recycle below bandgap radiation by reflecting it back to the source. Recently, the thermophotovoltaic efficiency record was attained using this by Zomair et al [106], ultimately achieving a power conversion efficiency of 29.1%. In particular, an InP/InGaAs-based photovoltaic cell ($E_g = 0.75\text{eV}$) was used with a gold-backed mirror, which has an average reflectivity of $\approx 95\%$ in the below bandgap energy range. However, if this reflectivity can be improved to 99% or better, thermophotovoltaic efficiency exceeding 50% is theoretically possible [105].

The merit function defining our problem is the average reflectivity over all sub-bandgap photon energies ($\hbar\omega < E_g$) and incident angles from the full hemisphere above the device. This merit function can be denoted quantitatively by:

$$R_{\text{average}} \equiv \frac{\iint R(\omega, \theta) b(\omega) 2\pi \sin\theta \cos\theta d\theta d\omega}{\iint b(\omega) 2\pi \sin\theta \cos\theta d\theta d\omega} \tag{7.32}$$

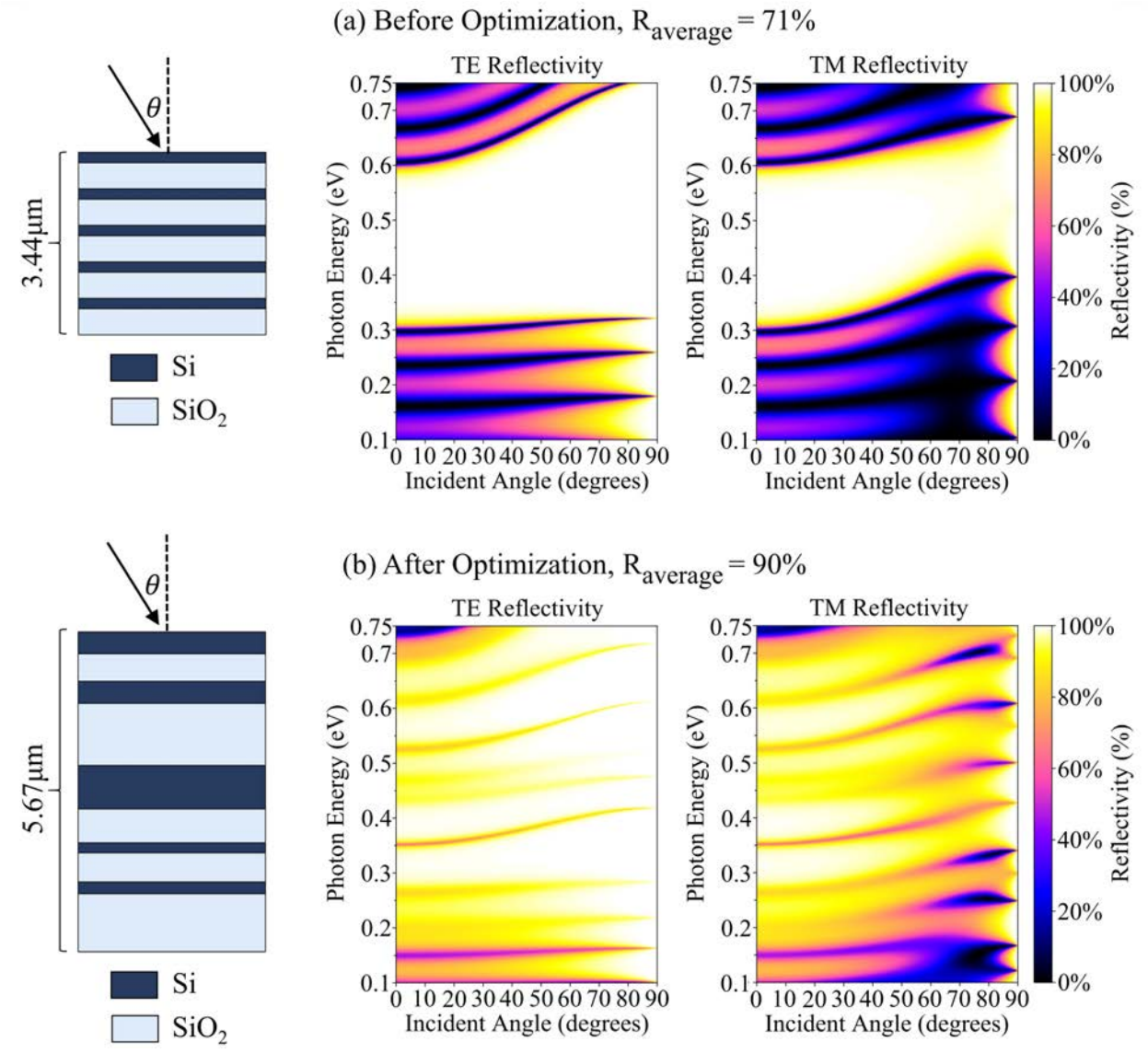

Figure 7.1: Optimization of a distributed Bragg reflector with alternating silicon and silica layers. A simple DBR achieves only 70% average reflectivity. After optimization, the bandwidth of the DBR can be expanded dramatically resulting in 90% reflectivity.

where $R(\omega, \theta)$ is the reflectivity at a particular angle of incidence (θ) and frequency (ω), and $b(\omega)$ is the Planck spectrum from Eq. 7.31. The $2\pi \sin\theta d\theta$ factor accounts for the differential solid angle subtended by the source at an angle θ with respect to normal. The $\cos\theta$ factor comes from Lambert's Cosine Law, where we assume the typical case that the blackbody source is a diffuse radiator. Thus, Eq. 7.32 can be regarded as a weighted average of the reflectivity. One additional average that must be taken into account is polarization, since on average the blackbody source is unpolarized and therefore the reflectivity may be written in terms of equal combinations of the TE and TM (s and p) polarizations: $R(\omega, \theta) = \frac{1}{2}(R^s(\omega, \theta) + R^p(\omega, \theta))$. where the $1/2$ factor takes into account the unpolarized light.

Distributed Bragg reflectors (DBRs) are known to provide high reflectivity with very little loss by exploiting thin-film interference. DBRs consist of low-loss dielectric layers with alternating refractive index. While near unity reflectivity can be achieved at a single wavelength with a few-layer DBR, a large number of layers are required to increase the reflectivity bandwidth. Consider the case of a 10 layer (5 pair) DBR in Fig. 7.1(a), which consists of alternating pairs of silicon and silicon dioxide floating in air. The right-hand side Fig. 7.1(a) shows the TE (s) and TM (p) reflectivity as functions of the incident angle and frequency of light. In the central band of the DBR, the reflectivity is near unity. However, the reflectivity quickly decreases outside of this band. Consequently, the average reflectivity according to Eq. 7.32 is just 70%,

The width of the reflectivity band (and thus the average reflectivity) can be increased by using more DBR layers. However, just increasing the number of layers provides diminishing returns for increasing the average reflectivity. Pragmatically speaking, we would like to limit the number of deposited layers in a process for reasons of cost and fabrication imperfection (such as deposition nonuniformity). Therefore, we may use the algorithm developed in this chapter to increase the reflectivity by optimizing the thicknesses of the DBR layers by gradient descent. Note that reflectivity can be rewritten as a function of the Fresnel reflectivity coefficient:

$$R(\omega, \theta) = r(\omega, \theta)\overline{r}(\omega, \theta) \tag{7.33}$$

where $\overline{r}$ is the complex conjugate of r. Therefore, using the formalism developed in Section 7.1, the derivative of $R(\omega, \theta)$ with respect to layer thickness d_i is given by:

$$\frac{\partial R(\omega, \theta)}{\partial d_i} = 2\mathrm{Re}\left(\frac{R(\omega, \theta)}{\partial r(\omega, \theta)}\frac{\partial r(\omega, \theta)}{\partial d_i}\right) \tag{7.34}$$

$$= 2\mathrm{Re}\left(\overline{r}(\omega, \theta)\frac{\partial r(\omega, \theta)}{\partial d_i}\right) \tag{7.35}$$

Note that $\frac{\partial r}{\partial d_i}$ may be found using Eq. 7.16 from the previous section. Thus, by linearity of derivatives, we may find the partial derivative of the average reflectivity from Eq. 7.32:

$$\frac{\partial R_{\text{average}}}{\partial d_i} \equiv \frac{\iint 2\mathrm{Re}\left(\overline{r}(\omega, \theta)\frac{\partial r(\omega, \theta)}{\partial d_i}\right) b(\omega) 2\pi \sin\theta \cos\theta d\theta d\omega}{\iint b(\omega) 2\pi \sin\theta \cos\theta d\theta d\omega} \tag{7.36}$$

In other words, we take the weighted average of the individual reflectivity derivatives. We apply the algorithm developed in Section 7.1 along with the gradients generated using this method to the uniform DBR

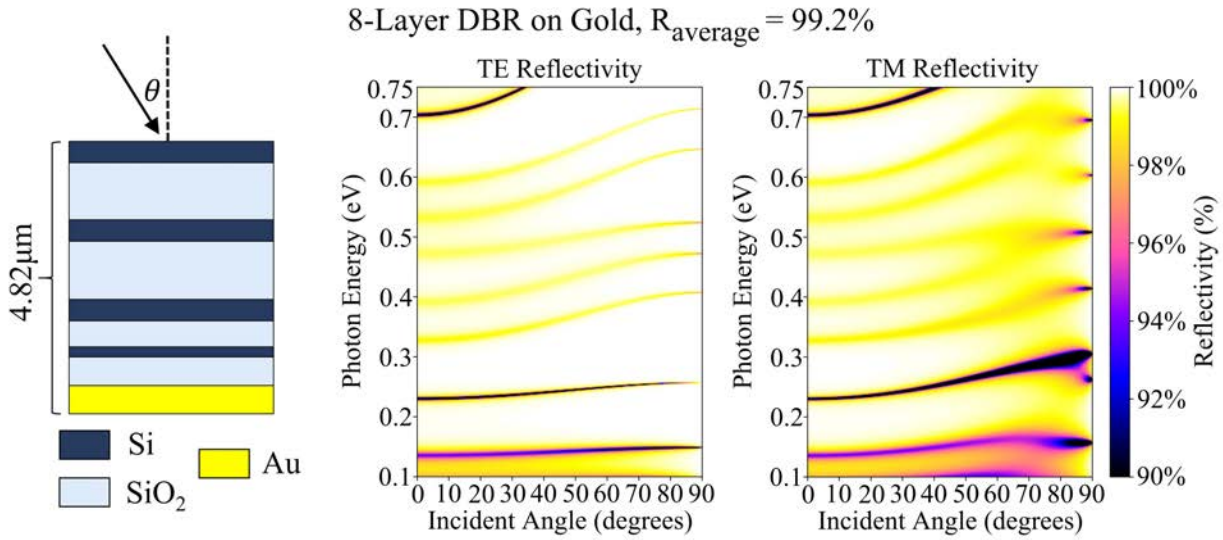

Figure 7.2: Result of an inverse design optimization of a distributed Bragg reflector on a gold substrate. Using just 8 total layers of alternating silicon and silica, we can achieve ¿99% reflectivity over a broad range.

from Fig. 7.1. Doing so, we achieve a final thickness profile and reflectivity in Fig. 7.1(b). This improves the average reflectivity to 90%, with a drastic change to the reflectivity across the broad range of incident photon angles and energies.

While Fig. 7.1(b) demonstrates a massive improvement over a simple DBR, 90% reflectivity is still not adequate nor competitive with a simple gold substrate for thermophotovoltaics. Thus, we also considered a DBR deposited on a gold substrate in order to achieve reflectivity in excess of 99%. In Ref [105], Omair et al showed that over 99% reflectivity could be achieved using a DBR with 40 Bragg pairs (80 layers) of alternating silicon and silicon dioxide. Furthermore, using inverse design optimization, we can achieve over 99% reflectivity with far fewer pairs. The result of an optimization using a distributed Bragg reflector on a gold substrate consisting of just 4 Bragg pairs (8 layers) is shown in Fig 7.2. Note the color scale on the reflectivity plots ranges from 90% to 100% as opposed to the color scale in the previous figure. Thus, we achieve an extremely reflective mirror with over 99% average reflectivity. Such a mirror could enable next-generation ultra-efficient thermophotovoltaics with conversion efficiency exceeding 50%.

7.3 Machine learning enhanced inverse design

In this section we will briefly describe an emerging class of inverse design methods that leverage machine learning techniques to potentially improve electromagnetic design capabilities. These methods seek to replace conventional gradient descent optimization (Eq. 6.17 from Chapter 6) with neural networks, which can potentially improve the search over the design space of interest.

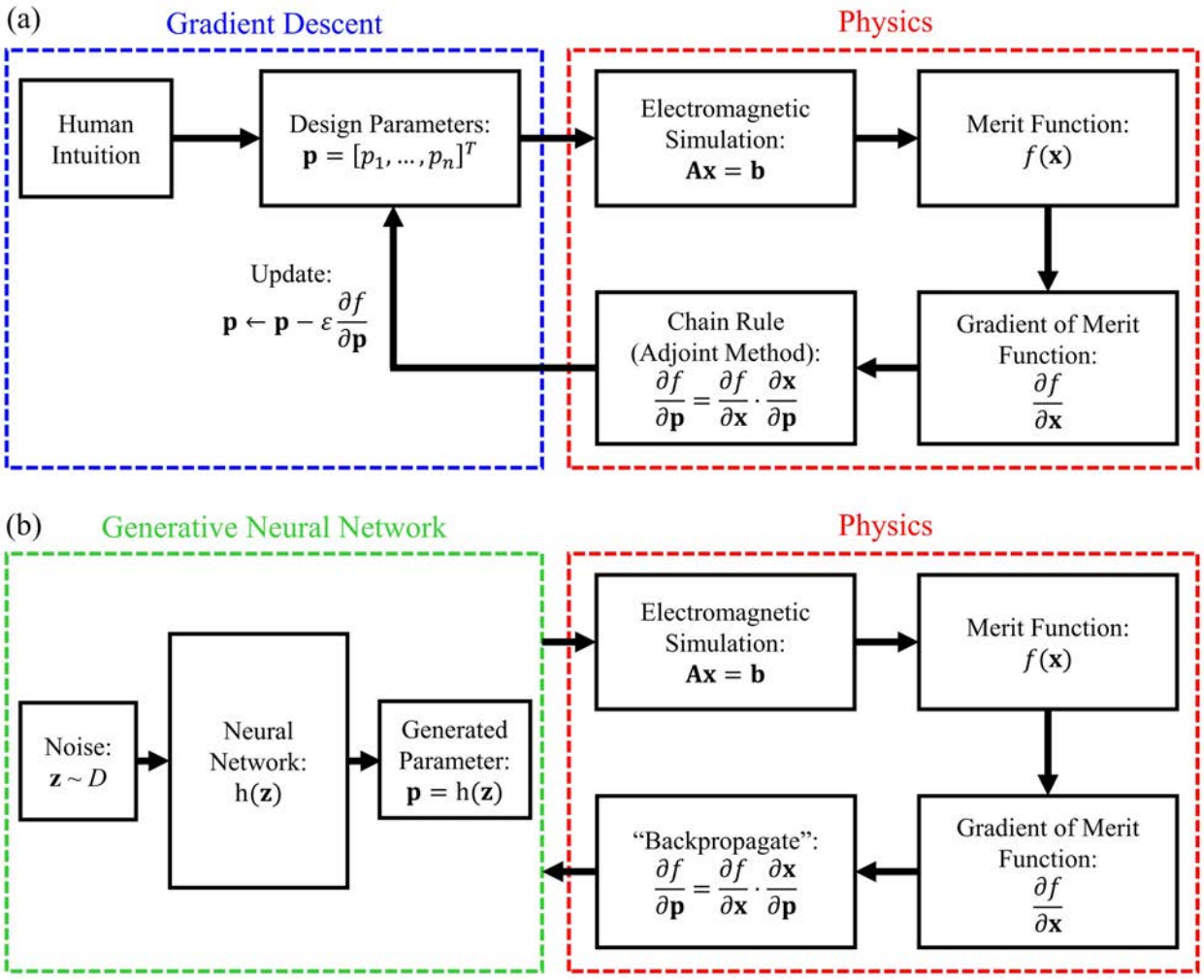

Figure 7.3: Neural network aided inverse electromagnetic design is a natural extension of the adjoint method. A conventional gradient-descent optimization loop interfaced with electromagnetic physics is provided in (a). Gradient-descent is replaced by a generative neural network in (b), which learns to generate design parameters using information from the physics solver.

This is illustrated qualitatively in Fig. 7.3. In Fig. 7.3(a) we depict the typical gradient-descent optimization loop for inverse design via the adjoint method. In the first step, a user selects candidate device parameters. These are subsequently simulated in a physics solver (such as FDTD), a merit function is calculated, and the gradient of that merit function is computed using the adjoint method. This gradient is used to update the design parameters using a deterministic update equation such as gradient descent.

By contrast, we may replace the gradient descent part of the optimization loop with a generative neural network, as depicted in Fig. 7.3(b). A generative neural network takes in noise as input, and performs a complex nonlinear function over that noise to output design parameters for the physics engine. By providing feedback from the merit function of interest as well as its gradient, the generative neural network will automatically learn to generate a distribution of device parameters that satisfy the desired merit function after many iterations. Generative neural networks are out-of-scope for this thesis, but a detailed discussion of this design method can be found in Ref. [57] where it originates for electromagnetic design.

Using the generative neural network method shown in Fig. 7.3(b), we hope to overcome a common issue in typical gradient-based design methods; namely, they are only capable of finding local optima. Local optima are regions in a design space where the gradient of the merit function is locally equal to zero. While these optima can be good, the global optimum cannot be found generally by gradient descent. Using neural networks, we might hope that the machine is capable of learning additional information and patterns in the design space that a human would otherwise be unable to discern, providing better optimization. Nevertheless, there is still no guarantee of finding a global optimum even when using neural networks, but results in the literature [57] and preliminary results by the current author are promising.

We applied a novel version of the generative neural network method[2] in Fig. 7.3(b) to the design of distributed Bragg reflectors for broadband reflection. This is shown in Fig. 7.4 where we compare the results of the neural network to the gradient descent method for DBR design detailed in the previous section. In this case, we optimize the DBR for a variable number of layers on a gold substrate, illustrated in the diagram on the left-hand side. Interestingly, the neural network is able to achieve more robust optima than conventional gradient descent. In particular, regular gradient descent provides diminishing returns for the 6- and 10- layer designs, indicating the solver was caught in a local optimum. By contrast, the mirrors produced by the neural network method demonstrate increased reflectivity with each addition of new layers – as one might expect should be possible given additional degrees of freedom. The best result achieved was the 10-layer neural network with an average reflectivity exceeding 99.5% over the frequency range of interest.

While these results are promising, a significant amount of trial-and-error is required to get neural networks working adeptly (commonly referred to as "hyperparameter tuning"). Consequently, neural networks that are capable of solving large-scale computational problems remain to be shown. Future work in this field will likely need to demonstrate neural networks that can apply to general electromagnetic design problems and optimize well with minimal user input.

[2] Work on this novel method is still in progress at the time of this thesis, and is not described in here for brevity.

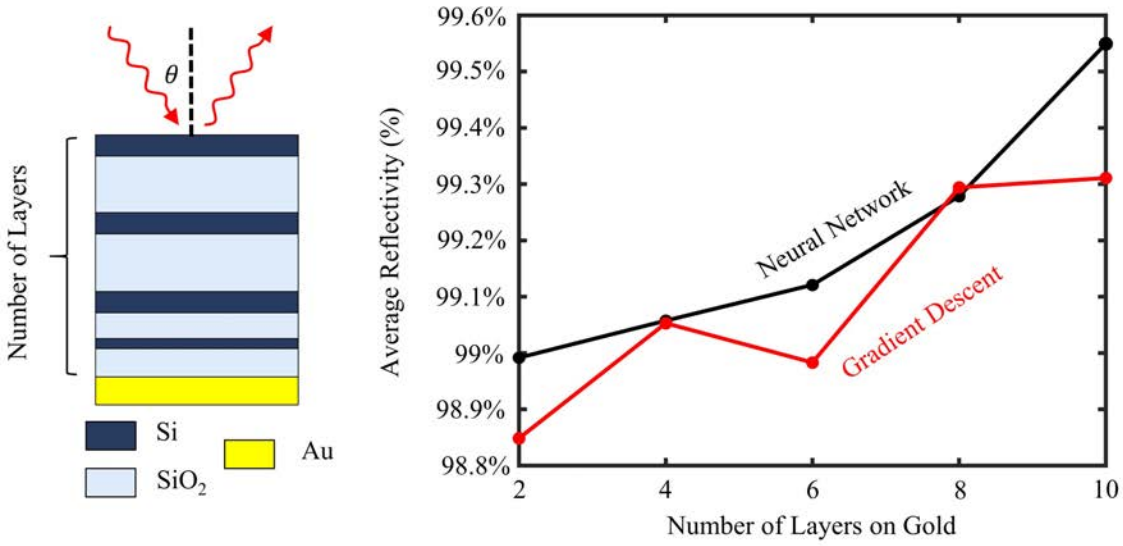

Figure 7.4: Neural network aided design of broadband distributed Bragg reflectors provides more robust solutions than regular gradient descent. The number of DBR layers on gold is illustrated qualitatively on the left-hand side of the diagram.

Chapter 8

Conclusion

The contributions of this thesis are summarized as follows:

1. We quantified the stimulated emission carrier lifetime in a saturated edge-emitting InGaAs laser using a novel method, ultimately finding a lifetime of 6ps.

2. We demonstrated that the spontaneous emission carrier lifetime in heavily-doped light-emitting diodes saturates to the fundamental spontaneous emission lifetime of a two-level system, which is 1ns for InGaAs. Consequently, we rejected the standard BNP model of the LED recombination rate.

3. We showed that optical antenna-enhanced spontaneous emission can be as fast as stimulated emission.

4. We examined the large-signal direct-electrical modulation rate of antenna-LEDs and lasers, finding that both devices are limited by their respective carrier lifetimes. Thus, we argued that antenna-LEDs can be fast as lasers.

5. We demonstrated that optical antenna-LEDs can reach practical and competitive internal quantum efficiency if III-V semiconductor surface treatment processes are improved.

6. A full simulation of the antenna-LED transmitter revealed that it is capable of direct-electrical modulation exceeding 50Gbit/s with 500 photons/bit signal.

7. We demonstrated novel metal-dielectric antennas that can overcome an efficiency versus antenna enhancement tradeoff using sharp dielectric tips.

8. We provided a tutorial of inverse design via the adjoint method.

9. Using inverse design, we showed that antenna-LEDs are capable of 94% single-mode waveguide coupling efficiency.

10. We demonstrated fabrication-friendly vertical grating couplers using inverse design, with industry-competitive insertion loss approaching 0.5dB and dense wavelength division multiplexing compatible bandwidth.

11. A novel semi-analytical transfer-matrix inverse design method for 1D interference filters was presented and applied to the design of ultra-broadband distributed Bragg reflectors with $>99\%$ average reflectivity.

12. We explored novel inverse design methods that leverage machine learning.

Based on the arguments in this thesis, ultra-fast optical antenna-LEDs may be feasible in next-generation on-chip optical interconnects. Furthermore, inverse electromagnetic design will continue to enable remarkable design capabilities, pushing the limits of photonic device engineering.

Appendix A

Microscopic Origin of Spontaneous and Stimulated Emission

There are several methods to derive the rate of spontaneous and stimulated emission from atoms and semiconductors. Perhaps the most fundamental comes from Fermi's Golden Rule:

$$\Gamma_{i\to f} = \frac{2\pi}{\hbar} \left| \langle f|H'|i\rangle \right|^2 \rho(E_f) \tag{A.1}$$

where H' is a perturbation Hamiltonian that causes a transition, with probability per unit time $\Gamma_{i\to f}$ of a transition from the the (joint) eigenstate i to a final (joint) eigenstate f. $\left| \langle f|H'|i\rangle \right|^2$ is known as the matrix element of this perturbation and $\rho(E_f)$ is called the density of final states (typically in units of states per unit energy, evaluated at E_f). In the case of light-matter interactions, the most common perturbation is the electric dipole interaction potential $H' = q\mathbf{x}\cdot\mathbf{E}$, where q is the elementary charge, $\mathbf{x}$ is the dipole moment, and $\mathbf{E}$ is the electric field. By second quantization, the electric field may be expressed as an operator on photon number states with,

$$\mathbf{E} = i\sqrt{\frac{\hbar\omega}{2\varepsilon V}} \left(\mathbf{a}^\dagger e^{-i\mathbf{k}\cdot\mathbf{r}} + \mathbf{a}e^{i\mathbf{k}\cdot\mathbf{r}}\right)\hat{\mathbf{e}} \tag{A.2}$$

where $\mathbf{a}^\dagger$ and $\mathbf{a}$ are the photon creation and destruction operators respectively, ω is the frequency of light, ε is the material permittivity, $\mathbf{k}$ is the wavevector, $\hat{\mathbf{e}}$ is the polarization direction, and V is a normalization volume. We can then express the joint eigenstates in terms of electron initial and final states, i and f, as well as photon number states, N_p and $N_p + 1$, since the electron transition either creates or destroys a photon. Hence,

$$\left| \langle f|H'|i\rangle \right|^2 \to \left| \langle f, N_p+1|q\mathbf{x}\cdot\mathbf{E}|i, N_p\rangle \right|^2 \tag{A.3}$$

$$= \frac{\hbar\omega}{2\varepsilon V} \left| \langle f, N_p+1|q\mathbf{x}\cdot\left(\mathbf{a}^\dagger e^{-i\mathbf{k}\cdot\mathbf{r}} + \mathbf{a}e^{i\mathbf{k}\cdot\mathbf{r}}\right)\hat{\mathbf{e}}|i, N_p\rangle \right|^2 \tag{A.4}$$

Since we are interested in light emission, we may drop the annihilation term which corresponds to absorption. Furthermore, we may employ the slowly-varying amplitude approximation by assumption that the

dipole moment is much shorter than the wavelength of light, allowing us to separate the joint eigenstates and remove the exponential term:

$$|\langle f|H'|i\rangle|^2 \to \frac{\hbar\omega}{2\varepsilon V}\,|\langle f|q\mathbf{x}\cdot\hat{\mathbf{e}}|i\rangle|^2\,\left|\langle N_p+1|a^\dagger|N_p\rangle\right|^2 \tag{A.5}$$

$$= \frac{\hbar\omega}{2\varepsilon V}(N_p+1)\,|\langle f|q\mathbf{x}\cdot\hat{\mathbf{e}}|i\rangle|^2 \tag{A.6}$$

where in the second step we used:

$$\left|\langle N_p+1|a^\dagger|N_p\rangle\right|^2 = \left|\langle N_p+1|\sqrt{N_p+1}|N_p+1\rangle\right|^2 = (N_p+1)|\langle N_p+1|N_p+1\rangle|^2 = (N_p+1) \tag{A.7}$$

This allows us to separate the matrix element into two terms,

$$|\langle f|H'|i\rangle|^2 = \frac{\hbar\omega}{2\varepsilon V}N_p\,|\langle f|q\mathbf{x}\cdot\hat{\mathbf{e}}|i\rangle|^2 + \frac{\hbar\omega}{2\varepsilon V}\,|\langle f|q\mathbf{x}\cdot\hat{\mathbf{e}}|i\rangle|^2 \tag{A.8}$$

where the term on the left represents stimulated emission (due to the presence of N_p photons), and the term on the right represents spontaneous emission (in absence of photons, $N_p = 0$). Furthermore, the matrix element now depends only on the transition dipole matrix element of the material. This demonstrates the fundamental relationship between stimulated and spontaneous emission.

Taking only the term for spontaneous emission, and expressing the optical density of states as:

$$\rho(\hbar\omega) = V\frac{\omega^2 n^3}{\pi^2\hbar c^3} \tag{A.9}$$

We find,

$$\Gamma_{i\to f} = \frac{2\pi}{\hbar}\frac{\hbar\omega}{2\varepsilon V}\,|\langle f|q\mathbf{x}\cdot\hat{\mathbf{e}}|i\rangle|^2\,V\frac{\omega^2 n^3}{\pi^2\hbar c^3} \tag{A.10}$$

$$= \frac{\omega^3 n}{\hbar\varepsilon_{\mathrm{o}}c^3}\,|\langle f|q\mathbf{x}\cdot\hat{\mathbf{e}}|i\rangle|^2 \tag{A.11}$$

Conventionally, we average over all possible polarizations of the electric field and the dipole moment directions, allowing us to remove the dot product;

$$\Gamma_{i\to f} = \frac{|qx_{i\to f}|^2\omega^3 n}{3\hbar\varepsilon_{\mathrm{o}}c^3} \tag{A.12}$$

where $x_{i\to f}$ is now a scalar operator and determined by the material. In the main body of this thesis, we express this final transition rate in terms of a spontaneous emission lifetime, $1/\tau_{\mathrm{o}} = \Gamma_{i\to f}$. Furthermore, we simplify the notation of the dipole matrix element with $|qx_{i\to f}| \to |qx_{21}|$. Thus, we may write:

$$\frac{1}{\tau_{\mathrm{o}}} = \frac{|qx_{21}|^2\omega^3 n}{3\hbar\varepsilon_{\mathrm{o}}c^3} \tag{A.13}$$

This concludes the proof.

Appendix B

Detailed Model of Spontaneous and Stimulated Emission with Arbitrary Doping Concentration

Several times throughout this thesis we have provided simulations of LEDs, lasers, and antenna-LEDs. In particular, we claimed that under heavy doping the LED spontaneous emission carrier lifetime saturates to the fundamental lifetime of a two-level system. In this Appendix, we will briefly describe how these simulations were constructed. In particular, the models follow from Chuang [21], but allow for arbitrary doping concentration. The code may be found at [45].

Step 1: Self-consistently solve for quasi-Fermi levels

Knowing the desired doping concentration, we can solve for the electron and hole concentrations with zero applied voltage. For example, if the hole dopant density $N_A \gg n_i$ where n_i is the intrinsic carrier concentration at room temperature with zero doping, then it is well known that:

$$P_{\text{o}} \approx N_A \tag{B.1}$$

$$N_{\text{o}} \approx \frac{n_i^2}{N_A} \tag{B.2}$$

Then, we may express the full carrier concentrations N and P by:

$$N = N_{\text{o}} + N' \tag{B.3}$$

$$P = P_{\text{o}} + P' \tag{B.4}$$

where N' and P' are the excess carrier concentrations under an applied voltage. By charge conservation, these must be equal:

$$N' = P' \tag{B.5}$$

$$\Rightarrow N - N_{\text{o}} = P - P_{\text{o}} \tag{B.6}$$

Keeping this in mind, we may use the definition of the carrier concentrations parameterized by the quasi-Fermi level in the respective conduction and valence bands, where we have:

$$N = N_c F_{1/2}\left(\frac{F_c - E_g}{kT}\right) \tag{B.7}$$

$$P = N_v F_{1/2}\left(\frac{-F_v}{kT}\right) \tag{B.8}$$

where N_c and N_v are the effective conduction band and valence band density of states respectively, $F_{1/2}$ is the Fermi-Dirac integral of order 1/2, F_c and F_v are the conduction band and valence band quasi-Fermi levels, E_g is the bandgap, and kT is the Boltzmann factor. Note that $F_{1/2}$ only applies for bulk semiconductors, but can be changed for quantum wells if necessary. Furthermore, it applies regardless of whether the semiconductor is degenerately or non-degenerately doped. Thus, using Eq. B.2–B.8 and by definition,

$$V = \Delta F = F_c - F_v \tag{B.9}$$

we may self-consistently (numerically) solve for the carrier concentrations of a semiconductor with arbitrary doping and applied voltage, because we have 4 equations and 4 unknowns (N, P, F_c, and F_v).

Step 2: Solve for the gain and spontaneous emission spectra

The (net) gain spectrum is given by [21]:

$$g(\hbar\omega) = \frac{\pi q^2 M_b^2}{nc\varepsilon_o m_o^2 \omega}\rho_r(\hbar\omega)\left(f_c - f_v\right) \tag{B.10}$$

where M_b^2 is the transition momentum matrix element, m_o is the electron mass, $\hbar\omega$ is the photon energy, f_c and f_v are the Fermi-Dirac functions (parameterized by the quasi-Fermi levels F_c and F_v), and ρ_r is the reduced density of states for a bulk crystal. These quantities are defined below:

$$\rho_r(\hbar\omega) = \frac{1}{2\pi^2}\left(\frac{2m_r^*}{\hbar^2}\right)^{3/2}\sqrt{\hbar\omega - E_g} \tag{B.11}$$

$$f_c = \frac{1}{\exp\{(E_2 - F_c)/kT\} + 1} \tag{B.12}$$

$$f_v = \frac{1}{\exp\{(E_1 - F_v)/kT\} + 1} \tag{B.13}$$

$$E_2 = E_c + \frac{m_r^*}{m_e^*}(\hbar\omega - E_g) \tag{B.14}$$

$$E_1 = E_v - \frac{m_r^*}{m_h^*}(\hbar\omega - E_g) \tag{B.15}$$

$$\hbar\omega = E_2 - E_1 \tag{B.16}$$

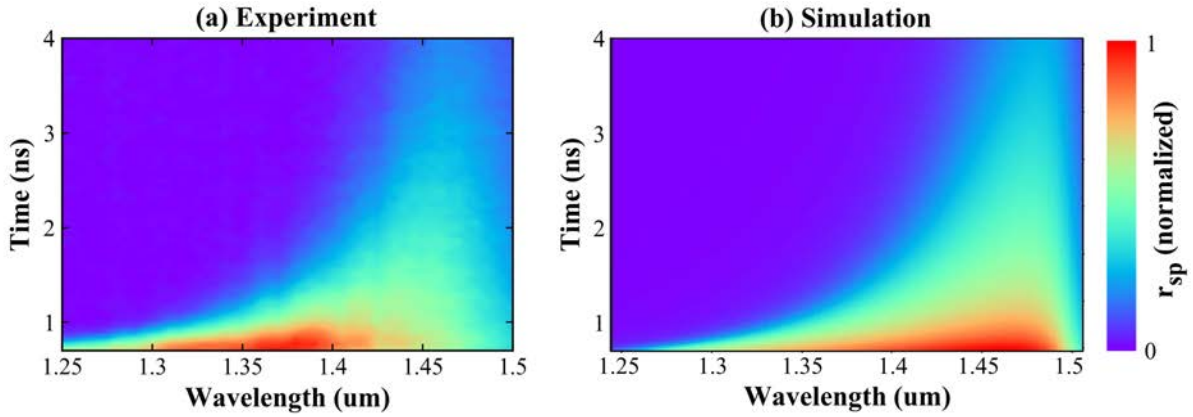

Figure B.1: Spontaneous emission spectrum versus time simulation shows excellent agreement with experiment. Experimental spectrum (a) was obtained by Fortuna [36] who used a time-correlated single-photon counting (TCSPC) setup along with a spectral filter. Simulation (b) was performed by using the analysis developed in this section to obtain the spontaneous emission spectrum (Eq. B.18) as a function of minority carrier concentration. The carrier dynamics with time were obtained using the spontaneous emission recombination rate (Eq. B.19) along with an assumed surface recombination lifetime of 4ns.

where m_e^* and m_h^* are the electron and hole effective masses, m_r^* is the reduced effective mass, E_c and E_v are the conduction and valence band energies, and E_2 and E_1 are energy levels in the conduction and valence bands parameterized by the photon energy $\hbar\omega$ after assuming conservation of momentum. Then, the spontaneous emission spectrum can be obtained by a simple transformation:

$$r_{\text{sp}}(\hbar\omega) = \left(\frac{n^2\omega^2}{\pi^2\hbar c^2}\right) g(\hbar\omega) \frac{1}{1-\exp\{(\hbar\omega-(F_c-F_v))/kT\}} \tag{B.17}$$

which follows from Einstein's AB analysis. Interestingly, the spontaneous emission spectrum can also be rewritten in the form:

$$r_{\text{sp}}(\hbar\omega) = \frac{1}{\tau_{\text{o}}(\hbar\omega)} \rho_r(\hbar\omega) f_c(1-f_v) \tag{B.18}$$

where $\tau_{\text{o}}(\hbar\omega)$ is the fundamental spontaneous emission lifetime of a two-level system (from Appendix A). Note that both the gain and spontaneous emission spectra are implicitly defined at given carrier concentration and doping, because f_c and f_v are defined by the quasi-Fermi levels from above.

Step 3: Solve for the spontaneous emission recombination rate and carrier lifetime

Finally, we solve for the recombination rate due to spontaneous emission using:

$$R_{\text{sp}} = \int r_{\text{sp}}(\hbar\omega) d\hbar\omega \tag{B.19}$$

where r_{sp} is the spontaneous emission spectrum from above, and we integrate over all photon energies. The carrier lifetime is defined as:

$$\frac{1}{\tau_{\text{sp}}} = \frac{\partial R_{\text{sp}}}{\partial N} \tag{B.20}$$

where N refers to the minority carrier in this case. Note that we have $R_{\text{sp}}(N)$ implicitly defined, so this carrier lifetime derivative can be calculated simply by finite-difference.

Agreement of simulated spontaneous emission spectrum with experiment

Since we have indexed the spontaneous emission spectrum, $r_{sp}(\hbar\omega)$ as an implicit function of the carrier concentration, N, we may use carrier dynamics to solve for $r_{sp}(\hbar\omega)$ as a function of time. To do so, we use the rate equations developed in Appendix E. The result versus a TCSPC experiment for InGaAs is provided in Fig. B.1. We obtain excellent agreement with experiment.

Appendix C

Heavily-Doped LED Saturation

In this section we prove the claim that the semiconductor spontaneous emission lifetime saturates to the fundamental spontaneous emission lifetime of a two-level system in the limit of heavy doping and low-level injection. In doing so, we will make several simplifying assumptions. The generality of the solution under a full numerical calculation can be found in Appendix B and the radiative lifetime versus dopant density curve from Fig. 3.7 in the main manuscript.

In Appendix B we show that the total spontaneous emission recombination rate is given generally by,

$$R_{\text{sp}} = \int_{E_g}^{\infty} \frac{1}{\tau_{\text{o}}(\hbar\omega)} \rho_r(\hbar\omega) f_c(1 - f_v) d\hbar\omega \tag{C.1}$$

where E_g is the bandgap energy, τ_{o} is the fundamental spontaneous emission lifetime of a two-level system (from Appendix A), ρ_r is the reduced density of states, $\hbar\omega$ is the photon energy, and f_c,f_v are the Fermi-Dirac distributions in the conduction band and valence band. These quantities are defined symbolically below:

$$\rho_r(\hbar\omega) = \frac{1}{2\pi^2} \left(\frac{2m_r^*}{\hbar^2} \right)^{3/2} \sqrt{\hbar\omega - E_g} \tag{C.2}$$

$$\tag{C.3}$$

$$f_c = \frac{1}{\exp\{(E_2 - F_c)/kT\} + 1} \tag{C.4}$$

$$f_v = \frac{1}{\exp\{(E_1 - F_v)/kT\} + 1} \tag{C.5}$$

where E_2 and E_1 are defined as:

$$\begin{aligned} E_2 &= E_c + \frac{\hbar^2 k^2}{2m_e^*} \\ E_1 &= E_v - \frac{\hbar^2 k^2}{2m_h^*} \end{aligned} \tag{C.6}$$

where m_h^* and m_e^* are the hole and electron effective masses respectively, m_r^* is the reduced effective mass, F_v and F_c are the quasi-Fermi levels for the valence and conduction bands respectively, and E_1 and E_2 parameterize the energies of the valence and conduction bands by the wavevector k under the assumption they are parabolic. In other words, by conservation of energy and momentum, we have that the photon energy due to electron-hole radiative recombination is given by,

$$\hbar\omega = E_2 - E_1 = (E_c - E_v) + \frac{\hbar^2 k^2}{2}\left(\frac{1}{m_e^*} + \frac{1}{m_h^*}\right) = E_g + \frac{\hbar^2 k^2}{2m_r^*} \tag{C.7}$$

We may then solve for k^2 in terms of $\hbar\omega$ and the reduced mass m_r^* plug the result into Eq. C.6 to find,

$$\begin{aligned} E_1 &= E_v - \frac{m_r^*}{m_h^*}(\hbar\omega - E_g) \\ E_2 &= E_c + \frac{m_r^*}{m_e^*}(\hbar\omega - E_g) \end{aligned} \tag{C.8}$$

Now we may begin to simplify these equations under three assumptions:

1. The hole effective mass is much larger than the electron effective mass, $m_h^* \gg m_e^*$
2. The quasi-Fermi level for holes is far below the valence band edge, $F_v \ll E_v$, or equivalently the doping density is very large and degenerate, $P_\text{o} \to \infty$.
3. The quasi-Fermi level for electrons is below the conduction band edge, $F_c \ll E_c$, or equivalently the minority carrier concentration is nondegenerate and we are in a low-level injection pumping condition

These are all reasonable assumptions for an unstrained III-V LED under heavy doping conditions $N_A > 10^{19}\text{cm}^{-3}$, since the valence band will typically be dominated by the contribution of heavy-hole states. The E-k diagram of a direct-gap semiconductor representing each of these assumptions is shown in Fig. C.1 for reference. We now detail the effects of these assumptions below.

Assumption 1: $m_h^* \gg m_e^*$

When the hole effective mass is much larger than the electron effective mass, its contribution to the reduced effective mass becomes negligible:

$$\frac{1}{m_r^*} \equiv \frac{1}{m_e^*} + \frac{1}{m_h^*} \tag{C.9}$$

$$\frac{1}{m_r^*} \approx \frac{1}{m_e^*} \tag{C.10}$$

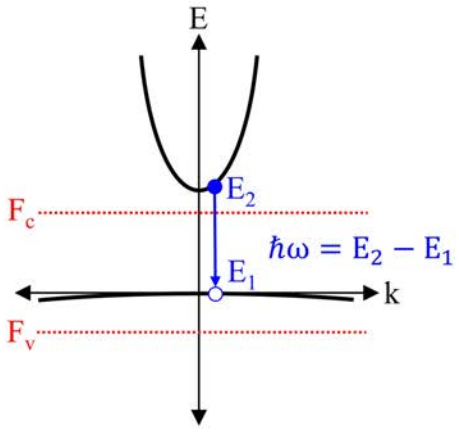

Figure C.1: E-k diagram representing our assumption that the hole effective mass is much larger than the conduction band effective mass.

Furthermore, the energy level equations in Eq. C.8 now become,

$$E_1 = E_v - \frac{m_r^*}{m_h^*}(\hbar\omega - E_g) \tag{C.11}$$

$$\approx E_v - \frac{m_e^*}{m_h^*}(\hbar\omega - E_g) \tag{C.12}$$

$$E_1 \approx E_v \tag{C.13}$$

where we used $m_h^* \gg m_e^*$. Similarly,

$$E_2 = E_c + \frac{m_r^*}{m_e^*}(\hbar\omega - E_g) \tag{C.14}$$

$$\approx E_c + \frac{m_e^*}{m_e^*}(\hbar\omega - E_g) \tag{C.15}$$

$$E_2 \approx E_c - E_g + \hbar\omega \tag{C.16}$$

Since we are only interested energy level differences, without loss of generality we may set the bandedge absolute energies to $E_v = 0$ and $E_c = E_g$, to which Eq. C.13 and Eq. C.16 become further simplified:

$$E_1 = 0 \tag{C.17}$$

$$E_2 = \hbar\omega \tag{C.18}$$

which is simply a statement that the valence band energy changes negligibly with k, so the energy differences $E_2 - E_1 = \hbar\omega$ occurs entirely in the conduction band.

Finally, using $m_r^* \approx m_e^*$, $E_c = E_g$, and $E_2 = \hbar\omega$, we may rewrite the joint density of states Eq. C.3, as,

$$\rho_r(\hbar\omega) \rightarrow \rho_c(E_2) = \frac{1}{2\pi^2}\left(\frac{2m_e^*}{\hbar^2}\right)^{3/2}\sqrt{E_2 - E_c} \tag{C.19}$$

which is exactly equal to the conduction band 3D density of states in the conduction band, ρ_c, under these assumptions.

Assumption 2: $F_v \ll E_v$

When the quasi-Fermi level is far below the valence band edge and using the assumption developed previously, $E_1 = E_v = 0$, the Fermi-Dirac probability distribution in the valence band becomes,

$$f_v = \frac{1}{\exp\{(E_1 - F_v)/kT\} + 1} \rightarrow 0 \tag{C.20}$$

which occurs because the exponential in the denominator becomes very large.

Assumption 3: $F_c \ll E_c$

Plugging Eq C.19, Eq. C.20, $E_2 = \hbar\omega$, and $E_c = E_g$ into the original spontaneous emission recombination rate integral from Eq. C.1, we may write:

$$R_{\text{sp}} \approx \int_{E_c}^{\infty} \frac{1}{\tau(E_2)}\rho_c(E_2) f_c dE_2 \tag{C.21}$$

Using the low-level injection assumption, we may approximate the Fermi-Dirac probability distribution in the conduction band as,

$$f_c = \frac{1}{\exp\{(E_2 - F_c)/kT\} + 1} \approx \exp\{(F_c - E_2)/kT\} \tag{C.22}$$

which is the common Boltzmann approximation. Substituting this back into Eq. C.21:

$$R_{\text{sp}} \approx \int_{E_c}^{\infty} \frac{1}{\tau_{\text{o}}(E_2)}\rho_c(E_2)\exp\{(F_c - E_2)/kT\}dE_2 \tag{C.23}$$

we can now make the observation that energies far above the bandedge $E_2 \gg E_c$ will not provide a meaningful contribution to the integral because of the exponential. Therefore, we may effectively truncate the upper limit of the integral after a few kT. Alternatively, we may effectively treat τ_{o} as a constant defined at the bandgap energy $\tau_{\text{o}}(\hbar\omega) \rightarrow \tau_{\text{o}}(E_g)$, and pull it out of the integral [1]. Therefore we have,

$$R_{\text{sp}} \approx \frac{1}{\tau_{\text{o}}(E_g)}\int_{E_c}^{\infty} \rho_c(E_2)\exp\{(F_c - E_2)/kT\}dE_2 \tag{C.24}$$

[1] Technically speaking, the average kinetic energy of holes in the conduction band is $\frac{3}{2}kT$, and similarly for holes in the valence ban, indicating that the lifetime should be defined at $\tau_{\text{o}}(\hbar\omega) \approx \tau_{\text{o}}(E_g + 3kT)$. This extra contribution is more-or-less negligible, and ignored for simplicity.

where the integral is now instantly recognizable: it represents the minority electron concentration in the conduction band under the Boltzmann concentration, N:

$$R_{\text{sp}} \approx \frac{N}{\tau_{\text{o}}(E_g)} \tag{C.25}$$

which is the desired behavior we wished to prove. We may can take this one step further to get the carrier lifetime:

$$\frac{1}{\tau_{\text{sp}}} \equiv \frac{\partial R_{\text{sp}}}{\partial N} = \frac{1}{\tau_{\text{o}}(E_g)} \tag{C.26}$$

or, in other words, the spontaneous emission lifetime of semiconductors becomes the fundamental spontaneous emission lifetime of a two-level system under conditions of heavy doping and low-level injection. This concludes the proof.

Appendix D

Internal Photon Density From External Laser Power

The internal photon density of a laser cavity may be found just from knowing the external laser power. Assuming a steady state condition with time-harmonic electric field within and outside the laser cavity $E = \widetilde{E}e^{-j\omega t}$, where ω is the laser frequency, we may write the external laser intensity (using the time-averaged Poynting vector):

$$I^{external} = \frac{1}{2}\text{Re}(\widetilde{E} \times \widetilde{H}^*) = \frac{1}{2}n_{ext}c\varepsilon_o|\widetilde{E}_{external}|^2 \tag{D.1}$$

where in the second equality we assume $\widetilde{H} = \widetilde{E}/\eta$ where $\eta = 1/nc\varepsilon_o$ is the characteristic impedance of the medium external to the laser which has refractive index n_{ext}. Furthermore, we may write the time-averaged energy density within the laser cavity as,

$$u^{internal} = \frac{1}{2}n^2_{internal}\varepsilon_o|\widetilde{E}_{internal}|^2 \tag{D.2}$$

where $n_{internal}$ is the (effective) index of the laser cavity. Note that this term includes contributions from both the electric and magnetic field energy densities, where we have assumed that the energy density in the magnetic field is equal to the energy density in the electric field (on average). The factor of $1/2$ appears from time-averaging the harmonic fields. Now to get the average photon density in the cavity, we must average over the energy density within the laser volume. That is, we take,

$$S^{internal} = \frac{\text{Photons}}{\text{Volume}} = \frac{\iiint u(\vec{r})d^3\vec{r}}{\hbar\omega \cdot \text{Volume}} \tag{D.3}$$

where $\hbar\omega$ is the photon energy (allowing us to convert from energy density to photon density), and $\vec{r}$ is the spatial position within the laser cavity. After substituting Eq. D.2 into Eq. D.3, we may take the ratio of Eq. D.3 and Eq. D.1. We find,

$$\frac{S^{internal}}{I^{external}} = \frac{1}{v_g\hbar\omega}\frac{n_{internal}}{n_{external}}\frac{1}{\text{Volume}}\iiint \frac{|\widetilde{E}_{internal}(\vec{r})|^2}{|\widetilde{E}_{external}|^2}d^3\vec{r} \tag{D.4}$$

where we replaced $c/n_{internal}$ with the group velocity v_g, and $|\widetilde{E}_{external}|^2$ was moved into the integral and treated as a constant (since the outgoing wave is planewave-like and does not depend on location).

If the incident beam on the laser facet within the cavity were a simple plane wave, the ratio of the refractive indices internal and external to the laser facet and the ratio of the electric field intensities could be recognized as the power transmission at normal incidence, $T = 1 - R$. However, since the reflected light also contributes to the overall internal photon density, we cannot simply use T. Indeed, for the cleaved facet laser with a TE mode, one actually finds that that the two electric field intensities at the laser facet are equal because of the tangential electric field boundary condition. One must average over the electric field within the full volume of the laser cavity because of the resulting spatial interference pattern in the Fabry-Perot cavity from reflections at the waveguide facets.

This is depicted more clearly in Fig. D.1 where we simulate the electric field intensity within a Fabry-Perot cavity. Fig. D.1(a) shows the geometry of a laser used for this toy example, with a length $L = 5\mu$m and reflectivity $R = 30\%$ at each facet [1]. We assume that the coherent beam in the laser is launched from the center of the cavity. Fig D.1(b-e) depict the resulting electric field intensity as a function of position in the laser cavity at a wavelength of 1570nm with absorption coefficients, α, of 50,000, 0, -1500, and -2407 1/cm respectively. Negative absorption can be interpreted as gain. Note that Fig. D.1(b) was normalized by the peak electric field intensity, while Fig D.1(c)-(e) were normalized by the electric field intensity external to the laser (after transmission). Because of the tangential E-field boundary condition, the electric field intensity outside of the laser is equal to the electric field intensity at the boundary of the laser facet. In Fig. D.1(b) the large absorption prevents the launched wave from reaching the laser facet. In Fig D.1(c) one can observe coherent oscillations in the laser mode, where the oscillation amplitude is smaller on the left half of the cavity because most of the initial wave power was lost upon transmission through the mirror at the right facet. In Fig D.1(d) gain is turned on, resulting in increasing amplitude of oscillations on both ends. Finally, Fig D.1(e) shows the gain threshold condition, where the net gain in the cavity is equal to the mirror loss $\alpha_m = \frac{1}{2L} \ln \frac{1}{R_1 R_2} = 2407$ 1/cm. As one might intuit, the spatial profile of the electric field at the gain threshold condition is independent of the location of the initial incident wave and always takes on this symmetric profile. For larger mirror reflectivity and longer Fabry-Perot cavity length, the oscillation amplitude becomes nearly constant throughout the cavity.

Returning to Eq. D.4 and using the insight from Fig. D.1(e) at the laser threshold condition, we may now evaluate the integral term. Assuming that the mode profile of the outgoing laser beam is about equal to the internal laser mode profile in the transverse directions, the integral in Eq. D.4 evaluates approximately to:

$$\frac{1}{\text{Volume}} \iiint \frac{|\widetilde{E}_{internal}(\vec{r})|^2}{|\widetilde{E}_{external}|^2} d^3\vec{r} \approx \frac{\iint_A dxdy}{A} \cdot \frac{1}{L} \int_0^L \cos^2\left(\frac{2\pi n}{\lambda} z\right) dz \approx \frac{1}{2} \tag{D.5}$$

where the approximation was deduced from the $\cos^2$-like profile of Fig. D.1(e) [2].

[1] The short length of the cavity is used for illustration purposes only, because the resulting spatial profile of the laser mode would be too difficult to see for a more reasonable length (like that taken in the main text, $L = 60\mu$m).

[2] Technically the integral is smaller than 1/2, but we will assume the conservative case for simplicity and generality. Note also that we ignored the spectral constructive interference condition in this analysis, but it was inconsequential and is assumed to hold at the lasing wavelength.

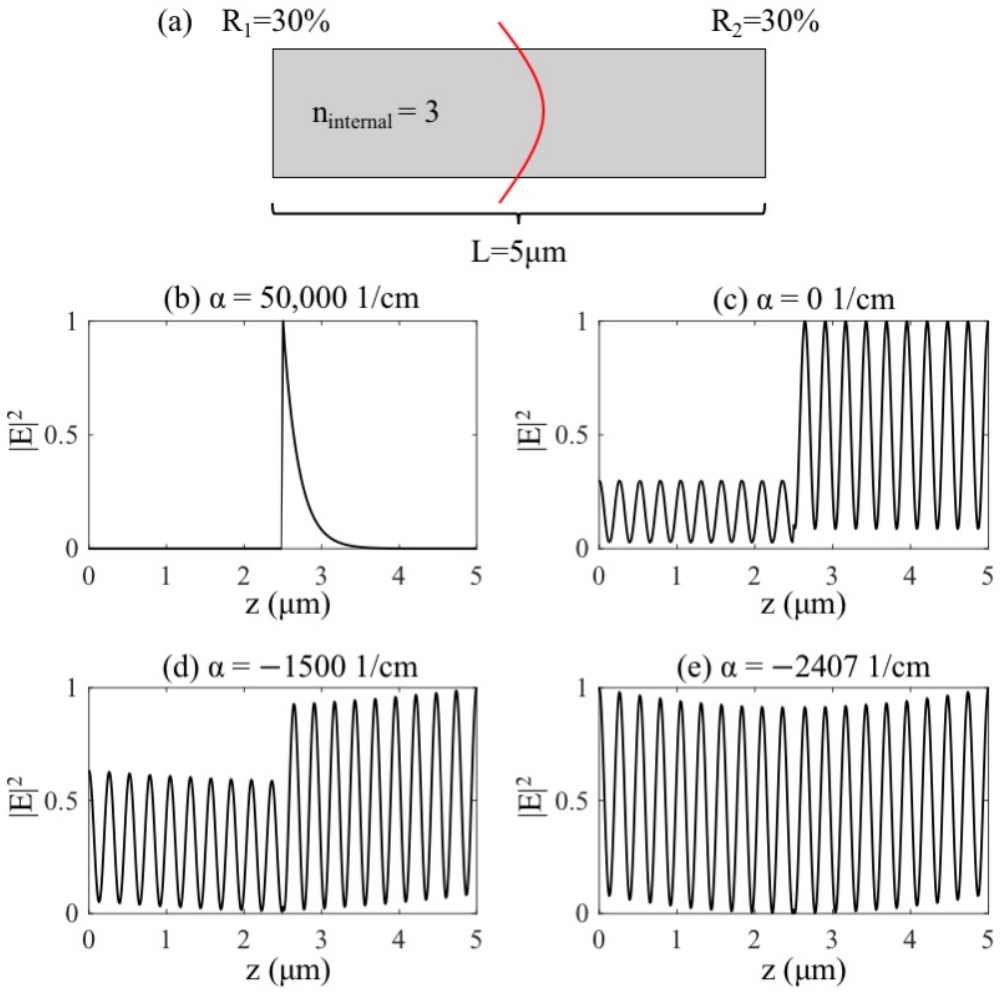

Figure D.1: (a) Simple one-dimensional Fabry-Perot optical cavity with 5μm length. Large absorption (b), zero absorption (c), some gain (d), and threshold gain (e) conditions respectively. In the case of threshold gain, the electric field intensity profile is independent of the incident source location.

Thus, we may solve Eq. D.4 for the average photon density internal to the Fabry-Perot cavity, given by:

$$S^{internal} \approx \frac{1}{2}\frac{1}{v_g \hbar\omega}\frac{n_{internal}}{n_{external}}\frac{P_{opt}}{A} \tag{D.6}$$

where we replaced $I^{external} = P_{opt}/A$ where P_{opt} is the laser power, and A is the modal area. Taking $n_{external} = 1$ for air and $n_{internal} = 3$ for the laser effective index, we find,

$$S^{internal} \approx \frac{3}{2}\frac{P_{opt}}{A \cdot v_g \cdot \hbar\omega} \tag{D.7}$$

which is the equation that was used in the main body of the text for calculation of the photon density from the laser power.

Appendix E

Dynamic Models of Antenna-LEDs and Lasers

In this Appendix we provide a detailed description of the time dynamics of antenna-LEDs and lasers. All carrier dynamics simulations performed in this thesis used the rate equations [22]:

$$\frac{\partial N}{\partial t} = G(I) - R_{\text{nr}}(N) - R_{\text{sp}}(N) - R_{\text{st}}(N,S) \tag{E.1}$$

$$\frac{\partial S}{\partial t} = \Gamma R_{\text{st}}(N,S) + \Gamma\beta R_{\text{sp}}(N) - \frac{S}{\tau_p} \tag{E.2}$$

where N is the density of minority carriers in the material (usually in units of electrons/cm^3), S is the density of photons in the optical cavity (in units of photons/cm^3), and each of the terms beginning with R correspond to recombination rates with units $1/(s \cdot cm^3)$. The recombination rates R_{nr} and R_{sp} correspond to nonradiative and radiative (spontaneous emission) respectively which depend only on the minority carrier density in the device. R_{st}, on the other hand, corresponds to recombination by stimulated emission and depends explicitly on both the minority carrier concentration and photon density terms. G is a generation term, which, for electrical injection, corresponds to the current pumped into the device per unit volume. Γ is known as the confinement factor, β is known as the spontaneous emission coupling factor, and τ_p is called the photon lifetime. A simplified version of Eq E.1 was provided in Chapter 3, and it describes the dynamics of the minority carrier population due to generation and recombination events. Eq. E.2 describes the dynamics of the photon population within an optical cavity, and is important for the description of lasers. Importantly, it indicates that the density of photons in the cavity increases with stimulate and spontaneous emission, and decreases as photons are lost at a rate $1/\tau_p$.

For simulations of large-signal modulation responses (such as Fig. 4.10 and Fig. 4.12), we used the detailed model of the spontaneous emission recombination rate (developed in Appendix B) along with the stimulated emission recombination rate and non-radiative recombination rate developed in Chapters 3 and 4. Eq. E.1 and Eq. E.2 were directly integrated using the backward Euler method for good accuracy.

E.1 Small-signal model of the antenna-LED

A small-signal modulation model for the antenna-LED is developed in this section. A small signal model of the laser modulation may be found in [22].

For the antenna-LED we can assume that $S \approx 0$ because generally-speaking, the photon lifetime, τ_p, is very small. This is because of the inherently small quality-factor, Q, of optical antennas due to both radiative and nonradiative damping in the presence of metal. Furthermore, because LEDs operate from the principle of spontaneous emission, a build-up of photons in an optical cavity is not needed (and in fact undesirable, because of the possibility of optical absorption) for high-speed operation. Thus, we need only consider Eq E.1, which describes the dynamics of the device minority carrier density. Moreover, we may assume the net stimulated emission rate $R_{\text{st}} = v_g g S \approx 0$ because of the small photon density.

Under direct electrical injection, we may take,

$$G(I) = \eta_i \frac{I}{qV} \tag{E.3}$$

where G is the generation rate due to electrical injection, I is the current, and V is the active volume. η_i is known as the injection efficiency, and describes what ratio of total current that ends up recombining in the active region of the device. We will now provide a step-by-step proof of the small-signal transfer function. Our goal is to find the peak-to-peak optical power, ΔP, induced by a small modulation in current, ΔI, at some frequency ω.

Small-signal assumption

We first assume that the carrier density and current satisfy,

$$n(t) = n_o + \Delta n e^{-j\omega t} \tag{E.4}$$

$$I(t) = I_o + \Delta I e^{-j\omega t} \tag{E.5}$$

where n_o corresponds to a steady-state carrier density due to I_o input current, and ΔI at frequency ω induces a small modulation in the carrier density Δn. We may then linearize the recombination terms by taking the first order Taylor expansion:

$$R_{\text{sp}} = R_{\text{sp,o}} + \frac{\partial R}{\partial n} \Delta n e^{-j\omega t} \tag{E.6}$$

$$R_{\text{nr}} \approx R_{\text{nr,o}} + \frac{\partial R}{\partial n} \Delta n e^{-j\omega t} \tag{E.7}$$

where $R_{\text{sp,o}}$ and $R_{\text{nr,o}}$ correspond to steady-state recombination rates, with linear modulation terms given by the differential terms $\partial R / \partial n$. Plugging Eq. E.7 into the rate equation Eq E.1 and rearranging, we find,

$$\frac{\Delta n}{\Delta I} = \frac{\eta_i}{qV} \frac{1}{j\omega + \frac{\partial R_{\text{sp}}}{\partial n} + \frac{R_{\text{nr}}}{\partial n}} \tag{E.8}$$

The optical power due to spontaneous emission is then given by,

$$P(t) = P_{\mathrm{o}} + \Delta P e^{-j\omega t} \tag{E.9}$$

where,

$$P_{\mathrm{o}} = \hbar\omega V R_{\mathrm{sp}} \tag{E.10}$$

$$\Delta P = \hbar\omega V \frac{\partial R_{\mathrm{sp}}}{\partial n}\Delta n \tag{E.11}$$

satisfy the steady-state power and differential power due to modulation respectively. We may then plug Eq. E.11 into Eq. E.8 to find,

$$\frac{\Delta P}{\Delta I} = \eta_i \frac{\hbar\omega}{q} \frac{\frac{\partial R_{\mathrm{sp}}}{\partial n}}{j\omega + \frac{\partial R_{\mathrm{sp}}}{\partial n} + \frac{\partial R_{\mathrm{nr}}}{\partial n}} \tag{E.12}$$

where we note that in the "DC" limit ($\omega \to 0$), we have:

$$\frac{\Delta P}{\Delta I}\Big|_{\omega=0} = \eta_i \frac{\hbar\omega}{q} \frac{\frac{\partial R_{\mathrm{sp}}}{\partial n}}{\frac{\partial R_{\mathrm{sp}}}{\partial n} + \frac{\partial R_{\mathrm{nr}}}{\partial n}} = \eta_i \eta_r \frac{\hbar\omega}{q} \tag{E.13}$$

Then, since in principle $\Delta P/\Delta I$ may be complex valued (accounting for a phase shift of the optical power with respect to the modulation current), we will take the modulus of Eq. E.12 and normalize by the DC limit to find:

$$\left|\frac{\Delta P}{\Delta I}\right| \Big/ \left|\frac{\Delta P}{\Delta I}\right|_{\omega=0} = \frac{1}{1 + \left(\dfrac{\omega}{\dfrac{\partial R_{\mathrm{sp}}}{\partial n} + \dfrac{\partial R_{\mathrm{nr}}}{\partial n}}\right)^2} \tag{E.14}$$

To solve for the $f_{3\mathrm{dB}}$ frequency, we then find the frequency such that the transfer function is equal to 1/2. Doing so, we find,

$$f_{3\mathrm{dB}} = \frac{\sqrt{3}}{2\pi}\left(\frac{\partial R_{\mathrm{sp}}}{\partial n} + \frac{\partial R_{\mathrm{nr}}}{\partial n}\right) \tag{E.15}$$

As shown in the main text, $\partial R_{\mathrm{sp}}/\partial n$, is defined as the carrier lifetime of spontaneous emission, $1/\tau_{\mathrm{sp}}$. Similarly, $\partial R_{\mathrm{nr}}/\partial n = 1/\tau_{\mathrm{nr}}$ can be thought of as a carrier lifetime due to nonradiative processes. Thus, Eq. E.15 becomes,

$$f_{3\mathrm{dB}} = \frac{\sqrt{3}}{2\pi}\left(\frac{1}{\tau_{\mathrm{sp}}} + \frac{1}{\tau_{\mathrm{nr}}}\right) \tag{E.16}$$

which concludes the proof. Note that one may obtain the $f_{3\mathrm{dB}}$ of an antenna-LED by replacing τ_{sp} with τ_{sp}^*, the enhanced spontaneous emission carrier lifetime.

Appendix F

Detailed Antenna Circuit Model

In Chapter 4 we demonstrated that a simplified circuit analysis of the dipole antenna suggests a peak enhancement factor of:

$$F = \left(\frac{l}{d}\right)^2 \tag{F.1}$$

where l is the antenna length and d is the antenna gap. However, in the course of this derivation we made a critical assumption: namely, we took the antenna to be a series RLC circuit with a current source. In this Appendix we will derive expressions for the dipole antenna enhancement and efficiency using the complete circuit model, depicted in Fig. F.1(a), including all antenna reactive effects. This complete circuit model was used to calculate the enhancement and efficiency curves from Chapter 5.2.

Consider the optical dipole antenna circuit model in Fig. F.1(a). The antenna length is l and there is a vacuum gap of width d. In this case, we will also need to explicitly define the antenna radius, where we choose $r = 25$nm. As depicted in the diagram, we include several additional lumped circuit elements. The antenna resistance includes several terms, including the radiation resistance R_{rad} and the antenna Ohmic resistance R_{loss}. Near the antenna gap is an additional source of resistance called the spreading resistance, R_{spread}, which was described originally in Chapter 5.2 to explain the decrease in antenna efficiency at small gap d. The reactive antenna elements include the intrinsic antenna capacitance C_{fringe}, a shunt capacitance in the antenna gap C_{gap}, the intrinsic Faraday inductance L_{F}, and an additional inductance called the kinetic inductance L_k. The kinetic inductance accounts for the lag of free electron motion in the antenna arms at large optical frequencies – in other words it accounts for the "plasmonic" nature of the antenna. All of these quantities are defined explicitly in [30]. For now, we will leave most of the circuit elements symbolic.

Fig. F.1(b) provides an equivalent circuit to the dipole antenna circuit model from Fig. F.1(a). Namely, the optical antenna is a current divider consisting of two impedance elements in parallel with the current

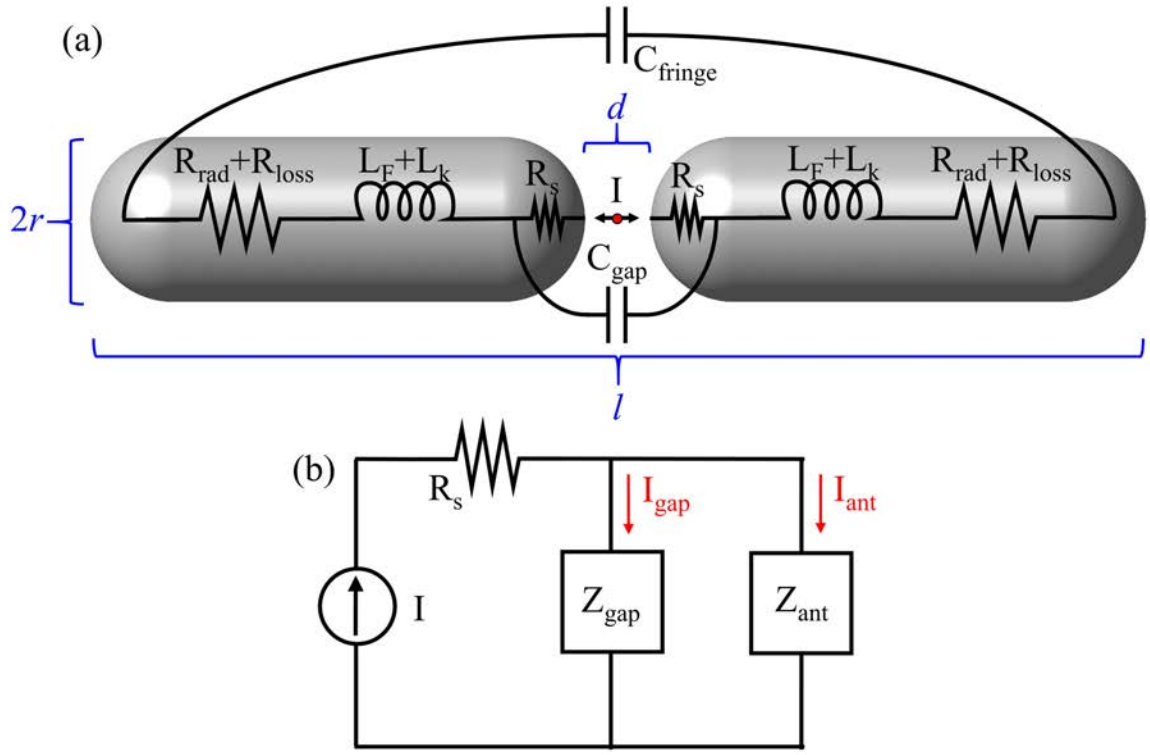

Figure F.1: (a) Circuit model of the dipole antenna, originally proposed by Eggleston et al [30]. (b) Equivalent optical antenna circuit is a current divider.

source, along with a series spreading resistance. The impedances are defined as,

$$Z_{\text{ant}} = R + j\omega L + \frac{1}{j\omega C_{\text{fringe}}} \tag{F.2}$$

$$Z_{\text{gap}} = \frac{1}{j\omega C_{\text{gap}}} \tag{F.3}$$

where $R = R_{\text{rad}} + R_{\text{loss}}$ and $L = L_{\text{k}} + L_{\text{F}}$. We are ultimately interested in the quantity:

$$P^{\text{tot}} = \frac{1}{2}|I_{\text{ant}}|^2 R + \frac{1}{2}|I|^2 R_{\text{s}} \tag{F.4}$$

because this will represent the total power drawn from the source, which can be used to obtain the enhancement factor and efficiency. Since the circuit is a current divider, we may easily write:

$$I_{\text{ant}} = \left(\frac{Z_{\text{gap}}}{Z_{\text{gap}} + Z_{\text{ant}}}\right) I \tag{F.5}$$

Then, we have,

$$P^{\text{tot}} = \frac{1}{2}|I|^2 R_{\text{s}} + \frac{1}{2}|I_{\text{ant}}|^2 R = \frac{1}{2}|I|^2 \left(R_{\text{s}} + \left| \frac{Z_{\text{gap}}}{Z_{\text{gap}} + Z_{\text{ant}}} \right|^2 R \right) \tag{F.6}$$

It remains to calculate the impedance ratio term. Note that we may write,

$$Z_{\text{gap}} + Z_{\text{ant}} = R + j\omega L + \frac{1}{j\omega} \left(\frac{1}{C_{\text{fringe}}} + \frac{1}{C_{\text{gap}}} \right) \tag{F.7}$$

We may then define an effective capacitance:

$$\frac{1}{C_{\text{eff}}} = \frac{1}{C_{\text{fringe}}} + \frac{1}{C_{\text{gap}}} \tag{F.8}$$

Then there is a resonance condition for the sum of the impedances in Eq. F.7 that occurs when

$$\omega_1^2 = \frac{1}{LC_{\text{eff}}} \tag{F.9}$$

On this condition:

$$|Z_{\text{gap}} + Z_{\text{ant}}|^2 = R^2 \tag{F.10}$$

$$|Z_{\text{gap}}|^2 = \frac{LC_{\text{eff}}}{C_{\text{gap}}^2} \tag{F.11}$$

This allows us to rewrite the total power from Eq. F.6 as (on resonance):

$$P_1^{\text{tot}} = \frac{1}{2}|I|^2 \left(R_{\text{s}} + \frac{LC_{\text{eff}}}{R^2 C_{\text{gap}}^2} R \right) \tag{F.12}$$

Interestingly, the second term depends on a ratio between reactive and resistive time constants. In Chapter 5.2, we expressed the enhancement factor as a ratio between the antenna radiated power and the radiated power from a Hertzian dipole. Note that the overall enhancement seen by a dipole source includes the resistive losses, and this was taken into account in our calculation of the average enhancement of the cavity backed slot antenna in Chapter 4. Nevertheless, more physical significance can be derived from considering only the radiated power. Using Eq. F.12 we may easily obtain the radiated power on resonance by disregarding R_{s}:

$$P_1^{\text{rad}} = \frac{1}{2}|I|^2 \frac{LC_{\text{eff}}}{R^2 C_{\text{gap}}^2} R_{\text{rad}} \tag{F.13}$$

The radiation resistance and current of the dipole antenna were provided in Chapter 4 and are repeated here:

$$I = 2q\omega \frac{|x_{21}|}{d} \tag{F.14}$$

$$R_{\text{rad}} = \frac{2}{3}\pi Z_{\text{o}} \left(\frac{l}{d} \right)^2 \tag{F.15}$$

Furthermore, the radiated power by a Hertzian dipole is given by:

$$P_{\text{o}} = \frac{|qx_{21}|^2\omega^4}{3\pi\varepsilon_{\text{o}}c^3} \tag{F.16}$$

After some rearrangement, we may express the (radiative) enhancement factor as:

$$F = \frac{P_1^{\text{rad}}}{P_{\text{o}}} = \frac{LC_{\text{eff}}}{R^2C_{\text{gap}}^2}\left(\frac{l}{d}\right)^2 \tag{F.17}$$

To which we recover the characteristic $(l/d)^2$ factor. However, There is an additional prefactor, which has a complicated dependence on the antenna reactive and resistance parameters. Each of these parameters, in turn, have dependencies on the antenna geometry and material. For brevity, we will not express these circuit elements in terms of the antenna parameters, but they may be found in [30].

It is interesting to note that there is an L/R factor, which can be thought of as an antenna quality factor, establishing some analog between this circuit model and the Purcell enhancement factor. However, in practice one generally does not want to decrease R because low quality factors are typically desirable in antennas. Furthermore, decreasing the total resistance is not trivial, because radiation resistance tends to trade off with Ohmic loss unfavorably. Another interesting feature is that the shunt capacitance C_{gap} can severely limit the antenna enhancement. It is desirable to lower this capacitance as much as possible without sacrificing the small gap parameter d. The best way to do this is to use sharp metallic or dielectric tips, as demonstrated in Chapter 5.2.

The full enhancement factor in Eq. F.17 was used in Chapter 5.2 in Fig. 5.9(c). Furthermore, we presented a circuit model of the antenna efficiency with and without the surface collision effect, in Fig. 5.9(d). This circuit model antenna efficiency is given by the ratio of the radiated power on resonance with the total power on resonance:

$$\text{Efficiency} = \frac{P_1^{\text{rad}}}{P_1^{\text{tot}}} = \frac{\frac{1}{2}|I|^2\frac{LC_{\text{eff}}}{R^2C_{\text{gap}}^2}R_{\text{rad}}}{\frac{1}{2}|I|^2\left(R_{\text{s}} + \frac{LC_{\text{eff}}}{R^2C_{\text{gap}}^2}R\right)} \tag{F.18}$$

Approximately speaking, for a well-designed antenna the prefactor becomes $\frac{LC_{\text{eff}}}{R^2C_{\text{gap}}^2} \approx 1$. Thus, the efficiency may be expressed as,

$$\text{Efficiency} \approx \frac{R_{\text{rad}}}{R_{\text{s}} + R_{\text{loss}} + R_{\text{rad}}} \tag{F.19}$$

Now, we note that the spreading resistance term is given by $R_{\text{s}} = 2\rho/d$ where ρ is the antenna resistivity, and d is the antenna gap. Furthermore, when including surface collisions, we modify the resistivity in the concentrated current region by a factor $(1 + l_\infty/\beta d)$ where l_∞ is the bulk electron mean free path, and βd is an empirical parameter that can be interpreted as the net mean free path in the confined geometry of the antenna defined by the vacuum gap parameter d. Thus, the efficiency may be approximately written:

$$\text{Efficiency} \approx \frac{R_{\text{rad}}}{2\frac{\rho}{d}\left(1 + \frac{l_\infty}{\beta d}\right) + R_{\text{loss}} + R_{\text{rad}}} \tag{F.20}$$

Hence, this stringent dependence on d emphasizes the need to optimize the gap distance to increase the antenna enhancement without sacrificing the antenna efficiency due to spreading resistance and surface collisions. This concludes our discussion of the optical antenna circuit model.

Appendix G

Average Enhancement Factor

In this section we will derive the average enhancement factor from the text in detail. As a reminder, we used,

$$\frac{1}{\tau_{\text{sp}}^*} = \frac{F_{\text{average}}}{\tau_{\text{sp}}} \tag{G.1}$$

$$F_{\text{average}} = F_{\text{peak}} \cdot \text{Spatial Average} \cdot \text{Spectral Average} \cdot \text{Polarization Average} \tag{G.2}$$

Thus, in particular we will show how each of the averages may be found.

Before we begin, we note that the recombination rate due to spontaneous emission in an intrinsic semiconductor is given by (from Appendix B),

$$R_{\text{sp}} = \int \frac{1}{\tau(\omega)} \rho_r(\omega) f_e(\omega) d\omega \tag{G.3}$$

where ρ_r is the reduced density of states and f_e is called the Fermi emission factor. These quantities account for the carrier concentration in the LED. τ is the spontaneous emission lifetime of a two level system, such that,

$$\frac{1}{\tau} = \frac{|H'_{21}|^2 \omega^3 n}{3\pi\varepsilon_o c^3} \tag{G.4}$$

where $|H'_{21}|^2 = |qx_{21}|^2$ is the matrix element. Note that in Eq. G.4 we have already implicitly performed polarization averaging over the material matrix element and zero-point electric field polarization. In other words, in the main text we showed that the matrix element depends on the incident electric field polarization in confined geometries, often called transition matrix element:

$$|q\vec{x} \cdot \hat{e}|^2 = K_i |qx_{21}|^2 = K_i |H'_{21}|^2 \tag{G.5}$$

where K_i accounts for the averaged dipole polarization seen by the incident light with polarization $\hat{e}$. However, for spontaneous emission from regular semiconductors, we must also average over the polarization of the zero-point electric field which we assume is equally partitioned in the three Cartesian

directions. Thus, the matrix element is once again averaged:

$$\frac{1}{3}\sum_{i\in[\hat{x},\hat{y},\hat{z}]} K_i|H'_{21}|^2 = \frac{1}{3}|H'_{21}|^2 \sum_{i\in[\hat{x},\hat{y},\hat{z}]} K_i = \frac{1}{3}|H'_{21}|^2 \tag{G.6}$$

where we used that $\sum_i K_i = 1$ regardless of the semiconductor crystal confinement. This derivation indicates that we always retrieve Eq. G.4 for the spontaneous emission lifetime of conventional LEDs[1]. We will return to this point about polarization later on, but for now we must keep the relative transition strengths, K_i, in mind.

Moving on to enhanced spontaneous emission, we first note that the enhancement factor (normalized power) from a point dipole is in general a function of frequency, position, and polarization with respect to an antenna or electromagnetic cavity. This can be easily confirmed in simulation. In other words, we may write

$$F \rightarrow F_i(\omega, \vec{r}) \tag{G.7}$$

where ω is the frequency, $\vec{x}$ is the 3D position, and $i \in [\hat{x}, \hat{y}, \hat{z}]$ is the polarization of a point dipole source. Therefore, the recombination rate of an idealized point source before taking into account spatial averaging is given by,

$$R^*_{\text{sp}}(\vec{r}) = \int \left(\sum_i K_i F_i(\omega, \vec{r})\right) \frac{1}{\tau(\omega)} \rho_r(\omega) f_e(\omega) d\omega \tag{G.8}$$

where have simply multiplied the integrand by the sum of the enhancement factors in each polarization, which also have their own spatial and spectral dependencies in general. Notice that if $F_i = 1$ for each polarization (which is the conventional LED), then we recover the regular recombination rate since $\sum_i K_i = 1$. In general, point dipole sources will be distributed spatially throughout the semiconductor, so we may simply perform a weighted average under the assumption that the carrier concentration does not considerably vary spatially:

$$R^*_{\text{sp}} = \frac{1}{V} \iint \left(\sum_i K_i F_i(\omega, \vec{r})\right) \frac{1}{\tau(\omega)} \rho_r(\omega) f_e(\omega) d\omega d^3\vec{r} \tag{G.9}$$

where V is the active volume of the semiconductor.

We now make the assumption that the enhancement factor F is a separable function in frequency and space. This is not necessarily true for arbitrary structures, but is approximately true for the optical

[1]Interestingly, this implies that the B_{o} radiative recombination coefficient is independent of quantum confinement. The author was unable to find references confirming this result, but typically B_{o} is not reported as contingent upon whether the device emits from quantum wells or bulk.

antennas studied in this work (confirmed by exhaustive simulation). Thus, for polarization $i \in [\vec{x}, \vec{y}, \vec{z}]$:

$$F_i(\omega, \vec{r}) = F_i^{peak} X_i(\vec{r}) W_i(\omega) \tag{G.10}$$

$$W_i(\omega) \equiv \frac{F_i(\omega, \vec{r} = \vec{r}_{peak})}{F_i^{peak}} \tag{G.11}$$

$$X_i(\vec{r}) \equiv \frac{F_i(\omega = \omega_{peak}, \vec{r})}{F_i^{peak}} \tag{G.12}$$

where $F_i^{peak} = F_i(\omega = \omega_{peak}, \vec{r} = \vec{r}_{peak})$ is the peak enhancement spatially and spectrally, and the X_i and W_i are normalized functions that take into account the separate spectral and spatial dependencies with respect to peak. Thus, substituting these expressions into Eq. G.9, we have,

$$R_{\text{sp}}^* = \frac{1}{V} \sum_i K_i F_i^{peak} \iint d\omega d^3\vec{r} X_i(\vec{r}) W_i(\omega) \frac{1}{\tau(\omega)} \rho_r(\omega) f_e(\omega) \tag{G.13}$$

where we pulled the polarization sum and F_i^{peak} terms out of the integral since they are constants that only depend on polarization. Now, we may separate the integral into its spatial and spectral parts, such that:

$$R_{\text{sp}}^* = \sum_i K_i F_i^{peak} \frac{1}{V} \int d^3\vec{r} X_i(\vec{r}) \int d\omega W_i(\omega) \frac{1}{\tau(\omega)} \rho_r(\omega) f_e(\omega) \tag{G.14}$$

Now, we observe that the first integral can be regarded as spatial average (indexed by polarization) of the enhancement factor,

$$\text{Spatial Average}_i = \frac{1}{V} \int d^3\vec{r} X_i(\vec{r}) = \frac{1}{V} \int d^3\vec{r} \frac{F_i(\omega = \omega_{peak}, \vec{r})}{F_i^{peak}} \tag{G.15}$$

Noting that $F \propto |E|^2$, we can write [2]:

$$\text{Spatial Average}_i = \frac{1}{V} \int d^3\vec{r} X_i(\vec{r}) = \frac{1}{V} \int d^3\vec{r} \frac{|E_i(\vec{r})|^2}{|E_i|^2_{peak}} \tag{G.16}$$

$$\text{Spatial Average}_i = \frac{1}{V} \iiint |E_i|^2_{\text{normalized}} dxdydz \tag{G.17}$$

where the bottom expression is the Spatial Average definition used in the main text. Thus, Eq. G.14 becomes,

$$R_{\text{sp}}^* = \sum_i F_i^{peak} K_i \times (\text{Spatial Average}_i) \times \int d\omega W_i(\omega) \frac{1}{\tau(\omega)} \rho_r(\omega) f_e(\omega) \tag{G.18}$$

[2] We may ignore changes in optical density of states because we are only considering spatial variations.

Now, to obtain the total average enhancement, we will first need to divide Eq. G.18 by the recombination rate of a regular LED (from Eq. G.3):

$$\frac{R^*_{\text{sp}}}{R_{\text{sp}}} = \sum_i F_i^{peak} K_i \times (\text{Spatial Average}_i) \times \frac{\int d\omega W_i(\omega) \frac{1}{\tau(\omega)} \rho_r(\omega) f_e(\omega)}{\int d\omega \frac{1}{\tau(\omega)} \rho_r(\omega) f_e(\omega)} \tag{G.19}$$

We may rewrite the right hand term by expanding $W(\omega)$ as the following:

$$\frac{\int d\omega W_i(\omega) \frac{1}{\tau(\omega)} \rho_r(\omega) f_e(\omega)}{\int d\omega \frac{1}{\tau(\omega)} \rho_r(\omega) f_e(\omega)} = \frac{1}{F_i^{peak}} \frac{\int d\omega F_i(\omega, \vec{r} = \vec{r}_{peak}) \frac{1}{\tau(\omega)} \rho_r(\omega) f_e(\omega)}{\int d\omega \frac{1}{\tau(\omega)} \rho_r(\omega) f_e(\omega)} \tag{G.20}$$

Letting $L_N(\omega) \equiv \frac{1}{\tau(\omega)} \rho_r(\omega) f_e(\omega)$ define the intrinsic spontaneous emission spectrum corresponding to minority carrier concentration N, this can be recognized as the spectral average from the main manuscript:

$$\text{Spectral Average}_i = \frac{1}{F_i^{peak}} \frac{\int d\omega F_i(\omega, \vec{r} = \vec{r}_{peak}) L_N(\omega)}{\int d\omega L_N(\omega)} \tag{G.21}$$

Thus, the recombination rate ratio in Eq. G.19 becomes:

$$\frac{R^*_{\text{sp}}}{R_{\text{sp}}} = \sum_i F_i^{peak} K_i \times (\text{Spatial Average}_i) \times \big(\text{Spectral Average}_i\big) \tag{G.22}$$

Finally, let,

$$F^{peak} = \max\{F_x^{peak}, F_y^{peak}, F_z^{peak}\} \tag{G.23}$$

in other words, the maximum enhancement over all polarizations. Then, we may write:

$$\frac{R^*_{\text{sp}}}{R_{\text{sp}}} = F^{peak} \sum_i \frac{F_i^{peak} K_i}{F^{peak}} \times (\text{Spatial Average}_i) \times \big(\text{Spectral Average}_i\big) \tag{G.24}$$

Observe that we may transform the leftmost term in the following way:

$$\sum_i \frac{F_i^{peak} K_i}{F^{peak}} = \frac{1}{F^{peak}} \frac{\sum_i F_i^{peak} K_i |H'_{21}|^2}{|H'_{21}|^2} = \frac{1}{F^{peak}} \frac{\sum_i F_i^{peak} K_i |H'_{21}|^2}{\sum_i K_i |H'_{21}|^2} \tag{G.25}$$

where we used the fact that $\sum_i R_i = 1$ and $|H'_{21}|^2$ is the matrix element. We called the right hand side of this expression the polarization average in the main text, but neglected the contribution of the spatial average and spectral average in the overall expression. Typically speaking, the spectral average and spatial averages do not have strong polarization dependencies, so we may drop it for simplicity. Moreover, one

enhancement polarization dominates over the others, e.g. $F_y \gg \{F_x, F_z\}$. Consequently, we may define the Polarization Average as,

$$\text{Polarization Average} \approx \frac{1}{F^{peak}} \frac{\sum_i F_i^{peak} K_i |H'_{21}|^2}{\sum_i K_i |H'_{21}|^2} \approx \frac{1}{F^{peak}} \frac{F^{peak} K_y |H'_{21}|^2}{\sum_i K_i |H'_{21}|^2} \tag{G.26}$$

$$\text{Polarization Average} = \frac{K_y}{K_x + K_y + K_z} = K_y \tag{G.27}$$

where the large enhancement in y has picked out the transition strength in that direction. Thus, the recombination rate ratio becomes:

$$\frac{R^*_{\text{sp}}}{R_{\text{sp}}} = F^{peak} \times (\text{Polarization Average}) \times (\text{Spatial Average}) \times (\text{Spectral Average}) \tag{G.28}$$

where we have dropped the subscripts on the Spectral and Spatial Average terms, where they are implicitly defined in the y direction.

To obtain the carrier lifetime, we note that we had implicitly assumed in the spectral average term that both the antenna-LED and reference LED are pumped to the same carrier concentration N. We may then represent the respective recombination rates as $R = N/\tau$, allowing the carrier concentration to cancel. This allows us to define:

$$F_{\text{average}} \equiv \frac{R^*_{\text{sp}}}{R_{\text{sp}}} = \frac{1/\tau^*_{\text{sp}}}{1/\tau_{\text{sp}}} \tag{G.29}$$

And by comparison with Eq. G.28, we have,

$$F_{\text{average}} \equiv F^{peak} \times (\text{Polarization Average}) \times (\text{Spatial Average}) \times (\text{Spectral Average}) \tag{G.30}$$

which concludes the proof.

A curiosity concerning the polarization average

From the beginning of the average enhancement proof, we implicitly made an assumption regarding polarization: namely that the zero-point electric field is equally partitioned in the three Cartesian directions. Indeed, dropping the spectral and spatial terms from Eq. G.8, we have

$$R^*_{\text{sp}} \propto \sum_i K_i F_i \frac{1}{\tau} \tag{G.31}$$

where τ is the fundamental spontaneous emission lifetime of a two-level system. Noting that $\tau \propto \frac{1}{3}|H'_{21}|^2$ according to our polarization average from before, we can rewrite Eq. G.31 as,

$$R^*_{\text{sp}} \propto \sum_i \frac{F_i}{3} K_i |H'_{21}|^2 \tag{G.32}$$

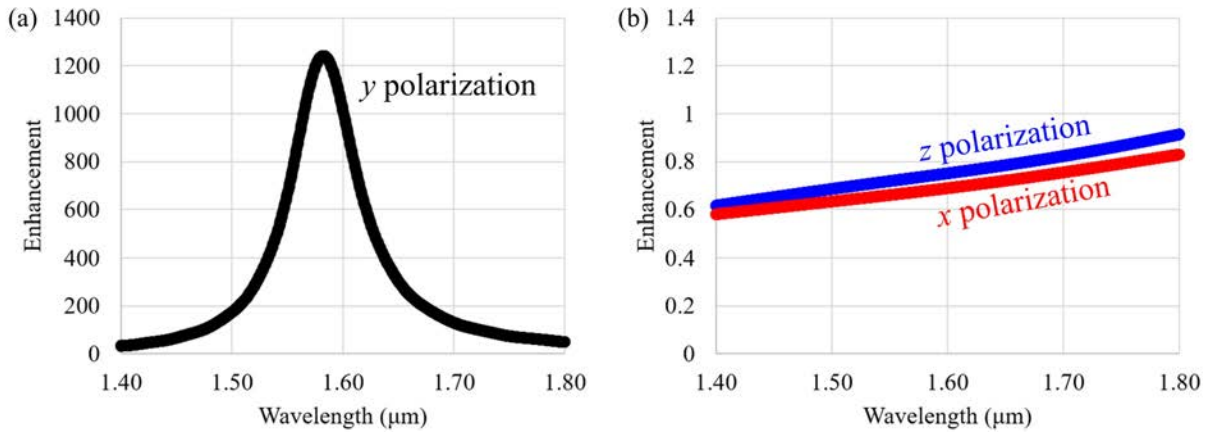

Figure G.1: Off-polarization enhancement spectrum indicates suppression of spontaneous emission, in contrast with the on-polarization enhancement.

In other words, we are taking a simple average over the three enhancement polarizations. Does this mean that we have implicitly assumed that the zero-point E-field energy is equal in all three directions even when there is enhancement present?

The curiosity that remains is whether we should take into account preferential polarization of the vacuum field in this average. In the extreme case of a 3D photonic crystal cavity there is a photonic bandgap, implying that the vaccuum field within the cavity is identically zero. Moreover, 2D photonic crystals exhibit polarized spontaneous emission. Some evidence that a similar effect is occurring in the antenna-LED can be seen in Fig. G.1. Here we show the enhancement spectra of a dipole oriented in x, y, and z within the cavity-backed slot antenna. As discussed previously, the y polarization exhibits very large enhancement factor on resonance. Meanwhile, the off-polarizations are not only small but actually dip below an enhancement factor of 1. This indicates that the dipole emission of these polarizations is actually suppressed, analagous to the inhibited spontaneous emission in photonic crystals.

Should this fact be taken into account in Eq. G.32? An alternative to a simple average would be to use a weighted average of each enhancement factor by $F_i/\sum_j F_j$. Then we would have,

$$R^*_{\text{sp}} \propto \sum_i \left(\frac{F_i}{\sum_j F_j} \right) F_i K_i |H'_{21}|^2 \approx F_y K_y |H'_{21}|^2 \tag{G.33}$$

In other words, the factor of 3 would drop out and the antenna-LED would receive the full benefit of the transition matrix element. This is only a guess, however, and we will leave this as an unsolved experimental question.

Bibliography

[1] S. Adachi. "Optical dispersion relations for GaP, GaAs, GaSb, InP, InAs, InSb, AlxGa1xAs, and In1xGaxAsyP1y". In: *Journal of Applied Physics* 66.12 (Dec. 1989), pp. 6030–6040. DOI: 10.1063/1.343580.

[2] N. M. Andrade et al. "Efficient and broadband single-mode waveguide coupling of electrically injected optical antenna based nanoLED". In: *2017 Fifth Berkeley Symposium on Energy Efficient Electronic Systems Steep Transistors Workshop (E3S)*. Oct. 2017, pp. 1–3. DOI: 10.1109/E3S.2017.8246159.

[3] N. M. Andrade et al. "Efficient single-mode waveguide coupling of electrically injected optical antenna based nanoLED". In: *2017 IEEE Photonics Conference (IPC)*. Oct. 2017, pp. 649–650. DOI: 10.1109/IPCon.2017.8116265.

[4] N. M. Andrade et al. "Inverse design optimization for efficient coupling of an electrically injected optical antenna-LED to a single-mode waveguide". In: *Optics Express* 27.14 (July 2019), pp. 19802–19814. DOI: 10.1364/OE.27.019802.

[5] N. M. Andrade et al. "Optical Antenna NanoLED Based Interconnect Design". In: *2018 IEEE Photonics Conference (IPC)*. Sept. 2018, pp. 1–2. DOI: 10.1109/IPCon.2018.8527169.

[6] A. S. G. Andrae and T. Edler. "On Global Electricity Usage of Communication Technology: Trends to 2030". In: *Challenges* 6.1 (June 2015), pp. 117–157. DOI: 10.3390/challe6010117.

[7] P. Anger, P. Bharadwaj, and L. Novotny. "Enhancement and Quenching of Single-Molecule Fluorescence". In: *Physical Review Letters* 96.11 (Mar. 2006), p. 113002. DOI: 10.1103/PhysRevLett.96.113002.

[8] G. Astfalk. "Why optical data communications and why now?" In: *Applied Physics A* 95.4 (June 2009), pp. 933–940. DOI: 10.1007/s00339-009-5115-4.

[9] S.-H. Bae et al. "Integration of bulk materials with two-dimensional materials for physical coupling and applications". In: *Nature Materials* 18.6 (June 2019), pp. 550–560. DOI: 10.1038/s41563-019-0335-2.

[10] C. Baudot et al. "Low cost 300mm double-SOI substrate for low insertion loss 1D 2D grating couplers". In: *11th International Conference on Group IV Photonics (GFP)*. Aug. 2014, pp. 137–138. DOI: 10.1109/Group4.2014.6961964.

[11] R. G. Beausoleil, M. Mclaren, and N. P. Jouppi. "Photonic Architectures for Data Centers". In: *IEEE Journal of Selected Topics in Quantum Electronics* 19.2 (2013), p. 3700109. DOI: 10.1109/JSTQE.2012.2236080.

[12] H. Becker et al. "Out-of-Plane Focusing Grating Couplers for Silicon Photonics Integration With Optical MRAM Technology". In: *IEEE Journal of Selected Topics in Quantum Electronics* 26.2 (Mar. 2020), pp. 1–8. DOI: 10.1109/JSTQE.2019.2933805.

[13] D. Benedikovic et al. "High-directionality fiber-chip grating coupler with interleaved trenches and subwavelength index-matching structure". In: *Optics Letters* 40.18 (Sept. 2015), pp. 4190–4193. DOI: 10.1364/OL.40.004190.

[14] S. Bhargava and E. Yablonovitch. "Lowering HAMR near-field transducer temperature via inverse electromagnetic design". In: *IEEE Transactions on Magnetics* 51.4 (Apr. 2015), pp. 1–7. DOI: 10.1109/TMAG.2014.2355215.

[15] A. del Campo and C. Greiner. "SU-8: a photoresist for high-aspect-ratio and 3D submicron lithography". In: *Journal of Micromechanics and Microengineering* 17.6 (May 2007), R81–R95. DOI: 10.1088/0960-1317/17/6/R01.

[16] E. Castanié, M. Boffety, and R. Carminati. "Fluorescence quenching by a metal nanoparticle in the extreme near-field regime". In: *Optics Letters* 35.3 (Feb. 2010), pp. 291–293. DOI: 10.1364/OL.35.000291.

[17] R. R. Chance, A. Prock, and R. Silbey. "Molecular Fluorescence and Energy Transfer Near Interfaces". In: *Advances in Chemical Physics*. Vol. 37. John Wiley & Sons, 1978, pp. 1–65.

[18] X. Chen et al. "Dual-etch apodised grating couplers for efficient fibre-chip coupling near 1310 nm wavelength". In: *Optics Express* 25.15 (July 2017), pp. 17864–17871. DOI: 10.1364/OE.25.017864.

[19] H. Choi, M. Heuck, and D. Englund. "Self-Similar Nanocavity Design with Ultrasmall Mode Volume for Single-Photon Nonlinearities". In: *Physical Review Letters* 118.22 (May 2017), p. 223605. DOI: 10.1103/PhysRevLett.118.223605.

[20] L. J. Chu. "Physical Limitations of Omni-Directional Antennas". In: *Journal of Applied Physics* 19.12 (Dec. 1948), pp. 1163–1175. DOI: 10.1063/1.1715038.

[21] S. L. Chuang. *Physics of Photonic Devices*. 2nd ed. Wiley, Jan. 2009. ISBN: 978-0-470-29319-5.

[22] L. A. Coldren, S. W. Corzine, and M. L. Mashanovitch. *Diode Lasers and Photonic Integrated Circuits*. 2nd ed. Wiley, Mar. 2012. ISBN: 978-0-470-48412-8.

[23] G. Crosnier et al. "Hybrid indium phosphide-on-silicon nanolaser diode". In: *Nature Photonics* 11.5 (May 2017), pp. 297–300. DOI: 10.1038/nphoton.2017.56.

[24] A. Demeter and S. Ruschin. "S-matrix absolute optimization method for a perfect vertical waveguide grating coupler". In: *arXiv:1903.11313 [physics]* (Mar. 2019).

[25] K. Ding and C. Z. Ning. "Fabrication challenges of electrical injection metallic cavity semiconductor nanolasers". In: *Semicond. Sci. Technol.* 28.12 (Nov. 2013), p. 124002. DOI: 10.1088/0268-1242/28/12/124002.

[26] H. M. Doeleman, E. Verhagen, and A. F. Koenderink. "Antenna–Cavity Hybrids: Matching Polar Opposites for Purcell Enhancements at Any Linewidth". In: *ACS Photonics* 3.10 (Oct. 2016), pp. 1943–1951. DOI: 10.1021/acsphotonics.6b00453.

[27] V. Dolores-Calzadilla et al. "Waveguide-coupled nanopillar metal-cavity light-emitting diodes on silicon". In: *Nature Communications* 8.1 (Feb. 2017), p. 14323. DOI: 10.1038/ncomms14323.

[28] R. H. Doremus. "Optical Properties of Small Silver Particles". In: *The Journal of Chemical Physics* 42.1 (Jan. 1965), pp. 414–417. DOI: 10.1063/1.1695709.

[29] M. S. Eggleston and M. C. Wu. "Efficient Coupling of an Antenna-Enhanced nanoLED into an Integrated InP Waveguide". In: *Nano Letters* 15.5 (May 2015), pp. 3329–3333. DOI: 10.1021/acs.nanolett.5b00574.

[30] M. S. Eggleston et al. "Optical antenna enhanced spontaneous emission". In: *Proceedings of the National Academy of Sciences* 112.6 (Feb. 2015), pp. 1704–1709. DOI: 10.1073/pnas.1423294112.

[31] M. S. Eggleston et al. "Ultrafast Spontaneous Emission from a Slot-Antenna Coupled WSe2 Monolayer". In: *ACS Photonics* 5.7 (July 2018), pp. 2701–2705. DOI: 10.1021/acsphotonics.8b00381.

[32] Y. Elesin et al. "Time domain topology optimization of 3D nanophotonic devices". In: *Photonics and Nanostructures - Fundamentals and Applications* 12.1 (Feb. 2014), pp. 23–33. DOI: 10.1016/j.photonics.2013.07.008.

[33] D. Englund et al. "Controlling the Spontaneous Emission Rate of Single Quantum Dots in a Two-Dimensional Photonic Crystal". In: *Physical Review Letters* 95.1 (July 2005), p. 013904. DOI: 10.1103/PhysRevLett.95.013904.

[34] A. W. Fang et al. "Electrically pumped hybrid AlGaInAs-silicon evanescent laser". In: *Optics Express* 14.20 (Oct. 2006), pp. 9203–9210. DOI: 10.1364/OE.14.009203.

[35] G. W. Ford and W. H. Weber. "Electromagnetic interactions of molecules with metal surfaces". In: *Physics Reports* 113.4 (Nov. 1984), pp. 195–287. DOI: 10.1016/0370-1573(84)90098-X.

[36] S. A. Fortuna. "Integrated Nanoscale Antenna-LED for On-Chip Optical Communication". PhD Thesis. UC Berkeley, 2017.

[37] S. A. Fortuna et al. "Large spontaneous emission rate enhancement from an electrically-injected nanoLED coupled to an optical antenna". In: *2015 IEEE Photonics Conference (IPC)*. Oct. 2015, pp. 172–173. DOI: 10.1109/IPCon.2015.7323683.

[38] L. F. Frellsen et al. "Topology optimized mode multiplexing in silicon-on-insulator photonic wire waveguides". In: *Optics Express* 24.15 (July 2016), pp. 16866–16873. DOI: 10.1364/OE.24.016866.

[39] H. Gehring et al. "Low-loss fiber-to-chip couplers with ultrawide optical bandwidth". In: *APL Photonics* 4.1 (Jan. 2019), p. 010801. DOI: 10.1063/1.5064401.

[40] B. Gelmont and M. Shur. "Spreading resistance of a round ohmic contact". In: *Solid-State Electronics* 36.2 (Feb. 1993), pp. 143–146. DOI: 10.1016/0038-1101(93)90132-A.

[41] V. Giannini and J. A. Sánchez-Gil. "Excitation and emission enhancement of single molecule fluorescence through multiple surface-plasmon resonances on metal trimer nanoantennas". In: *Optics Letters* 33.9 (May 2008), pp. 899–901. DOI: 10.1364/OL.33.000899.

[42] M. T. Hill et al. "Lasing in metallic-coated nanocavities". In: *Nature Photonics* 1.10 (Oct. 2007), pp. 589–594. DOI: 10.1038/nphoton.2007.171.

[43] T. B. Hoang et al. "Ultrafast spontaneous emission source using plasmonic nanoantennas". In: *Nature Communications* 6.1 (July 2015), p. 7788. DOI: 10.1038/ncomms8788.

[44] G. B. Hoffman et al. "Improved broadband performance of an adjoint shape optimized waveguide crossing using a Levenberg-Marquardt update". In: *Optics Express* 27.17 (Aug. 2019), pp. 24765–24780. DOI: 10.1364/OE.27.024765.

[45] S. Hooten. *reqns*. 2018. URL: https://github.com/smhooten/reqns.

[46] S. Hooten and Z. Omair. *TMatrixOpt*. 2020. URL: https://gihub.com/smhooten/TMatrixOpt.

[47] S. Hooten et al. "Adjoint Optimization of Efficient CMOS-Compatible Si-SiN Vertical Grating Couplers for DWDM Applications". In: *Journal of Lightwave Technology* 38.13 (July 2020), pp. 3422–3430. DOI: 10.1109/JLT.2020.2969097.

[48] S. Hooten et al. "nanoLED Wavelength Division Multiplexer Analysis". In: Optical Society of America, May 2019, FW3C.6. DOI: 10.1364/CLEO_QELS.2019.FW3C.6.

[49] S. Hu and S. M. Weiss. "Design of Photonic Crystal Cavities for Extreme Light Concentration". In: *ACS Photonics* 3.9 (Sept. 2016), pp. 1647–1653. DOI: 10.1021/acsphotonics.6b00219.

[50] D. Huang et al. "High-power sub-kHz linewidth lasers fully integrated on silicon". In: *Optica* 6.6 (June 2019), pp. 745–752. DOI: 10.1364/OPTICA.6.000745.

[51] K. C. Y. Huang et al. "Antenna electrodes for controlling electroluminescence". In: *Nature Communications* 3 (Aug. 2012), p. 1005. DOI: 10.1038/ncomms1985.

[52] T. W. Hughes et al. "Adjoint Method and Inverse Design for Nonlinear Nanophotonic Devices". In: *ACS Photonics* 5.12 (Dec. 2018), pp. 4781–4787. DOI: 10.1021/acsphotonics.8b01522.

[53] Luxtera Inc. *Luxtera Debuts Duplex 100G-CWDM2 Optical Transceiver Module at OFC 2017*. Mar. 2017. URL: https://www.globenewswire.com/news-release/2017/03/20/1228601/0/en/Luxtera-Debuts-Duplex-100G-CWDM2-Optical-Transceiver-Module-at-OFC-2017.html (visited on 04/27/2021).

[54] H. Iwase, D. Englund, and J. Vučković. "Analysis of the Purcell effect in photonic and plasmonic crystals with losses". In: *Optics Express* 18.16 (Aug. 2010), pp. 16546–16560. DOI: 10.1364/OE.18.016546.

[55] D. J. Jackson. *Classical Electrodynamics*. 3rd ed. Wiley, Aug. 1998. ISBN: 978-0-471-30932-1.

[56] J. B. Jackson and N. J. Halas. "Surface-enhanced Raman scattering on tunable plasmonic nanoparticle substrates". In: *Proceedings of the National Academy of Sciences* 101.52 (Dec. 2004), pp. 17930–17935. DOI: 10.1073/pnas.0408319102.

[57] J. Jiang and J. A. Fan. "Global Optimization of Dielectric Metasurfaces Using a Physics-Driven Neural Network". In: *Nano Letters* 19.8 (Aug. 2019), pp. 5366–5372. DOI: 10.1021/acs.nanolett.9b01857.

[58] Y.-H. Jin, B. J. Park, and M.-K. Kim. "Extreme field enhancement in nano-gap plasmonic cavity via 90% efficient coupling with silicon waveguide". In: *Opt. Express, OE* 24.22 (Oct. 2016), pp. 25540–25547. DOI: 10.1364/OE.24.025540.

[59] N. Jones. "How to stop data centres from gobbling up the world's electricity". In: *Nature* 561.7722 (Sept. 2018), pp. 163–166. DOI: 10.1038/d41586-018-06610-y.

[60] S. C. Kan et al. "On the effects of carrier diffusion and quantum capture in high speed modulation of quantum well lasers". In: *Applied Physics Letters* 61.7 (Aug. 1992), pp. 752–754. DOI: 10.1063/1.107787.

[61] Y. A. Kelaita et al. "Hybrid metal-dielectric nanocavity for enhanced light-matter interactions". In: *Optical Materials Express* 7.1 (Jan. 2017), pp. 231–239. DOI: 10.1364/OME.7.000231.

[62] J. Kern et al. "Electrically driven optical antennas". In: *Nature Photonics* 9.9 (Sept. 2015), pp. 582–586. DOI: 10.1038/nphoton.2015.141.

[63] M.-K. Kim et al. "Engineering of metal-clad optical nanocavity to optimize coupling with integrated waveguides". In: *Opt. Express, OE* 21.22 (Nov. 2013), pp. 25796–25804. DOI: 10.1364/OE.21.025796.

[64] A. F. Koenderink. "On the use of Purcell factors for plasmon antennas". In: *Optics Letters* 35.24 (Dec. 2010), pp. 4208–4210. DOI: 10.1364/OL.35.004208.

[65] J. Komma et al. "Thermo-optic coefficient of silicon at 1550 nm and cryogenic temperatures". In: *Applied Physics Letters* 101.4 (2012), p. 041905. DOI: 10.1063/1.4738989.

[66] A. N. Koya et al. "Novel Plasmonic Nanocavities for Optical Trapping-Assisted Biosensing Applications". In: *Advanced Optical Materials* 8.7 (2020), p. 1901481. DOI: https://doi.org/10.1002/adom.201901481.

[67] A. E. Krasnok et al. "An antenna model for the Purcell effect". In: *Scientific Reports* 5.1 (Aug. 2015), p. 12956. DOI: 10.1038/srep12956.

[68] U. Kreibig and M. Vollmer. *Optical Properties of Metal Clusters*. 1st ed. Vol. 25. Springer-Verlag Berlin Heidelberg, 1995. ISBN: 978-3-540-57836-9.

[69] P. Lalanne et al. "Light Interaction with Photonic and Plasmonic Resonances". In: *Laser & Photonics Reviews* 12.5 (2018), p. 1700113. DOI: https://doi.org/10.1002/lpor.201700113.

[70] C. M. Lalau-Keraly. "Optimizing Nanophotonics: from Photoreceivers to Waveguides". PhD thesis. University of California, Berkeley, May 2017.

[71] C. M. Lalau-Keraly et al. "Adjoint shape optimization applied to electromagnetic design". In: *Opt. Express, OE* 21.18 (Sept. 2013), pp. 21693–21701. DOI: 10.1364/OE.21.021693.

[72] L. D. Landau and E. M. Lifshitz. *Electrodynamics of Continuous Media*. Pergamon, 1960. ISBN: 978-0-08-030275-1.

[73] X. Li and Q. Gu. "Ultrafast shifted-core coaxial nano-emitter". In: *Optics Express* 26.12 (June 2018), pp. 15177–15185. DOI: 10.1364/OE.26.015177.

[74] P. Lodahl, S. Mahmoodian, and S. Stobbe. "Interfacing single photons and single quantum dots with photonic nanostructures". In: *Reviews of Modern Physics* 87.2 (May 2015), pp. 347–400. DOI: 10.1103/RevModPhys.87.347.

[75] Y. London et al. "Energy Efficiency Analysis of Comb Source Carrier-Injection Ring-Based Silicon Photonic Link". In: *IEEE Journal of Selected Topics in Quantum Electronics* 26.2 (Mar. 2020), pp. 1–13. DOI: 10.1109/JSTQE.2019.2934121.

[76] J. Lu and J. Vučković. "Objective-first design of high-efficiency, small-footprint couplers between arbitrary nanophotonic waveguide modes". In: *Optics Express* 20.7 (Mar. 2012), pp. 7221–7236. DOI: 10.1364/OE.20.007221.

[77] J. C. C. Mak et al. "Multi-layer silicon nitride-on-silicon polarization-independent grating couplers". In: *Optics Express* 26.23 (Nov. 2018), pp. 30623–30633. DOI: 10.1364/OE.26.030623.

[78] J. C. C. Mak et al. "Silicon nitride-on-silicon bi-layer grating couplers designed by a global optimization method". In: *Optics Express* 26.10 (May 2018), pp. 13656–13665. DOI: 10.1364/OE.26.013656.

[79] N. Mangal et al. "Expanded Beam Backside Coupling Interface for Alignment-Tolerant Packaging of Silicon Photonics". In: *IEEE Journal of Selected Topics in Quantum Electronics* (2019), pp. 1–1. DOI: 10.1109/JSTQE.2019.2934161.

[80] R. Marchetti et al. "High-efficiency grating-couplers: demonstration of a new design strategy". In: *Scientific Reports* 7.1 (Nov. 2017), pp. 1–8. DOI: 10.1038/s41598-017-16505-z.

[81] E. Masanet et al. "Recalibrating global data center energy-use estimates". In: *Science* 367.6481 (Feb. 2020), pp. 984–986. DOI: 10.1126/science.aba3758.

[82] S. Mathai et al. "Detachable 1x8 single-mode optical interface for DWDM microring silicon photonic transceivers". In: *2020 SPIE Photonics West Conference (OPTO, Optical Interconnects XX)*. Feb. 2020.

[83] Y. Matsui et al. "30-GHz bandwidth 1.55-μm strain-compensated InGaAlAs-InGaAsP MQW laser". In: *IEEE Photonics Technology Letters* 9.1 (Jan. 1997), pp. 25–27. DOI: 10.1109/68.554159.

[84] Y. Matsui et al. "55 GHz Bandwidth Distributed Reflector Laser". In: *Journal of Lightwave Technology* 35.3 (Feb. 2017), pp. 397–403.

[85] A. Mekis et al. "A Grating-Coupler-Enabled CMOS Photonics Platform". In: *IEEE Journal of Selected Topics in Quantum Electronics* 17.3 (May 2011), pp. 597–608. DOI: 10.1109/JSTQE.2010.2086049.

[86] D. Melati et al. "Mapping the global design space of nanophotonic components using machine learning pattern recognition". In: *arXiv:1811.01048 [physics]* (Oct. 2018).

[87] A Michaels and E. Yablonovitch. "Leveraging continuous material averaging for inverse electromagnetic design". In: *Opt. Express, OE* 26.24 (Nov. 2018), pp. 31717–31737. DOI: 10.1364/OE.26.031717.

[88] A. Michaels. *EMopt*. May 2019. URL: https://github.com/anstmichaels/emopt.

[89] A. Michaels, M. C. Wu, and E. Yablonovitch. "Hierarchical Design and Optimization of Silicon Photonics". In: *IEEE Journal of Selected Topics in Quantum Electronics* 26.2 (Mar. 2020), pp. 1–12. DOI: 10.1109/JSTQE.2019.2935299.

[90] A. Michaels and E. Yablonovitch. "Inverse design of near unity efficiency perfectly vertical grating couplers". In: *Opt. Express, OE* 26.4 (Feb. 2018), pp. 4766–4779. DOI: 10.1364/OE.26.004766. (Visited on 03/11/2019).

[91] A. Michaels et al. "Fabrication-Tolerant Efficient Dual-Etch Grating Couplers with Low Back Reflections". In: *2018 IEEE Photonics Conference (IPC)*. Sept. 2018, pp. 1–2. DOI: 10.1109/IPCon.2018.8527318.

[92] D. A. B. Miller. "Attojoule Optoelectronics for Low-Energy Information Processing and Communications". In: *Journal of Lightwave Technology* 35.3 (2017), pp. 346–396. DOI: 10.1109/JLT.2017.2647779.

[93] D. A. B. Miller. "Device requirements for optical interconnects to silicon chips". In: *Proceedings of the IEEE* 97.7 (July 2009), pp. 1166–1185. DOI: 10.1109/JPROC.2009.2014298.

[94] D. A. B. Miller. "Rationale and challenges for optical interconnects to electronic chips". In: *Proceedings of the IEEE* 88.6 (June 2000), pp. 728–749. DOI: 10.1109/5.867687.

[95] D. A. B. Miller and H. M. Ozaktas. "Limit to the Bit-Rate Capacity of Electrical Interconnects from the Aspect Ratio of the System Architecture". In: *Journal of Parallel and Distributed Computing* 41.1 (Feb. 1997), pp. 42–52. DOI: 10.1006/jpdc.1996.1285.

[96] J. Missinne et al. "Alignment-tolerant interfacing of a photonic integrated circuit using back side etched silicon microlenses". In: *Silicon Photonics XIV*. Vol. 10923. International Society for Optics and Photonics, Mar. 2019, p. 1092304. DOI: 10.1117/12.2506159.

[97] S. Molesky et al. "Inverse design in nanophotonics". In: *Nature Photonics* 12.11 (Nov. 2018), pp. 659–670. DOI: 10.1038/s41566-018-0246-9.

[98] P. Mühlschlegel et al. "Resonant Optical Antennas". In: *Science* 308.5728 (June 2005), pp. 1607–1609. DOI: 10.1126/science.1111886.

[99] O. L. Muskens et al. "Strong Enhancement of the Radiative Decay Rate of Emitters by Single Plasmonic Nanoantennas". In: *Nano Letters* 7.9 (Sept. 2007), pp. 2871–2875. DOI: 10.1021/nl0715847.

[100] S. Nambiar, P. Sethi, and S. K. Selvaraja. "Grating-Assisted Fiber to Chip Coupling for SOI Photonic Circuits". In: *Applied Sciences* 8.7 (July 2018), p. 1142. DOI: 10.3390/app8071142.

[101] S. Nie and S. R. Emory. "Probing Single Molecules and Single Nanoparticles by Surface-Enhanced Raman Scattering". In: *Science* 275.5303 (Feb. 1997), pp. 1102–1106. DOI: 10.1126/science.275.5303.1102.

[102] L. Novotny and N. van Hulst. "Antennas for light". In: *Nature Photonics* 5.2 (Feb. 2011), pp. 83–90. DOI: 10.1038/nphoton.2010.237.

[103] K. Nozaki et al. "Amplifier-Free Bias-Free Receiver Based on Low-Capacitance Nanophotodetector". In: *IEEE Journal of Selected Topics in Quantum Electronics* 24.2 (Mar. 2018), pp. 1–11. DOI: 10.1109/JSTQE.2017.2777105.

[104] K. Nozaki et al. "Forward-biased nanophotonic detector for ultralow-energy dissipation receiver". In: *APL Photonics* 3.4 (Apr. 2018), p. 046101. DOI: 10.1063/1.5022074.

[105] Z. Omair, S. Hooten, and E. Yablonovitch. "Optimized Optics for Highly Efficient Photovoltaic Devices". In: *2020 47th IEEE Photovoltaic Specialists Conference (PVSC)*. June 2020, pp. 1813–1815. DOI: 10.1109/PVSC45281.2020.9300579.

[106] Z. Omair et al. "Ultraefficient thermophotovoltaic power conversion by band-edge spectral filtering". In: *Proceedings of the National Academy of Sciences* 116.31 (July 2019), pp. 15356–15361. DOI: 10.1073/pnas.1903001116.

[107] R. F. Oulton et al. "A hybrid plasmonic waveguide for subwavelength confinement and long-range propagation". In: *Nature Photonics* 2.8 (Aug. 2008), pp. 496–500. DOI: 10.1038/nphoton.2008.131.

[108] E. D. Palik. *Handbook of Optical Constants of Solids*. Elsevier, 1997.

[109] H.-G. Park et al. "Electrically Driven Single-Cell Photonic Crystal Laser". In: *Science* 305.5689 (Sept. 2004), pp. 1444–1447. DOI: 10.1126/science.1100968.

[110] M. Pelton. "Modified spontaneous emission in nanophotonic structures". In: *Nature Photonics* 9.7 (July 2015), pp. 427–435. DOI: 10.1038/nphoton.2015.103.

[111] A. Y. Piggott et al. "Fabrication-constrained nanophotonic inverse design". In: *Scientific Reports* 7.1 (May 2017), p. 1786. DOI: 10.1038/s41598-017-01939-2.

[112] E. M. Purcell, H. C. Torrey, and R. V. Pound. "Resonance Absorption by Nuclear Magnetic Moments in a Solid". In: *Physical Review* 69.1-2 (Jan. 1946), pp. 37–38. DOI: 10.1103/PhysRev.69.37.

[113] S. Ramo. "Currents Induced by Electron Motion". In: *Proceedings of the IRE* 27.9 (Sept. 1939), pp. 584–585. DOI: 10.1109/JRPROC.1939.228757.

[114] W. D. Sacher et al. "Wide bandwidth and high coupling efficiency Si3N4-on-SOI dual-level grating coupler". In: *Opt. Express* 22.9 (May 2014), pp. 10938–10947. DOI: 10.1364/OE.22.010938.

[115] N. V. Sapra et al. "Inverse Design and Demonstration of Broadband Grating Couplers". In: *IEEE Journal of Selected Topics in Quantum Electronics* 25.3 (May 2019), pp. 1–7. DOI: 10.1109/JSTQE.2019.2891402.

[116] C. Sauvan et al. "Theory of the Spontaneous Optical Emission of Nanosize Photonic and Plasmon Resonators". In: *Physical Review Letters* 110.23 (June 2013), p. 237401. DOI: 10.1103/PhysRevLett.110.237401.

[117] C. Scarcella et al. "Pluggable Single-Mode Fiber-Array-to-PIC Coupling Using Micro-Lenses". In: *IEEE Photonics Technology Letters* 29.22 (Nov. 2017), pp. 1943–1946. DOI: 10.1109/LPT.2017.2757082.

[118] S. A. Schelkunoff and H. T. Friis. *Antennas: Theory and Practice*. John Wiley & Sons, 1952. ISBN: 978-0-471-75900-3.

[119] T. J. Seok et al. "Wafer-scale silicon photonic switches beyond die size limit". In: *Optica* 6.4 (Apr. 2019), pp. 490–494. DOI: 10.1364/OPTICA.6.000490.

[120] K. T. Settaluri et al. "First Principles Optimization of Opto-Electronic Communication Links". In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 64.5 (May 2017), pp. 1270–1283. DOI: 10.1109/TCSI.2016.2633942.

[121] B. Shen et al. "An integrated-nanophotonics polarization beamsplitter with 2.4 $\times$ 2.4 ^{2}$ footprint". In: *Nature Photonics* 9.6 (June 2015), pp. 378–382. DOI: 10.1038/nphoton.2015.80.

[122] Y. Shen et al. "Deep learning with coherent nanophotonic circuits". In: *Nature Photonics* 11.7 (July 2017), pp. 441–446. DOI: 10.1038/nphoton.2017.93.

[123] G. Son et al. "High-efficiency broadband light coupling between optical fibers and photonic integrated circuits". In: *Nanophotonics* 7.12 (2018), pp. 1845–1864. DOI: 10.1515/nanoph-2018-0075.

[124] M. Staffaroni et al. "Circuit analysis in metal-optics". In: *Photonics and Nanostructures - Fundamentals and Applications* 10.1 (Jan. 2012), pp. 166–176. DOI: 10.1016/j.photonics.2011.12.002.

[125] L. Su et al. "Fully-automated optimization of grating couplers". In: *Optics Express* 26.4 (Feb. 2018), pp. 4023–4034. DOI: 10.1364/OE.26.004023.

[126] T. Suhr et al. "Modulation response of nanoLEDs and nanolasers exploiting Purcell enhanced spontaneous emission". In: *Optics Express* 18.11 (May 2010), pp. 11230–11241. DOI: 10.1364/OE.18.011230.

[127] S. Sun et al. "Metal–Dielectric Hybrid Dimer Nanoantenna: Coupling between Surface Plasmons and Dielectric Resonances for Fluorescence Enhancement". In: *The Journal of Physical Chemistry C* 121.23 (June 2017), pp. 12871–12884. DOI: 10.1021/acs.jpcc.7b02593.

[128] Y. Sun et al. "Metal-dielectric nanoantenna for radiation control of a single-photon emitter". In: *Optical Materials Express* 10.1 (Jan. 2020), pp. 29–35. DOI: 10.1364/OME.10.000029.

[129] D. Taillaert, P. Bienstman, and R. Baets. "Compact efficient broadband grating coupler for silicon-on-insulator waveguides". In: *Optics Letters* 29.23 (Dec. 2004), pp. 2749–2751. DOI: 10.1364/OL.29.002749.

[130] K. Takeda et al. "Heterogeneously integrated photonic-crystal lasers on silicon for on/off chip optical interconnects". In: *Optics Express* 23.2 (Jan. 2015), pp. 702–708. DOI: 10.1364/OE.23.000702.

[131] T. H. Taminiau et al. "Optical antennas direct single-molecule emission". In: *Nature Photonics* 2.4 (Apr. 2008), pp. 234–237. DOI: 10.1038/nphoton.2008.32.

[132] D. Thomson et al. "Roadmap on silicon photonics". In: *Journal of Optics* 18.7 (June 2016), p. 073003. DOI: 10.1088/2040-8978/18/7/073003.

[133] R. S. Tucker. "Large-signal switching transients in index-guided semiconductor lasers". In: *Electronics Letters* 20.19 (Sept. 1984), pp. 802–803.

[134] L. Verslegers et al. "Design of Low-Loss Polarization Splitting Grating Couplers". In: *Advanced Photonics for Communications*. Optical Society of America, 2014, JT4A.2.

[135] M. T. Wade et al. "75% efficient wide bandwidth grating couplers in a 45 nm microelectronics CMOS process". In: *2015 IEEE Optical Interconnects Conference (OI)*. Apr. 2015, pp. 46–47. DOI: 10.1109/OIC.2015.7115679.

[136] Y. Wan et al. "1.3μm submilliamp threshold quantum dot micro-lasers on Si". In: *Optica* 4.8 (Aug. 2017), pp. 940–944. DOI: 10.1364/OPTICA.4.000940.

[137] T. Watanabe, Y. Fedoryshyn, and J. Leuthold. "2-D Grating Couplers for Vertical Fiber Coupling in Two Polarizations". In: *IEEE Photonics Journal* 11.4 (Aug. 2019), pp. 1–9. DOI: 10.1109/JPHOT.2019.2926823.

[138] T. Watanabe et al. "Perpendicular Grating Coupler Based on a Blazed Antiback-Reflection Structure". In: *Journal of Lightwave Technology* 35.21 (Nov. 2017), pp. 4663–4669.

[139] S. Weisser et al. "Damping-limited modulation bandwidths up to 40 GHz in undoped short-cavity In/sub 0.35/Ga/sub 0.65/As-GaAs multiple-quantum-well lasers". In: *IEEE Photonics Technology Letters* 8.5 (May 1996), pp. 608–610. DOI: 10.1109/68.491554.

[140] H. A. Wheeler. "Fundamental Limitations of Small Antennas". In: *Proceedings of the IRE* 35.12 (Dec. 1947), pp. 1479–1484. DOI: 10.1109/JRPROC.1947.226199.

[141] Q. Wilmart et al. "A Versatile Silicon-Silicon Nitride Photonics Platform for Enhanced Functionalities and Applications". In: *Applied Sciences* 9.2 (Jan. 2019), p. 255. DOI: 10.3390/app9020255.

[142] X. Xu et al. "11 TOPS photonic convolutional accelerator for optical neural networks". In: *Nature* 589.7840 (Jan. 2021), pp. 44–51. DOI: 10.1038/s41586-020-03063-0.

[143] E. Yablonovitch and E. Kane. "Reduction of lasing threshold current density by the lowering of valence band effective mass". In: *Journal of Lightwave Technology* 4.5 (May 1986), pp. 504–506. DOI: 10.1109/JLT.1986.1074751.

[144] Y. Yamamoto et al. "Coherent Ising machines—optical neural networks operating at the quantum limit". In: *npj Quantum Information* 3.1 (Dec. 2017), pp. 1–15. DOI: 10.1038/s41534-017-0048-9.

[145] S. Yamaoka et al. "Directly modulated membrane lasers with 108 GHz bandwidth on a high-thermal-conductivity silicon carbide substrate". In: *Nature Photonics* 15.1 (Jan. 2021), pp. 28–35. DOI: 10.1038/s41566-020-00700-y.

[146] B. Yang et al. "Sub-nanometre resolution in single-molecule photoluminescence imaging". In: *Nature Photonics* 14.11 (Nov. 2020), pp. 693–699. DOI: 10.1038/s41566-020-0677-y.

[147] Y. Yang et al. "Low-Loss Plasmonic Dielectric Nanoresonators". In: *Nano Letters* 17.5 (May 2017), pp. 3238–3245. DOI: 10.1021/acs.nanolett.7b00852.

[148] A. Yariv. *Quantum Electronics*. 3rd ed. John Wiley & Sons, 1989.

[149] J. Yu and H. Yamada. "Design and investigation of a dual-layer grating coupler for efficient vertical fiber-chip coupling". In: *Applied Physics Express* 12.1 (Dec. 2018), p. 012004. ISSN: 1882-0786. DOI: 10.7567/1882-0786/aaf21f.

[150] W. S. Zaoui et al. "Bridging the gap between optical fibers and silicon photonic integrated circuits". In: *Optics Express* 22.2 (Jan. 2014), pp. 1277–1286. DOI: 10.1364/OE.22.001277.

[151] E. Zielinski et al. "Excitonic transitions and exciton damping processes in InGaAs/InP". In: *Journal of Applied Physics* 59.6 (Mar. 1986), pp. 2196–2204. DOI: 10.1063/1.336358.

[152] D. A. Zuev et al. "Fabrication of Hybrid Nanostructures via Nanoscale Laser-Induced Reshaping for Advanced Light Manipulation". In: *Advanced Materials* 28.16 (Apr. 2016), pp. 3087–3093. DOI: 10.1002/adma.201505346.