7-21-2023

Design and Optimization of Devices and Architectures for Neuromorphic Silicon Photonics

Matthew van Niekerk
mv7146@rit.edu

RIT

Design and Optimization of Devices and Architectures for Neuromorphic Silicon Photonics

by

Matthew van Niekerk

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctorate of Philosophy in Microsystems Engineering

Microsystems Engineering Program
Kate Gleason College of Engineering

Rochester Institute of Technology
Rochester, New York
July 21, 2023

Design and Optimization of Devices and Architectures for Neuromorphic Silicon Photonics

by

Matthew van Niekerk

Committee Approval:

We, the undersigned committee members, certify that we have advised and/or supervised the candidate on the work described in this dissertation. We further certify that we have reviewed the dissertation manuscript and approve it in partial fulfillment of the requirements of the degree of Doctorate of Philosophy in Microsystems Engineering.

Dr. Stefan F. Preble — Date
Director, Microsystems Engineering

Dr. Gregory A. Howland — Date
Associate Professor, Physics

Dr. Rui Li — Date
Associate Professor, Computer Science

Dr. Kai Ni — Date
Associate Professor, Microsystems Engineering

Certified by:

Dr. Stefan F. Preble — Date
Director, Microsystems Engineering Program

ABSTRACT

Kate Gleason College of Engineering
Rochester Institute of Technology

Degree: Doctorate of Philosophy **Program:** Microsystems Engineering
Author's Name: Matthew van Niekerk
Advisor's Name: Dr. Stefan F. Preble
Dissertation Title: Design and Optimization of Devices and Architectures for Neuromorphic Silicon Photonics

Neuromorphic photonics is an exciting field at the intersection of neuroscience, integrated photonics, and microelectronics. To realize large-scale neural network-inspired systems, we must have a toolbox of linear and nonlinear operators. Here, we review state of the art for integrated photonic linear operators, which can perform linear operations using interference or wavelength-division-multiplexing (WDM) techniques, and nonlinear operators, which perform nonlinear operations with a combination of photodetection and modulation. We present some devices that fit into the needed toolbox. A new type of directional coupler, using skin depth engineering of electromagnetic waves, suppresses optical crosstalk and eliminates the need for waveguide bends. An optimized high extinction ratio microring modulator, with which we demonstrate a high operating modulation frequency while having a large extinction ratio > 25 dB. We develop and improve a technique for wafer-scale thermal isolation of optical components, allowing the demonstration of a highly efficient thermo-optic phase shifter which exhibits a measured $\sim 30\times$ improvement on the power efficient and $\sim 25\times$ improvement on thermal parasitic cross talk. In addition to device-level optimization and design, we create a new neuromorphic photonics architecture that combines interference with WDM to create a massively scalable, wavelength-diverse integrated photonic linear neuron. This architecture enables dramatic parallelism and physical footprint reduction, resulting in a

highly scalable design system. We demonstrate this architecture with devices we optimized, showing single wavelength operation of a full neural network algorithm that adapts to three logic gates (AND, OR, XOR) with reconfiguration, achieving on-chip accuracies of $96.8\%, 99\%$, and 98.5%, respectively. We also demonstrate this architecture by simultaneously implementing four separate logic gates (OR, AND, XOR, NAND), projecting the outputs at four distinct wavelengths, and achieving on-chip accuracies of $99.87\%, 99.05\%, 98.05\%$ and 99.73%, respectively. Finally, we developed packages and implemented packaging techniques to address these increasingly complicated circuits like wire bonding, printed circuit board design, and flip-chip thermo-compression bonding. These efforts represent progress towards fully integrated neuromorphic photonic systems.

Acknowledgments

Naturally, there is a long list of people to thank and acknowledge for the various roles of support and encouragement they played or provided during my time both as a person and as a student.

Firstly, I thank my parents for instilling a strong sense of intrigue and academic interest from a young age. I will be forever grateful for their relentless support, love, and sacrifice for our family. My brother, rounding out the family unit, has always been kind and thoughtful and can provide a good distraction when needed. Also, I think it is important to thank my in-laws, who have always been welcoming and kind.

From there, I would acknowledge the various schools I attended throughout my academic career. My high school, Allendale Columbia, was a small school that allowed my individual interests to develop. My teachers, Mr. Fujita, Mrs. Broberg, Mr. Cruz, Dr. Jones and Mr. Hunt, for teaching problem-solving and critical thinking. I went to the University of Rochester in my first years of undergrad. While I personally have no pleasant academic memories of this place, I made a lifelong friend on the golf team, who has been a great part of my support system. I then transferred to Roberts Wesleyan College (now University). I took up Physics under Prof. Candice Fazar, who encouraged and supported me in a way I will always be grateful for – and she is ultimately responsible for introducing me to my main focus of Optics.

My time at the Rochester Institute of Technology has been extraordinary and formative. Prof. Stefan Preble and I met on my tour of the school when I was looking for a graduate program, and he took what I like to call, “a chance" on me as a student. Other than modest grades and an extended summer project, I had not been around the field of researching science or integrated photonics. Nevertheless, he found a place for me and allowed me to get into this rich and satisfying work. It is also worth mentioning my gratitude for his guidance, academic curiosity, and care not only for the work but for me to engage and understand at every stage of the journey. My groupmates that have been there, as well, are all worth thanking: Hector Rivera, Venkatesh Deenadalayan, Karl McNulty, Vijay Sundaram, and Lilian Neim. Our wonderful technical engineers offered wisdom, experience, and mostly patience: Thomas Palone, Mario Ciminelli, and Randy Kennard. Past group members who I thank for their patience in explaining things once and then again slowly: Dr. Jeff Steidle, Dr. Peichuan Yin, Dr. David Starling, Dr. John Serafini, and Dr. Paul Thomas. I want to also thank Prof. Gregory Howland for many things, especially for being a veritable well of information, common sense, and humor. His student Evan Manfreda is also a great collaborator and labmate.

Outside of RIT, we worked with several other teams, groups, and schools, including Air Force Research Labs (AFRL), Purdue University, Army Research Labs (ARL), Columbia University, MIT, SUNY Polytechnic, L3 Harris, and AIM Photonics to highlight the key collaborators I interacted with. From these institutions, I want to thank Dr. Anthony Rizzo (AFRL/Columbia), Dr. Navin Lingaraju (Purdue), Dr. Michael Fanto (AFRL), Dr. Christopher Tison (AFRL), Karthik Myilswamy (Purdue), Lucas Cohen (Purdue), Dr. Saman Jahani (Purdue), Gerald Leake (SUNY Polytechnic), Prof. Zubin Jacob (Purdue), Prof. Keren Bergman (Columbia). All of these people contributed or collaborated on much of the work I was involved with, and without these and my groupmates at RIT, none of the work would have been interesting or completed. In addition, I want to thank the funding agencies, namely the NSF (DMR-1747426, 1809695) and AFRL. I also want to thank my therapist for hours of working through life's challenges.

I also began working full-time towards the end of my academic career for Nokia as a PIC design engineer. This has been a very challenging yet rewarding experience. I want to thank many folks, especially those I have worked closely with, Ajay Mistry, Sequoia Ploeg, Matt Streshinsky, Xiaochen Ge, Subhojit Dutta, and Christian Carver.

Finally, I want to thank my wife, Shannon. It's hard to capture in words what her support has meant throughout the last five years, through which we have gone through a doctorate program, a global pandemic, moving, buying a house, getting a dog (also massive thanks to our dog, Ernie), not to mention other ebbs and flows of life's journey. The work here is impossible without her.

Contents

List of Tables

List of Figures

List of Conference Proceedings

** - corresponding author, † - authors contributed equally*

1. V. Deenadayalan, M. Fanto, M. van Niekerk, and S. Preble, 'Ultra-low voltage silicon photonic mems phase shifter', in 2020 IEEE Photonics Conference (IPC), 2020, pp. 1–2.

2. M. Van Niekerk* et al., 'Experimental Evolutionary Optimization of an Active Multimode Interferometer', in CLEO: Fundamental Science, 2020, pp. JTh2B-17.

3. M. van Niekerk* et al., 'Demonstration of Two-Dimensional Extreme Skin Depth Engineering in CMOS Photonics Foundry', in Frontiers in Optics, 2020, pp. FTh1C-5.

4. V. Deenadayalan*,†, M. van Niekerk†, M. Fanto, and S. Preble, 'Silicon photonic MEMS phase shifter using gradient electric force actuation', in Frontiers in Optics, 2020, pp. FW5D-3.

5. G. Bond, T. Palone, M. van Niekerk, et al., 'Direct attachment of optical fibers to photonic integrated circuits with in situ UV curing', in CLEO: Fundamental Science, 2021, pp. JW1A-29.

6. J. Serafini, M. van Niekerk, M. Fanto, and S. Preble, 'Tunable, High Purity Two-Photon Interference From Independent Sources On a Silicon Photonic Chip', in 2021 IEEE Photonics Conference (IPC), 2021, pp. 1–2.

7. L. G. Carpenter*,†, M. van Niekerk†, et al., 'Towards low propagation losses in active photonic multi-project wafer runs', in Integrated Photonics Research, Silicon and Nanophotonics, 2021, pp. ITu3A-5.

8. H. A. R. Rivera*,†, M. van Niekerk†, and S. F. Preble, 'Silicon photonic optical-electrical-optical modulator neuron verilog-a model', in CLEO: Fundamental Science, 2021, pp. JTu3A-125.

9. E. Manfreda-Schulz et al., 'Generation of High-Dimensional Entanglement on a Silicon Photonic Chip', in Quantum 2.0, 2022, pp. QTu4B-4.

10. J. Monteleone et al., 'Packaged foundry-fabricated silicon spiral photon pair source', in Quantum Nanophotonic Materials, Devices, and Systems 2022, 2022, vol. 12206, pp. 29–34.

11. V. Murthy et al., 'Mitigation of parasitic junction formation in compact resonant modulators with doped silicon heaters', in Laser Resonators, Microresonators, and Beam Control XXIV, 2022, vol. 11987, pp. 114–125.

12. K. McNulty, M. van Niekerk, et al., 'Wafer scale fabrication of silicon nitride MEMS phase shifters with XeF2 dry vapor release etch process', in Silicon Photonics XVII, 2022, vol. 12006, pp. 74–79.

13. M. van Niekerk*, C. Cheng, G. Leake, D. Coleman, M. L. Fanto, and S. F. Preble, 'High Extinction Ratio Microring Modulator', in CLEO: Science and Innovations, 2022, pp. STh4K-5.

14. M. van Niekerk* et al., 'Wafer-scale-compatible substrate undercut for ultra-efficient SOI thermal phase shifters', in 2022 Conference on Lasers and Electro-Optics (CLEO), 2022, pp. 1–2.

15. H. A. R. Rivera*,†, M. van Niekerk†, and S. F. Preble, 'Silicon Photonic Optical-Electrical-Optical Conversion with Gain Over Unity', in Frontiers in Optics, 2022, pp. FM1E-3.

16. M. van Niekerk* et al., 'Learning Bit-Gates With A Resonant Photonic Linear Neuron', in Laser Science, 2022, pp. JTu4A-54.

List of Publications

* - *corresponding author*, † - *authors contributed equally*

1. M. van Niekerk* et al., 'Two-dimensional extreme skin depth engineering for CMOS photonics', *JOSA B*, vol. 38, no. 4, pp. 1307–1316, 2021.

2. M. van Niekerk* et al., 'Massively Scalable Wavelength Diverse Integrated Photonic Linear Neuron', Neuromorphic Computing and Engineering, 2022.

3. A. Rizzo* et al., 'Petabit-scale silicon photonic interconnects with integrated Kerr frequency combs', IEEE Journal of Selected Topics in Quantum Electronics, vol. 29, no. 1: Non Linear Integrated Photonics, pp. 1–20, 2022.

4. S. Preble et al., 'Optical fiber attachment to a photonic integrated circuit using optical fiber-directed curing'. US Patent App. 17/720,989 2022.

Chapter 1

Introduction

1.1 Deviating from the Path of Digital Electronics

Current computing architecture harkens back to the 1945 seminal report from John von Neumann, entitled "First Draft of a Report on the EDVAC" [131]. In this report, von Neumann lays the foundation for a computing scheme that would go on to share his name. The von Neumann architecture for digital computing comprises a simple structure:

1. A CPU that handles instruction, logic, and calculations.

2. Memory for data storage.

3. Input/output streams.

Modern computers continue to rely on this architecture. Sitting on top of the von Neumann architecture, digital electronics has continuously improved over the decades along with the trend of Moore's law — which estimates the number of transistors in a central processing unit (CPU) will double every 18-24 months [24]. This trend has been steady for decades (Fig. 1.1); however, the transistor cannot continue to shrink past

the fundamental limit of atomic size [132]. Other factors become immediately important when considering future electronic scaling and improvement, especially if cramming more transistors on the chip isn't significantly improving performance. Such factors include

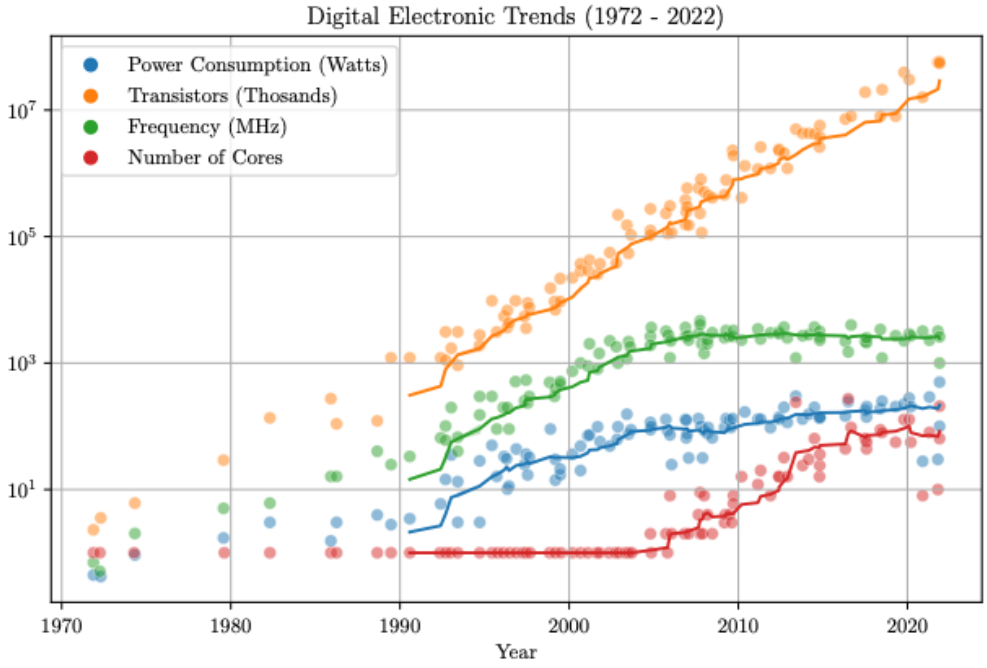

Figure 1.1: A 50-year survey of Digital Electronic trends, namely the number of transistors (in thousands) in a chip, the frequency (MHz) of the processor, the total power consumption (Watts), and the single thread performance (in kSPECint) [98]. Data Availability: `https://www.spec.org`

operating frequency, power consumption, and overall CPU performance. The operating frequency is one of the clear examples of the interconnect bottleneck (or the von Neumann bottleneck) that *architecturally* limits computing. Fig. 1.1 shows how the frequency in CPUs has plateaued in the last decade (near 4 GHz), primarily due to the physical bandwidth limitation of metallic interconnects. Power consumption is also related to this

limitation Fig. 1.1, since power (P) and frequency (f) are related by

$$P = CV^2 f, \tag{1.1}$$

for C and V as capacitance and voltage, respectively [22]. Power cannot be decreased by voltage, as voltage scaling is rapidly approaching its limit, and, conversely, it cannot be increased due to heat generation (with an upper limit near 200W) [26, 73]. Finally, the Standard Performance Evaluation Corporation (SPEC) benchmark SPECint evaluates CPUs' performance and has similarly seen a plateau over the past two decades. Performance has also begun to succumb to a barrier – in part indicated by the plateau in Fig. 1.1 of the Standard Performance Evaluation Corporation (SPEC) benchmark SPECint evaluates the performance of CPUs – where in this case software has allowed the increase instead of physical improvements. Koomey's law of computational efficiency states that the Multiply Accumulate (MAC) operations per joule of energy dissipated will double every 1.57 years [59]. Unfortunately, MAC/J deviated from this rule in the past decade, approaching an asymptote of 100 pJ per MAC. Ultimately, as electronics continue to shrink, the laws that govern the devices have not followed suit (the breaking of Dennard's law [27]). This has invigorated alternative and creative options for the future of computing.

1.2 Introduction to Neuromorphic Computing

In the age of rapidly evolving technology and information, we turn to the most powerful computer (the human brain) we know of for providing new insights and inspiration. The human brain has been a point of comparison since the very first computer. In 1947, Alan Turing wanted a "machine that can learn from experience" [125]. As the field of computing has matured, we see that the human brain serves as a direct source of inspiration, particularly from an *architectural* perspective. The rationale is simple: the efficiency of the human brain (<aJ per MAC) far exceeds digital electronics ($\sim$ 100pJ per MAC) [102]. *Neuromorphic Computing* is a non-traditional computing platform that aims to create biologically inspired hardware to overcome the issues present in von Neumann systems. In a remarkable twist of irony, von Neumann was one of the originators of the models inspiring neuromorphic computing [130]. The promise of neuromorphic computing is leveraging the brain's structure for solving certain types of problems [102]. In recent years, neuromorphic computing has become a front-runner for developing ultralow power devices with high energy efficiency and large bandwidth. Performance is one motivation for neuromorphic computing, but significant incentives are the onset of "Big Data" and the "internet of things" (IoT), wherein deployable devices are everywhere *and* are high performing [142]. As the various fields of neuroscience, computer science, device engineering, electronic engineering, biology, and the mathematics of artificial intelligence have matured, the vision for neuromorphic computing has become attainable.

The brain model can be simplified into two main components: synapses and neurons.

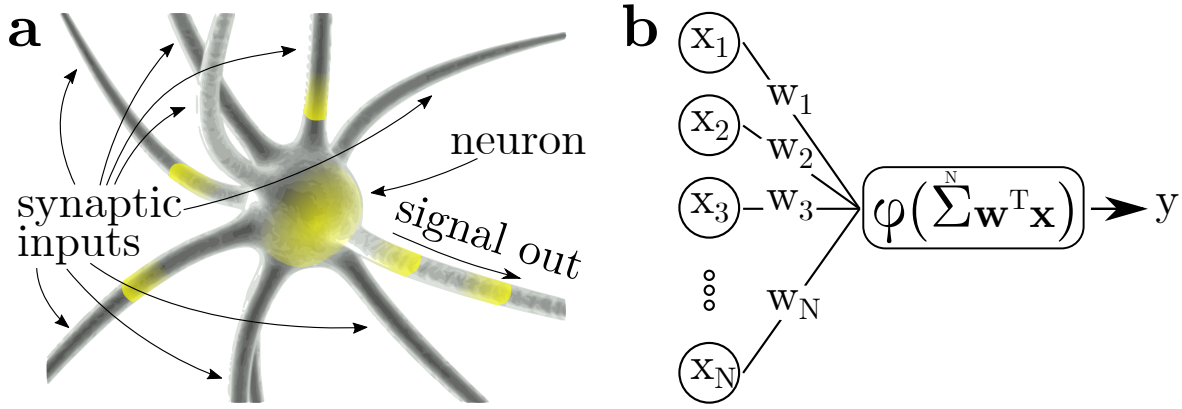

Figure 1.2: a) A biological sketch of a neuron, where the inputs "fan-in" from preceding neurons, then this neuron "decides" if the signal should be sent along, and "activates" or "fires" to the next neuron. b) The corresponding mathematical model, often called the perceptron, wherein data arrives in a layer, $\mathbf{x}$, and multiplies across the weights $\mathbf{w}$ via fan-in, where all these multiplications are summed and passed to an activation function ψ.

The synapse is responsible for connecting neurons. In contrast, the neuron is responsible for "deciding" whether or not to pass the message along to the next neuron (by performing an activation or firing). It is important to note that a neuron firing can be considered a MAC operation [102, 142]. A biological neuron (Fig. 1.2) can have a high number of input synapses (fan-in) that provide the stimulus, determining if the neuron fires its signal or not (activation). Fig. 1.2 (b) depicts a simple mathematical model (perceptron) of the biological neuron, where we perform a sum over the (linear) synaptic weights, $\mathbf{w}^T$, applied to the inputs, $\mathbf{x}$, and then subject this to a (non-linear) activation function, φ, to arrive at an output y – or:

$$y = \psi\left(\sum^{N} \mathbf{w}^T \mathbf{x}\right). \tag{1.2}$$

While models may vary throughout machine learning, this simple model (Fig. 1.2) demonstrates the representation of the synapses and neurons, the key defining features of any neuromorphic computing scheme. These two features can also be considered linear

operators (synapses) and non-linear operators (neurons).

Although neural network non-linear can operate on top of modern digital computers, the fundamental limitations of the interconnect inhibit the investigation of complex systems without an architectural overhaul [90]. The underlying mathematical model notwithstanding, physically implementing hardware with these models is critical to breaking through the computing barriers. The current state of the art in neuromorphic computing includes research in digital electronics, neuromorphic electronics (CMOS-based, phase change, memristive, spintronic, ferroelectric, metal-insulator transition), and neuromorphic photonics (subwavelength, integrated silicon photonics, hybrid silicon/III-IV photonics) [10,35,41,56,71,80,85,100]. Shown in Fig. 1.3, these three categories have distinct differences in efficiency and density. Neuromorphic electronics (projects like TrueNoth, HICANN, Loihi) have been more successful in immediate implementation as expected due to similarities with the digital electronics manufacturing process flow [90,142]. However, as Fig. 1.3 demonstrates, neuromorphic photonics has a distinct advantage in both efficiency and density.

1.3 The Case for Silicon Photonics in Neuromorphic Computing

Neuromorphic photonics itself can mean many different things. For example, it could mean free-space optical neural networks created on optical tables with lenses, mirrors, spatial light modulators, 3D-printed diffractive plates, etc. These systems have attracted

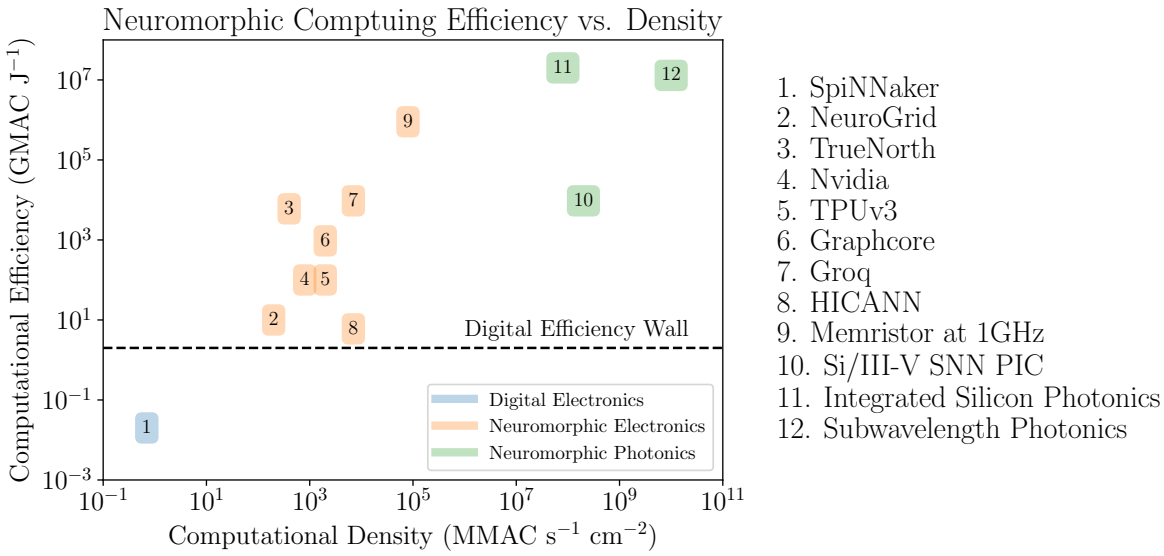

Figure 1.3: Effiency vs. Density for different approaches to Neuromorphic Computing. Adapted from Ref. [26]

some attention recently due to the heuristic similarities between quantum optics and neural networks. This approach tends to be simple for reasonably small experiments or concepts; however, creating a feasible computing system with free-space optics is ultimately limited by the physical scale [112]. Additionally, non-linear operations are traditionally challenging in anon-linear systems, wherein a typical Kerr medium high electric field intensities are needed to create relatively small effects (around the inverse of the order of the Kerr susceptibility, i.e., $\chi^{(-n)}$) [60]. While free space is a viable option for developing ideas and performing meaningful experiments, a different approach is needed for large-scale implementation, programmability, and, crucially, non-linearity.

As far as integrated photonics goes, silicon photonics has a distinct advantage over alternative approaches (i.e., III-V's, Thin Film Lithium Niobate, etc.). Silicon photonics has inherited the entire semiconductor processing industry "for free" [90]. This *gift*

cannot be understated or undervalued. The growing pains of manufacturing silicon photonics are drastically reduced, thereby accelerating the field's development, innovation, and future possibilities. The form factor, scalability, and packaging are already within the same context as digital computing. Integrated optoelectronics dramatically enhances the possibilities for non-linear operators due to photo-absorption and modulation effects. Silicon photonics enables hybrid silicon/III-IV technologies to introduce lasers within the same platform [102]. On the horizon, silicon photonics offers a platform to explore subwavelength effects for neuromorphic computing, where optimized photonic crystals or wave-shaping metamaterials can be integrated into the processing [26]. Many of these devices, circuits, or concepts are unavailable in free-space systems [112]. For these reasons, we assume that neuromorphic photonics is best suited to neural-inspired circuit design within the context of the silicon photonic platform.

1.4 Silicon Photonics Relevant Background & Theory

Silicon photonics (alternatively integrated photonics, silicon-on-insulator photonics) was driven into existence by the telecommunications industry, following a vision to integrate optics and electronics on the micro-scale [6, 109]. Silicon photonic systems are employed in the commercial arena as high-speed optical interconnects and biosensors, and, more recently, startup companies are introducing it in quantum computing and artificial intelligence applications [37, 50, 61, 62, 92, 95, 105]. In this section, we cover the relevant background for neuromorphic silicon photonics. More in-depth background can be found in these fantastic resources for silicon photonic devices [89, 134], silicon photonic circuits

and modeling [19], neuromorphic engineering [142], neuromorphic photonics [90].

1.4.1 Waveguides

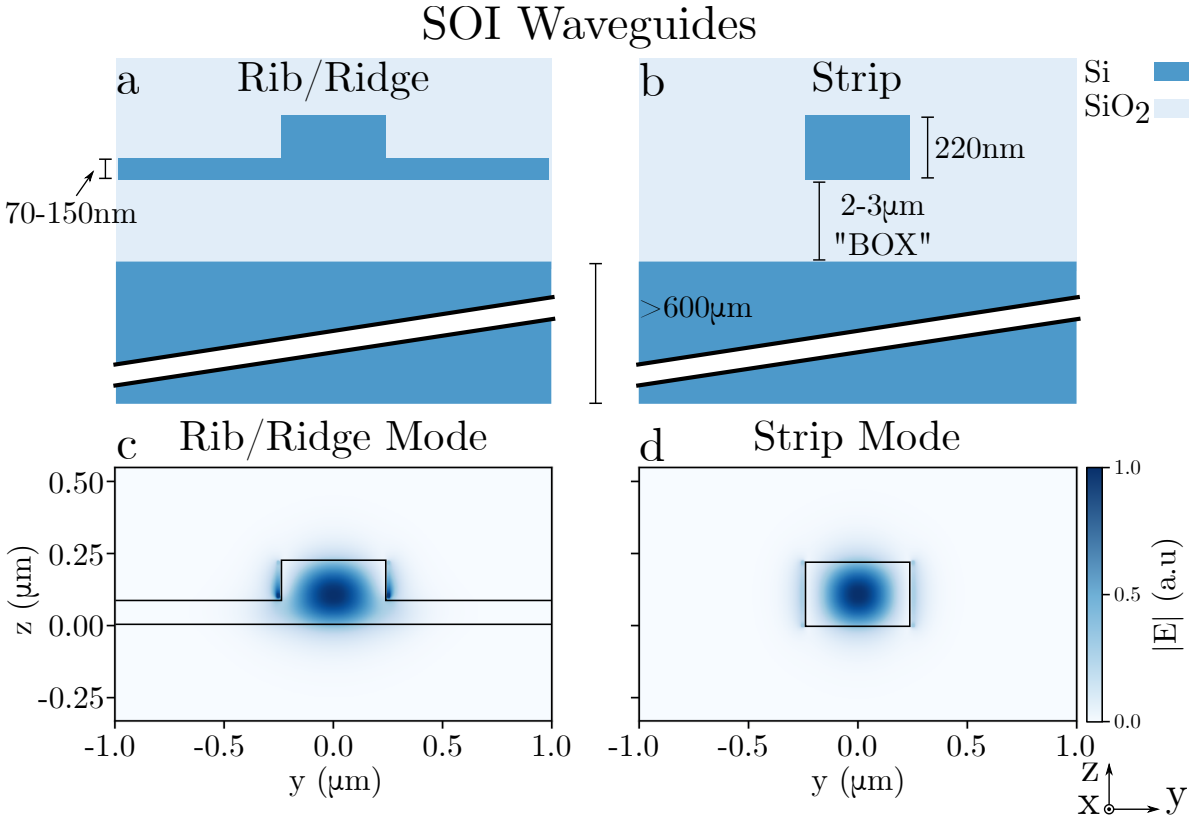

Figure 1.4: a) General cross-section of a rib/ridge waveguide, where the optical guiding portion is the raised silicon. b) General cross-section of a strip waveguide, where the optical guiding is kept to a more traditional core/cladding design. c) The electric field profile of the fundamental transverse electric (TE) mode of a). d) The electric field profile of the fundamental TE mode of b).

Waveguides operate on the fundamental concept of total internal reflection, where the light stays within a dielectric medium if its refractive index is higher than the surrounding (dielectric) medium. In such a way, light can be guided over long distances at a low loss — fiber optics — or in integrated circuits. The waveguides are created on silicon-on-insulator (SOI) wafers in the integrated circuits we will study. SOI wafers are made

up of a thick ($\sim$ 700 μm) silicon substrate, a buried silicon dioxide (BOX) layer ($\sim$ 2-3 μm), and a thin ($\sim$ 220 nm) top silicon layer (ridge and strip waveguides in Fig. 1.4 (a,b) demonstrate these physical features in SOI waveguides). Silicon has a much higher refractive index than silicon dioxide ($n_{Si,\lambda=1550\text{ nm}} = 3.45$, $n_{SiO_2,\lambda=1550\text{ nm}} = 1.44$), which creates strong confinement of the guided waveguide modes. It is worth noting the principles discussed below apply to any dielectric material pair where $n_{core} > n_{cladding}$; however, the rest of this work concerns silicon and silicon dioxide primarily.

Total internal reflection does not describe the whole picture, and here we turn to an electromagnetic image. Waveguide modes are distinct solutions to the Helmholtz wave equations for homogeneous, isotropic media (complete theoretical derivation is available in Ref. [134]). Ultimately, solving these equations for a given dielectric waveguide exposes discrete solutions known as modes. In general, waveguides support quasi transverse electric (TE), where the electric field is oriented in $\hat{y}$ (Fig.1.4), or quasi transverse magnetic (TM), where the electric field is oriented in $\hat{z}$. Generally, optical modes of silicon waveguides are simulated numerically (Fig. 1.4 (c,d)), and these simulations determine the effective index for each supported mode of a given geometry. Waveguides support as many modes as allowed by the relationship between the frequency and geometry, and single-mode operation is possible below the second mode cutoff frequency. The electromagnetic field has a single phase or propagation constant at single-mode propagation. We can mathematically describe the propagation along a waveguide of length L as:

$$\frac{E_{out}}{E_{in}} = e^{-\alpha L} e^{i\beta L}, \tag{1.3}$$

where α is the lumped propagation loss of the waveguide, β is the propagation constant for the waveguide.

For waveguide routing on-chip, bends are necessary for creating circuits. In opposition to electronics, but similarly to radio-frequency (RF) high-speed design, we must gently curve the waveguide to ensure that the optical mode is supported and is not radiated from the structure. Therefore, careful design must be taken to ensure the overlap between the straight waveguide mode and bent waveguide mode is sufficiently high. Because SOI waveguides have a high index contrast between core and cladding, the minimum bend radius can be as low as 3 μm for TE polarized light (minimum of $\sim$ 10 μm for TM). However, advanced techniques are often employed to overcome these size limitations or to reduce transmission losses. Specifically successful techniques are clothoid/Euler bends, where the radius of curvature consistently varies to decrease the mode mismatch entering and exiting the bends. Adjoint optimization, which often arrives at non-intuitive solutions for small dielectric spaces or so-called advanced bends, employs width variation along the curve to minimize radiative losses and exploit the single mode criterion [17, 81, 107].

1.4.2 Inter-connect Couplers

Interconnect couplers such as edge-couplers (or end-fire) and grating-couplers (or vertical) allow optical signals on and off the integrated photonic circuit [68]. Generally speaking, fiber is aligned (edge-coupling) or near-orthogonal (grating-coupling) to the optical axis of the waveguide. Edge-coupling is enabled by mode-matching between the

fiber and waveguide modes and grating-coupling is enabled by periodic structures (grating teeth) that phase-match the incident optical beam and a desired waveguide mode through controlled scattering. As silicon maintains an indirect bandgap in the near-infrared wavelength spectrum, inter-connect coupling is crucial due to the lack of on-chip grown lasers. Recent packaging efforts enable increased reliability for interconnect couplers and promising improvements of bonding III-V lasers directly on chip [78, 86].

1.4.3 Directional Couplers

Directional couplers consist of two parallel waveguides which are brought close together (for reference, Fig. 1.5 (a) helps to guide this discussion). The evanescent wave of the mode, even though it carries no power across the boundary, causes coupling between parallel guides if the overlap between the evanescent waves of supported modes is large enough [49]. We can describe the directional coupler similarly to a free-space beam splitter with reflection/pass and transmission/coupling coefficients t, κ, respectively. This allows for a transfer matrix representation of a simple two port device

$$\begin{bmatrix} E_3 \\ \\ E_4 \end{bmatrix} = \begin{bmatrix} t & \kappa \\ \\ \kappa & t \end{bmatrix} \begin{bmatrix} E_1 \\ \\ E_2 \end{bmatrix}. \tag{1.4}$$

In this equation, $E_{3,4}, E_{1,2}$ represent the complex electric field modal amplitudes at the outputs and inputs, respectively. The reflection and transmission coefficients here encompass the device parameters such as gap, length and waveguide dimensions. Additionally, the condition of energy conservation in a lossless coupler is $t^2 + \kappa^2 = 1$.

Alternatively, coupled mode theory offers a different perspective wherein specific parameters remain variable [49]. We define the output powers (modulus squared) as

$$|E_3|^2 = |E_1|^2 \cos^2(\kappa L) = |E_1|^2 \cos^2\left(\frac{\pi}{2}\frac{L}{L_x}\right), \tag{1.5}$$

$$|E_4|^2 = |E_1|^2 \sin^2(\kappa L) = |E_1|^2 \sin^2\left(\frac{\pi}{2}\frac{L}{L_x}\right), \tag{1.6}$$

for L as the coupling length, $|E_1|^2$ as the injected power, κ as the coupling coefficient, L_x as the crossover length such that $\kappa = \pi/(2L_x)$, and bar and cross refer to the light remaining in the injected waveguide or transitioning to the other waveguide, respectively. The crossover length is defined so that when $L = L_x$, there is complete power transfer from waveguide one to two. This approach allows for an intuitive understanding of the device such that the length is decoupled from κ, t. The crossover length is defined as

$$L_x = \frac{\lambda}{2(n_{even,\lambda} - n_{odd,\lambda})}, \tag{1.7}$$

where λ is the free space wavelength and $n_{even,\lambda}, n_{odd,\lambda}$ are the effective indices of the even and odd modes, respectively [19]. Both approaches capture the phase difference of $\pi/2$ introduced by the directional coupler either by the $-\kappa$ in Eq. 1.4 or the natural phase relationship of cosine and sine (i.e. $sin(x) = cos(x - \pi/2)$) in Eqs. 1.5, 1.6.

The matrix approach in Eq. 1.4 is useful for considering larger circuits consisting of multiple couplers. However, often understanding the directional coupler more intimately through coupled mode theory illuminates parameters prevalent in actual devices for design and fabrication. In particular, wavelength dependence is an ever present factor that

complicates circuit performance. For example, if we design a 50:50 splitter that, after fabrication, becomes 55:45 the circuit performance will be greatly diminished — particularly if this is repeated over many devices. Recent research from our group aims to create more broadband directional couplers and can be explored for larger circuits [128].

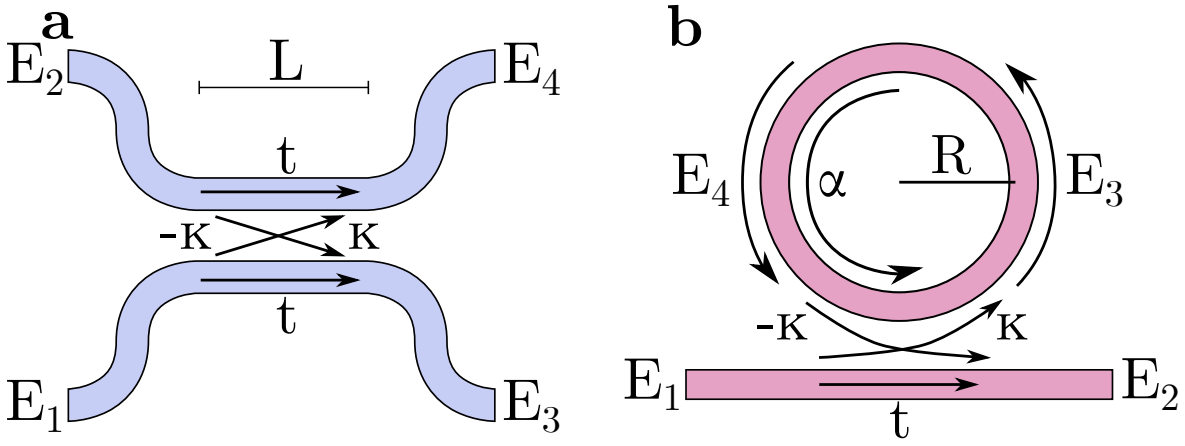

Figure 1.5: a) Diagram of an integrated directional coupler. Inputs E_1, E_2 feed through the coupling region, with pass and transmission coefficients κ, t, and are transformed into outputs E_3, E_4. b) Diagram of an integrated MRR, with input field E_1 interacting with reflection and transmission coefficients κ, t passing to a measurable output E_2. E_3 and E_4 represent the field in the ring.

1.4.4 Ring Resonators

Micro-ring resonators (MRR) are another fundemental device available in the integrated photonic platform. MRRs are deceivingly nuanced, and as such more comprehensive treatments are available in these works: Ref. [11] , Ref. [89], Ref. [134], and Ref. [47]. For our purposes, we only uncover the relevant aspects of MRRs. Similar to Eq. 1.4, we *could* describe the MRR using a full matrix formalism. This entails treating the coupling region as a directional coupler with a waveguide looped on itself. However, without re-inventing the wheel, we can define the key parameters of the ring directly. Using Fig. 1.5 (b) as a

starting place, we wish to probe E_2 assuming incident light from E_1. In this diagram, E_1 is considered the input "port," E_2 is considered the through "port." Additionally, there are coupling coefficients κ, t, for each coupling region, the ring radius, R (such that one round trip is $L = 2\pi R$), and the propagation loss in the ring, α. The analysis naturally follows (for example in Refs. [11, 47]), and we state the through port field

$$E_2 = E_1 \frac{t - e^{-\alpha L} e^{i\beta L}}{1 - t e^{-\alpha L} e^{i\beta L}}, \tag{1.8}$$

where β represents the propagation constant of the propagating mode. Here we also note that the resonant condition (or the resonant wavelength) can be described as

$$\lambda_{res} = \frac{n_{eff} L}{m}, \tag{1.9}$$

for a given resonant mode, m.

1.4.5 Mach-Zehnder Interferometers

If we assume two directional couplers have precisely 50:50 coupling ratios (implying $\kappa^2 = t^2 = 1/2$) and our waveguides have low loss (i.e. $\alpha = 0$), we can construct an integrated Mach-Zehnder Interferometer (MZI). Fig. 1.6 shows the diagram of a constructed MZI, where the stages transition from two electric field inputs, E_1 and E_2, through the first directional coupler, then each path has waveguides which consist of independent refractive indices (n_1, n_2) and path lengths (L_1, L_2), through a second directional coupler to a final

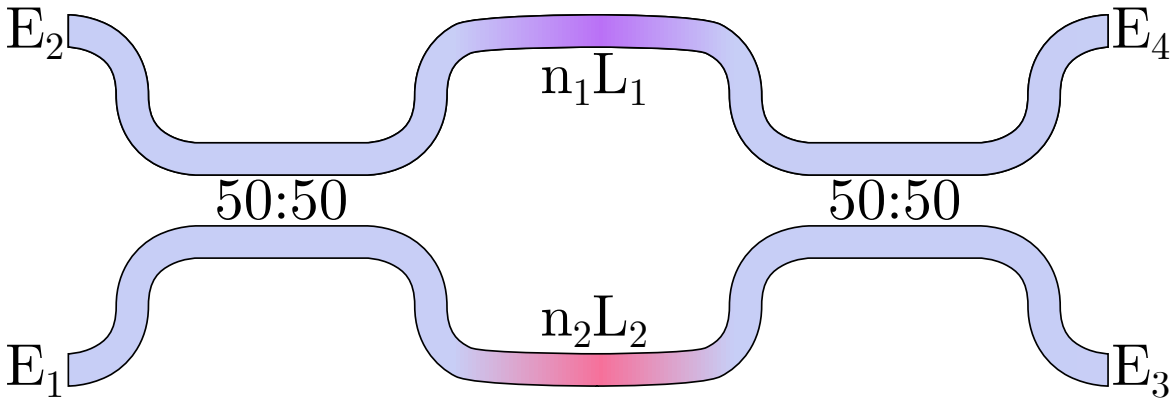

Figure 1.6: Diagram of an integrated MZI, where inputs E_1, E_2 feed through a balanced directional coupler two two waveguides, with distinct indices of refraction and lengths (which allow for capturing modulators with this derivation), to another balanced directional coupler and finally to outputs E_3, E_4

result of E_3, E_4. We translate this into an equation using the transfer matrix method:

$$\begin{bmatrix} E_3 \\ E_4 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix} \begin{bmatrix} e^{i\phi_1} & 0 \\ 0 & e^{i\phi_2} \end{bmatrix} \begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix} \begin{bmatrix} E_1 \\ E_2 \end{bmatrix}, \tag{1.10}$$

where $\phi_{1,2} = \beta_{1,2} L_{1,2}$. Perhaps more practically useful, we can express the transmission at each output of the MZI as:

$$|E_3|^2 = |E_1|^2 \left(\frac{1}{2} - \frac{1}{2} \cos(\phi) \right), \tag{1.11}$$

$$|E_4|^2 = |E_1|^2 \left(\frac{1}{2} + \frac{1}{2} \cos(\phi) \right), \tag{1.12}$$

where $\phi = \phi_1 - \phi_2$. To tune the performance of the MZI, we can have different waveguide lengths, so that a different form of Eqs. 1.11, 1.12 provides intuitive understanding

$$|E_3|^2 = |E_1|^2 \left(\frac{1}{2} - \frac{1}{2} \cos(\beta \Delta L) \right), \tag{1.13}$$

$$|E_4|^2 = |E_1|^2 \left(\frac{1}{2} + \frac{1}{2} \cos(\beta \Delta L) \right). \tag{1.14}$$

where $n = n_1 = n_2$ and $\Delta L = L_1 - L_2$. This form represents the interference relationship we can easily measure, which is helpful to our work in determining the manufacturing quality of waveguides and filtering in general.

1.4.6 Phase Shifters

The intuitive concept of a phase shifter is quite devishly simple, yet can amount to many practical challenges. We take the equation for light propagation along a waveguide from equation Eq. 1.3 and give it a slightly different look

$$\frac{E_{out}}{E_{in}} = Ae^{i\theta} \tag{1.15}$$

where we lump the propagation loss into the term A and describe the *phase-shift* along the length as θ. The β we dropped from Eq. 1.3 is referred to as the propagation constant, with the functional form of $\beta = 2\pi n_{eff}/\lambda$. The key realization here, is not only can we allow the phase to freely rotate along the propagation length as mentioned in the previous section, but if we can find a way to change the material's characteristics, in particular the index of refraction, then we can maniuplate the phase via other means

than length. This gives us the form of the phase angle θ as

$$\theta = \frac{2\pi n_{eff} L}{\lambda} \tag{1.16}$$

which can also describe the phase shift with respect to effective index

$$\Delta\theta = \frac{2\pi \Delta n_{eff} L}{\lambda}. \tag{1.17}$$

It is important to grasp that the length is something we can change in the design stage, but after fabrication, we are unable to change the length of our waveguides. For example, with the MZI, we would pre-determine the length difference so as to achieve a particular interference pattern or operation point. With phase shifters, we impart the change *after* fabrication via an exterior stimulus – most often through power dissipation, carrier depletion or carrier injection via a voltage bias across two terminals, but this can also be imparted by fluid, vibration, stress, plasmons, among other sources [14, 111, 113, 119]. These technologies are the backbone of telecommuncations and information processing application, in addition to many sensing applications. Phase shifting holds the key to large scale implementation of silicon photonics into a variety of different systems, an authentic attestation to the technologies versatility.

1.4.7 Modulators

Traditional modulators impose an electronic signal on top of an optical carrier signal. For example, telecommunications implements modulators in conjunction with fiber optics

to convey data over large distances, leveraging the optical fiber's low-loss propagation. Silicon photonic modulators are no different, at least in principle. From a material perspective, silicon has some exciting properties to aid modulation. For instance, the thermo-optic coefficient of silicon's refractive index is $1.8 \cdot 10^{-4} K^{-1}$ [58]. For example, thermo-optic modulators can be created with MZIs where the phase relationship in Eqs. 1.11, 1.12 is controlled with varying index $\phi = \beta L(n_1(T) - n_2(T))$. Unfortunately, the thermo-optic effect in silicon is slow (with rise times on the order of microseconds). It is susceptible to instability via thermal crosstalk, making it a poor candidate for many modulation purposes. Instead, circuits use the thermo-optic effect to perform more "static" operations like setting or reconfiguring conditions (i.e., overcoming manufacturing imperfections, low/no loss tuning) [19].

Electro-optic modulators in silicon also use the refractive index modulation approach. However, this is enabled by the free-carrier plasma dispersion effect. The free-carrier plasma dispersion effect maps the relationship of free carriers within silicon to changes in refractive index and absorption [82]. As the concentration of free carriers increases, the change in effective index increases linearly, while the absorption increases logarithmically. The concentration can be changed via an external voltage bias at speeds up to 10s - 100s of GHz. This effect is achieved by varying the concentration where the optical mode is guided so that the variation in carrier concentration results in the maximum variation in optical signal.

In an MZI, we can create this modulation effect by the addition of a lateral positive-intrinsic-negative (PIN) diode across the waveguide in the arms between couplers – Fig. 1.7 (d,g). By applying a bias, we can control the phase in the two arms to modulate

the output signal. MZI modulators in silicon can suffer from a hefty $V_\pi L_\pi$ (the voltage and length required to get a π phase shift or to switch between a 0 and 1 in data terms) on the order of $V\cdot$ cm, this modulation can be seen in Fig. 1.7 (h) [94]. However, MZI modulators have low wavelength dependence, high bandwidth, and generally low drive voltages [7, 9]

MRRs offer a slightly different approach to silicon photonic modulation. If we apply the PIN diode radially to an MRR, we can also create a modulator – Fig. 1.7 (d,e). This modulator has a highly compact footprint, is suited to wavelength division multiplexing (WDM) tasks, and has low modulation energy [29, 139, 140]. It is worth noting that a large difference between MZMs and MRRs are that MRRs require less V_π to switch between "0" and "1," however, this is counterbalanced by the fact that V_π is generally much larger in MRRs – Fig. 1.7 (f). In general, the on/off voltages can be quite similar. A detailed resource for electro-optic silicon modulators is available in Ref. [94].

1.4.8 Photodetectors

In contrast to manipulating optical signals, photodetectors intend to *detect* these signals. Photodetectors operate by absorbing photons and converting them into free carriers, which are swept out of the active region by an electric field and measured as current. In silicon photonics processes with intended operation in the c-band, we are typically limited to photodetectors with vertical PIN diodes where the P is p-type silicon, I is intrinsic germanium, and N is n-type germanium. Because silicon is a bad absorber at the telecom wavelengths (conversely, because silicon can guide light efficiently at the

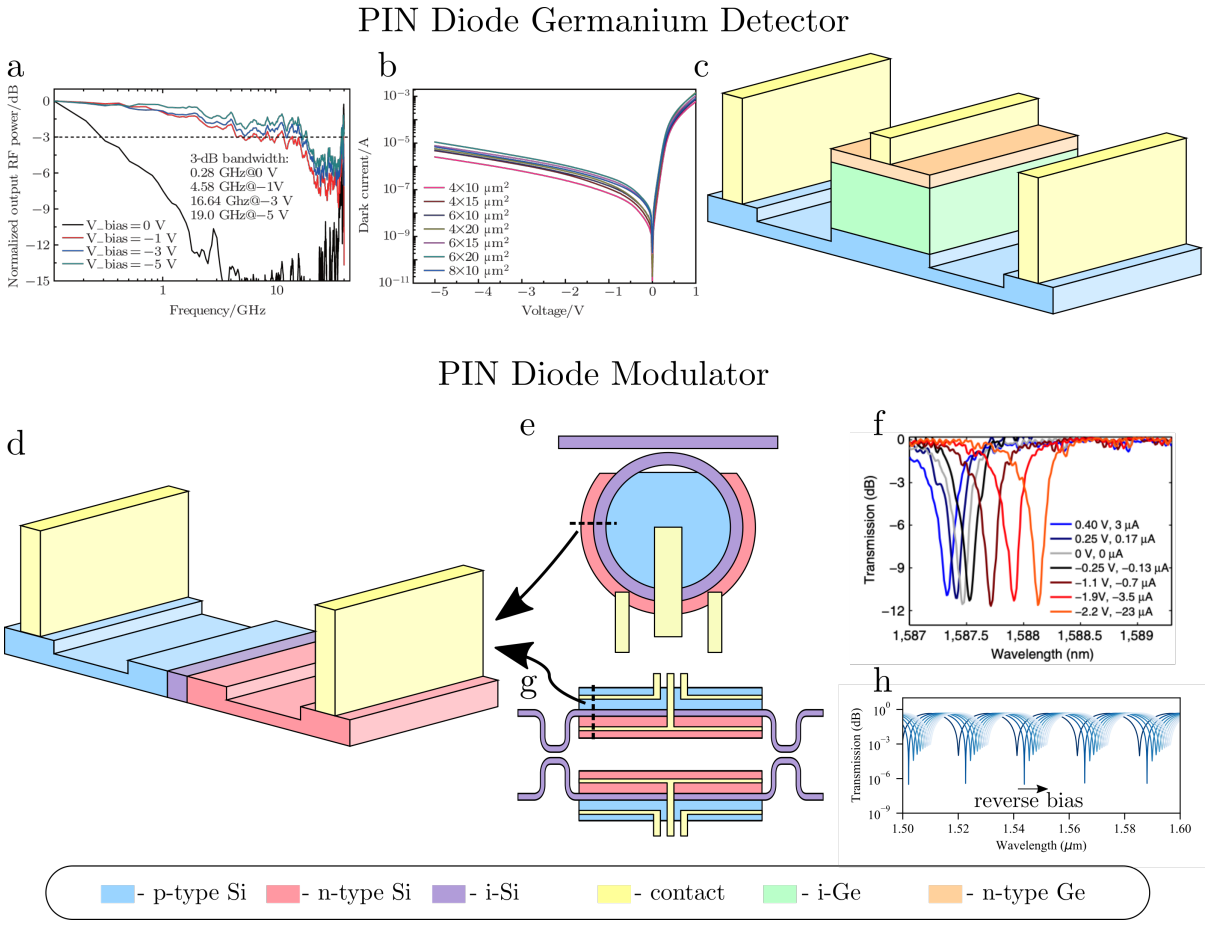

Figure 1.7: a) Demonstrating a photoresponse as a function of frequency [18]. b) Demonstrating the effect of dimensions of the photodetector on the electrical characteristics [18]. c) A diagram of a typical germanium-on-silicon photodetector. d) A diagram of a typical PIN diode modulator in silicon. e) A top-down perspective of a typical PIN MRR modulator. g) A top-down perspective of a typical PIN MZM. f) E/O response of an MRR modulator [120]. h) E/O response of an MZM modulator.

telecom wavelengths), it is not a good material for detecting photons. Therefore, we have germanium because the absorption coefficient is high for the telecom band and it is available in the standard CMOS foundry process kit.

Germanium is generally epitaxially grown on silicon, which can be leveraged for photodetectors in the near-infrared or for optical waveguides in the mid-infrared. As germanium is a semiconductor like silicon, it can accept donors to alter the electrical or optical characteristics. For photodetectors, a lateral or vertical (Fig. 1.7 (c)) PIN junction can

be formed by locally doping the germanium (vertical requires p-type doping of silicon) with n-type or p-type dopants. Characteristic performance traits of photodetectors are the responsitivity, dark current, and the 3 dB bandwidth. Responsivity refers to the conversion efficiency of incident optical power to electrical current with units A/W. Generally, germanium on silicon photodetectors has a theoretical maximum responsivity of 1.25 A/W. The dark current is usually defined as the leakage current at 1 V reverse bias. These photodetectors generally have small dark currents (Fig. 1.7 (b)) between 1 - 1000 nA. 3 *dB* bandwidth is the frequency at which the photodetector fails to maintain 50% (or 3 dB) of the signal for a given power and contact bias – Fig. 1.7 (a). There is more variation per design for germanium on silicon detectors, where the bandwidth ranges from 1 - 100 GHz depending on method and fabrication quality. A review of germanium on silicon photodetectors is available in Ref. [63].

1.5 Neuromorphic Photonics

If we parse the phrase "neural networks," we see that *networks of neurons* is perhaps the most straightforward, suitable explanation. As the introduction states, we can structurally separate these two components into the network and the neurons. The network comprises a trainable (configurable, programmable, etc.) linear, many-to-one, or many-to-many connection scheme [90]. Each neuron is a relatively static (i.e. perforones a single operation), non-linear and one-to-one connection scheme [115]. In neuromorphic photonics, we see linearity naturally map to optical propagation, and nonlinearity naturally maps the difference between the optical-electrical conversion (detection stage) and

the electrical-optical conversion (modulation stage).

1.5.1 Linear Operators

This section summarizes the current state of the art in linear photonic operators. We can also refer to this as the "linear neuron" or a neuron with linear activation. All of these names are equivalent to a linear operator. We explore the current linear photonic operators in literature. Primarily, there are two operation mechanisms for linear operators in this platform, namely wavelength incoherent and coherent.

1.5.1.1 Wavelength Incoherent Linear Operators

Broadcast & Weight The broadcast & weight protocol represents the first attempt to integrate neuromorphic computing onto the silicon photonic platform [117]. Broadcast & weight is an idea that leverages the dispersive guiding capacity of silicon waveguides. Simply put, an incoming optical signal is multiplexed (broadcast) into many wavelengths using a wavelength division multiplexer (WDM), and then each wavelength's amplitude is weighted by a filter (MRR) to interact with downstream nodes or detectors [117]. This circuit can be configured to perform both feed-forward (i.e., with many linked layers) or in recurrent architectures (Fig. 1.8 (a)).

Operationally, the weighting mechanism consists of a tunable bank of filters (MRRs). The incoming signal is multiplexed on N wavelengths and is injected into the MRR weight bank (Fig. 1.8 (a)). Each unique wavelength is selected from the signal via the corresponding tunable filter and given weight. This weight is imparted by tuning/de-tuning the MRR concerning the resonance, wherein the resonance bandwidth and the linewidth

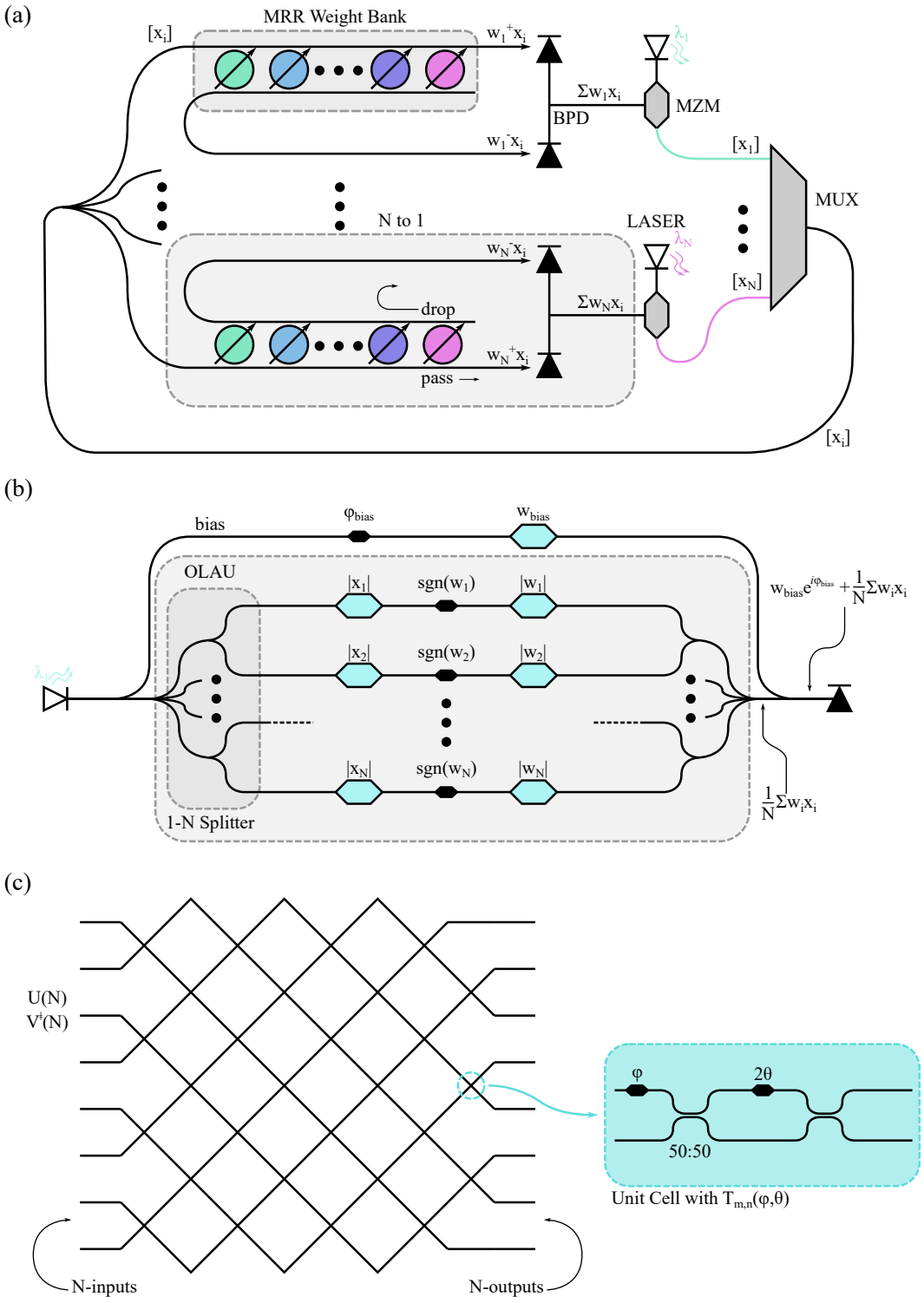

Figure 1.8: (a) The Broadcast & Weight protocol in a recurrent schematic [117]. (b) The coherent optical linear neuron (COLN) schematic [76]. (c) The universal multiport interferometer mesh schematic, where the unit cell is described as an inset [20, 93]. Adapted from Ref. [124]

of the corresponding wavelength, λ_j, are similar. Herein lies a practical difficulty with this method, as it requires a small radius (i.e., large FSR), and high Q resonators with distinguishable resonances. The scalability of these networks for a large N can be a true concern unless improved by filtering with multiple tunable resonators to mitigate wavelength cross talk, which may increase the signal integrity at the cost of increased control components [118]. However, in this way, the weight of filter j on the i^{th} signal, $w_{j,i}^{+}$, will pass by the filter and $w_{j,i}^{-}$ will be routed to the drop, such that the overall weight is denoted by the difference $w_{j,i}^{+} - w_{j,i}^{-}$. This allows the weights to be fully tuned from -1 to 1 [116]. The separately filtered signals from $0, ..., N$ occupy space within the same pass and drop waveguides concerning the MRR weight bank. In a general neural network sense, this represents the weight matrix operation or the linear synapse.

Following this weighting procedure, the pass/drop signals are detected by a balanced photodetector (BPD) pair. The BPD is designed to detect small signal changes such that equal signals sent to the independent detectors will register no signal. A signal is detected by differentiating the amplitude of the signals sent to each detector. Along with being a complementary detector to the MRR weight bank, enabling the weight range of -1 to 1, the photodetector performs another essential task. In this architecture, the detector sums the signals with the O/E conversion. In addition to representing the necessary weight and sum for a neural network, it is essential to note that herein lies the entire MAC operation for this architecture; the multiplying stage is represented by weighting and accumulates by O/E conversion. Due to the wavelength incoherence, the independent wavelengths, $\lambda_{0,...,N}$, do not interfere with one another and cannot be summed. The detector acts as

the summing operator for the wavelengths visible by the material bandgap, so for Ge-on-Si photodetectors, the visibility window extends roughly from the near-infrared up to 1600 nm with some variation possible due to specific material composition [25]. From here, the signal can be passed into an amplifier or onto a modulator (as in Fig. 1.8 (a)) to connect to the next node. A fully connected broadcast & weight circuit of size N to M will require N lasers and M modulators, $N \times M$ tunable weights and $2M$ photodetectors (or M BPDs) [117, 124].

1.5.1.2 Wavelength Coherent Linear Operators

Coherent Optical Linear Neuron Fig. 1.8 (b) demonstrates a coherent optical linear neuron (COLN) proposed in Ref. [76]. Fundamentally, this architecture performs MAC operations via a single wavelength, λ_1, or is simply wavelength coherent. A signal is passed through a 1 to N splitter, where N modulators impart the $x_{0,...,N}$ input signals onto each waveguide. From here, a phase shifter, $sgn(w_i)$, imparts a $0/\pi$ shift for the direction (positive/negative) of the weight, and a second modulator imparts the magnitude of the importance, $|w_i|$. Each of the N signals then has the form $\text{sign}(w_i) \cdot |w_i| \cdot x_i/N$, here representing the multiplying stage of the operation. All N signals are then combined, which due to the coherent wavelength scheme, represents the accumulated stage of operation. A bias is recombined with the entire signal, which is crucial to providing a reference level and converting the sign of the sum from phase to amplitude.

The COLN architecture does not require O/E conversion for its accumulation stage, which allows flexibility for optical or electrical nonlinearities - however, optical nonlinearities have a poor history for high-efficiency conversion on-chip, and likely some O/E

will be needed in practical demonstrations [124]. This method requires only one laser. However, it requires N times more laser power equally, so this is not an efficiency gain. MZMs add insertion loss to the system, particularly within an integrated injection/depletion mode modulation framework. Insertion losses will require higher power with the additional burden of decreased scalability, i.e., the ability of the initial input to modulate downstream nodes. Of course, this can be remedied by employing slower photonic devices (using thermo-optic effects) in favor of lower insertion loss.

Universal Multiport Interferometer In a largely impactful work, another coherent linear neuron was introduced to neuromorphic photonics as the universal multiport interferometer (UMIAnotherarn from quantum optics that any arbitrary unitary operator can be crafted with simple optical components (i.e., phase shifters and beam splitters) [93]. Each unit cell (Fig. 1.8 (c) inset) is comprised of two-phase shifters and two 50:50 beam splitters. These rotation cells can be represented by a transfer matrix, $T_{m,n}(\phi, \theta)$. We can cascade these unit cells in specific configurations [20, 93] to form a larger matrix represented physically by the device Fig. 1.8 (c). The key behind this method is to create a vector-by-matrix multiplication (VMM) operator by singular value decomposition such that a matrix, M, is denoted by $M = U\Sigma V^{\dagger}$. U represents a signal routing or mixing stage. A diagonal matrix Σ (physically represented by another OIU or a column of optical attenuators) represents the weight application (or the multiplying stage). Finally, the matrix $V^{\dagger}$ can represent the wavelength coherent interference operation (or the accumulation stage). With miniaturization in mind, we can apply this concept to the integrated photonic domain to create a linear operator in a more *classical* sense.

While the approach *is* ultimately applicable for neuromorphic photonics, the UMI has shortcomings in this application that are hard to overlook. Primarily, these issues are the configuration/control complexity (i.e., $\sim N^2$ phase shifters) and the physical size ($\sim N^2$ fan-in scaling compared to linear N for both Broadcast & Weight and COLN). In addition, these factors increase the power consumption of the physical circuit because of the individual device needs. These complications create a space for improvement in many different architectures and highlight the previous architectures' utility.

1.5.2 Non-Linear Operators

Non-linear operators are essential for neural network behavior since linear activation functions serve no purpose outside of increasing the network size. From an optic perspective, the non-linear operation can be hard to achieve simply because photons generally interact linearly. There have long been dreams of purely optical computers, which stand like glass monoliths, performing complicated tasks at ultra-high efficiency and speed. Unfortunately, these dreams have *generally* remained dreams [74].

Primarily, there are a few necessary characteristics for any non-linear operator (or, in the context of a neural network, a full processing node including the weighting stage) [74, 90]:

- **Cascadability**. The output of any operator must be capable of driving downstream operators, whereas polarization must be maintained for purely optical signals wavelength.

- **Fan-in/Fan-out**. The operator's output must be capable of driving at least two

downstream operators, i.e., gain greater than 2. Additionally, it can represent a many-to-one structure of connection.

- **Signal Restoration or Cleanup**. The operator must clean the input signal into logic-level restoration by embedded thresholding, resetting, and restoring pulse quality.

- **Input/Output Integrity**. There must be distinct input and output ports, so reflections or overlaps cannot create false outputs.

There are also requirements for the absence of high-precision biasing and independence of transmission loss.

This section reviews current approaches for silicon photonic neurons, representing the second constituent part of a neural network architecture. The following concepts are a survey of the current state of the art, representing promising and practical approaches.

1.5.2.1 Micro-Ring Resonator Modulator Neuron

The silicon modulator-based neuron, described and tested in Ref. [115], represents the first completely monolithic integrated photonic neuron with nonlinearity, scalability, and fan-in/out. Other classes of neurons proceed with this with architectural similarity (i.e., the laser spiking neuron [91]). Still, this neuron can be fabricated in any standard CMOS process without complex post-processing or laser attachment. However, the operating principles and requirements are pretty similar - the key difference is the nonlinearity for the modulator acts on top of a continuous-wave laser instead of switching the laser between "on" and "off."

The MRR modulator neuron is directly applicable in the broadcast & weight architecture context due to the BPD scheme. Fig. 1.9 (a) details the design considerations for the MRR modulator. The BPD acts as the accumulate (summation) stage in the neural network scheme of broadcast & weight, but in addition, it acts as a small signal detector for sensitive photodetection. For any small signal difference in the two detectors (i.e., from optical powers z^+, z^- such that the detected signal is $z^+ - z^-$), a signal current is generated, whereas when the signals are equal, no signal is generated. Again, this allows for filter-based weights, such as broadcast & weight. With the signal converted to the electronic domain, a potential bias current, I_{bias}, is added to the call and passed through to a diode. This diode represents the electronic feature of the MRR modulator. As current is applied, the effective capacitance of the diode is changed or "modulated." As this phenomenon occurs, the continuous wave optical signal *in* passes through the time-varyings a non-linear function, $f(z)$. This non-linear function maps the input power, namely $z^+ - z^-$, to the energy passed through the ring. Such functionality is not generally available in all-optical domains.

A device like this can act highly non-linear and, in addition, is reconfigurable. The reconfigurability comes from the fact that an MRR's spectral response is peaked-Lorenztian. Therefore, we can tune the laser (or thermally tune the resonance of the MRR) such that the wavelength acts as a starting point for various non-linear functions. In Ref. [115], optical-optical sigmoids, rectified linear units (ReLU), and radial basis functions (RBF) were demonstrated.

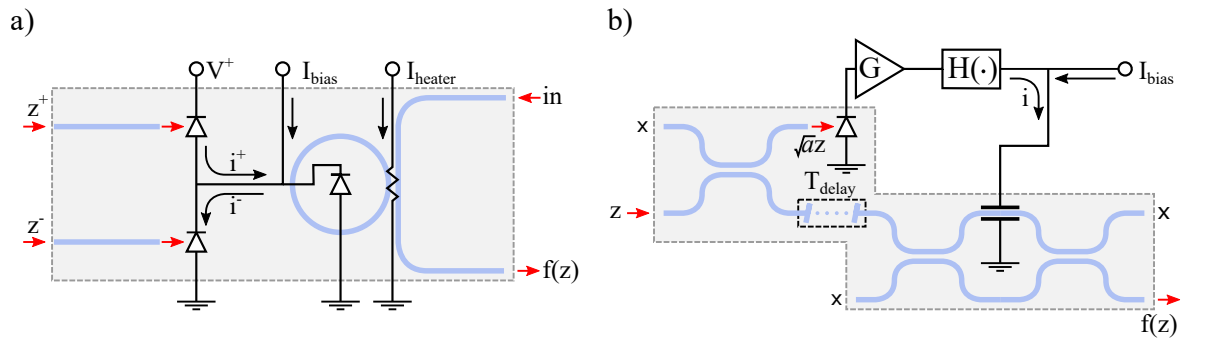

Figure 1.9: a) The MRR Modulator Neuron proposed in Ref. [115]. Two optical signals, z^+ and z^-, are detected by a balanced pair of photodetectors, which are specifically useful for detecting small signal differences. From an electronic perspective, the current detected is fed to an MRR modulator, a PIN diode. In addition, there is a I_{bias} which can inject extra current into the circuit (at the cost of latency). Acting on the diode in reverse bias, the current can affect the diode capacitance's size, modulating the optical power passing through the MRR (optical signal in). The modulation will impose a function onto in of the shape $f(z)$, which depends on the optical inputs. The MRR can be set by I_{heater} to perform various activation functions, like the sigmoid, ReLU, and RBF. b) The MZM Neuron proposed in Ref. [135]. An optical signal, z, is injected into an optical splitter such that a small fraction of light, $\sqrt{a}z$, is removed from the circuit and detected, while the remainder of the signal goes into a delay line of time, T. The delay preserves the majority signal while the detected light is fed into electronic gain, G, and processed with a signal conditioner, H, where a non-linear function is applied. The current is then fed to an optical phase shifter placed in an MZM. This MZM is optically excited by the light exiting the delay, and ultimately, the light leaves the MZM. The detected light at the output represents a non-linear activation function of the form $f(z)$.

1.5.2.2 Mach-Zehnder Modulator Neuron

Another fundamental photonic circuit is the MZI, as discussed above,e. In Ref. [135] a neuron is described that leverages the MZI and is demonstrated in [33]. This neuron is detailed in Fig. 1.9 (b). A small optical tap is placed on an incoming signal with power z, so only $\sqrt{a}$ arrives at the photodiode, where a describes the coupling coefficient of the coupler. This electrical signal is passed through a signal gain stage (G), represented by a signal amplifier, then a voltage signal conditioner transfer of $H()$. After a current bias injection, finally, the signal reaches the MZI in the form of a phase shift. Meanwhile, the rest of the input signal, namely $\sqrt{1-a}z$, is sent into the MZI through a delay line,

allowing the electrical signal to catch up. Therefore, the output, $f(z)$, at the exit stage of the MZI is non-linear proportional to the input z.

This non-linear function is a neat approach that utilizes an inventive combination of optics and electronics. However, the overall architecture falls out of favor with the broad strokes goal of neuromorphic photonics. Firstly, requiring an electrical signal gain and a voltage signal conditioner increase the energy needed per operation and slows the speed due to multiple transfer stages. Secondly, utilizing a simple phase shifter on an MZI arm requires an electro-optic device with a considerable length (> 500 μm) or a thermo-optic device with slow response speed (> 1 μs). Finally, this approach loses out on the spirit of integrated photonics by requiring multiple off-chip electronic stages, which, if integrated on-chip in the future, would increase the footprint of the neuron and decrease its scalability. This approach does have the potential to be reshaped into an improved non-linear operator by crafting the electronic stage to act in a more "receiver-less" paradigm.

1.5.2.3 Amplifier-Free, Bias-Free Optical Switch

One class of optical non-linearity is currently gaining momentum in the literature that seemingly can meet requirements outlined by Sec. 1.5.2; however, generally sits outside of the neuromorphic regime [84]. The approach is crafted by combining (a) photodetector(s) with a modulator to create a coupled optoelectronic effect or an O/E/O transfer. Strictly from an electronic perspective, Fig. 1.10 (a), we can think of a photodetector ideally as a current source with some parasitic effects (i.e., shunt resistance, series resistance, load capacitance, dark current, etc.). We can approximate the photodetector circuit as a

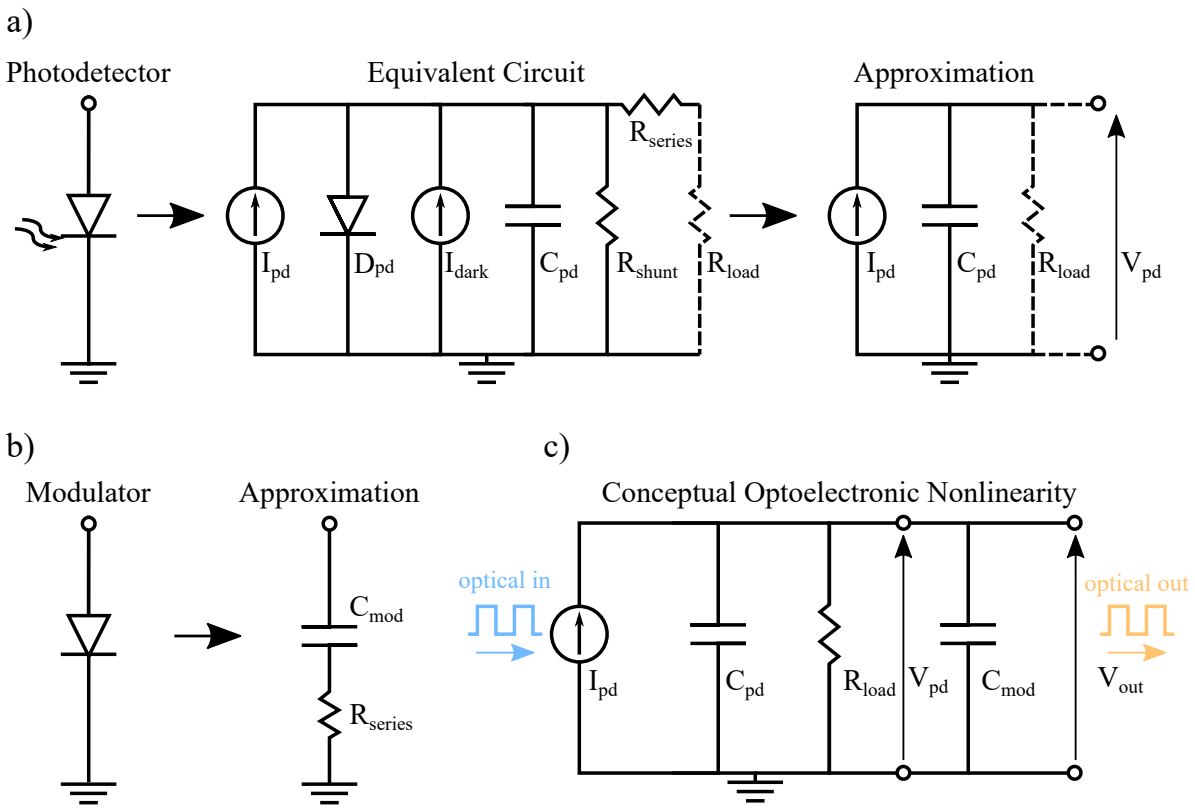

Figure 1.10: a) Here we detail the approach to simplifying the electronic circuit model of a photodetector. If we look at the full picture, the photodiode consists of the generated photocurrent, I_{pd}, the diode, D_{pd}, the dark current, I_{dark}, the internal capacitance, C_{pd}, the shunt resistance, R_{shunt}, the series resistance, R_{series}, and the load resistance R_{load}. In general, we can safely ignore all of these except the generated photocurrent, the capacitance, and the load resistance. b) A similar approximation of an integrated modulator's electronic perspective – we see that it reduces down to a series resistor, R_{series}, and a capacitor, C_{mod}. c) If we connect these, we see that an O/E/O switch can be modeled simply by the capacitors and resistors of the modulator and photodiode [84].

capacitor in parallel with a current source and load resistor or simply an output voltage dependent on the input optical power, V_{pd}. Similarly, we can think of a modulator as a diode, simplifying it further to a capacitor in reverse bias (and a series resistor for completeness). This circuit represents the simplest optoelectronic non-linearity, where the photodetector drives a modulator, and the electronics exist solely as an agent of the optics or in an amplifier-free, bias-free configuration.

Fig. 1.10 (a) depicts the equivalent electronic circuit for a simple PIN photodiode. The photodiode can be considered a "photon counter," wherein each photon is converted to an electron through the photoelectric effect by some efficiency η_{pd}. Quite simply, then, we can determine the photodetector voltage, V_{pd} as [84]

$$V_{pd} = \eta_{pd} R_{load} P_{optical}. \tag{1.18}$$

In simple terms, the modulator can be thought of primarily as a capacitor in series with a resistor (Fig. 1.10 (b)). We can combine these two devices to create an O/E/O conversion, Fig 1.10 (c), where the optical input affects the modulator device to transduce the signal onto optical out. It is vital to understand that the signals are all time-varying, and we can use this to derive the bandwidth response of such a device. The RC frequency response is given as [84]

$$f_{RC} = \frac{1}{2\pi R_{load}(C_{pd} + C_{mod})}. \tag{1.19}$$

The RC response can be added to the carrier transient characteristics to give a device responsivity of

$$f_{total} = \sqrt{\frac{1}{1/f_{RC}^2 + 1/f_{transient}^2}}. \tag{1.20}$$

1.6 Dissertation Organization

With the background and motivation covered, the remainder of this dissertation is organized into chapters representing the distinctly different projects pursued. In Chapter 2, we explore the 2D skin depth engineering concept. This early project resulted in a new

directional coupler and waveguide routing platform, targetted as a crosstalk suppressive and bendless paradigm. In , we delve into work concerning the design and test of a high extinction ratio microring modulator. This device has specific requirements for its primary applications in artificial intelligence and quantum information, marking it as unique against traditional high-speed WDM systems. In Chapter 4, we explore a process for thermally isolating optical phase shifters. In Chapter 5, we reimagine the COLN with microring modulators. This allows us to introduce a new architecture capable of using coherent and incoherent attributes for highly scalable neural network applications. In Chapter 6, we conclude by providing an overview of the packaging efforts undertaken in the course of the work presented and the process development needed to implement the packaging.

Chapter 2

2D Skin Depth Engineering

2.1 Evanescent Waves in Silicon Photonics

Evanescent waves in silicon photonic waveguides tend to cause parasitic optical crosstalk. In traditional photonic circuits, design strategies must consider minimum separation distances between closely spaced waveguides to prevent unwanted coupling [19]. This problem is inhibiting many photonic circuits due to cost and size constraints. Many efforts have been made to overcome these issues, battling size constraints by employing inverse design [88, 103] or implementing metamaterials to increase performance [53, 65, 110].

Recent work introduced a new, metamaterial paradigm for waveguiding that fundamentally suppresses coupling between waveguides [54]. In this subwavelength approach, multi-layer cladding is placed in-plane and parallel with the waveguide, decreasing the skin depth of the fundamental transverse-electric (TE) mode's evanescent field. The concept is called *extreme skin depth engineering*, or *e-skid*. *E-skid* has been employed as cross-talk suppression [54, 72], and for high performance polarization splitting [16, 138].

The *e-skid* features are created in the same processing step as the waveguide itself, allowing this to be an innate no-cost addition to any design. Adding these features can reduce the crosstalk between waveguides by more than three orders of magnitude, dramatically reducing the photonic design footprint. [54].

Here we expand on this work by using *e-skid* engineering in *two directions*, within the same plane as the waveguide. Using both a parallel (as in [54]) and perpendicular *e-skid* cladding, we engineer the coupling between waveguides throughout a photonic circuit. Specifically, the perpendicular *e-skid* cladding (with features orthogonal to the waveguide) can dramatically increase the evanescent tail of the mode (or equivalently decrease its decay constant - as shown in Fig 2.1 (c)). This then enables arbitrary enhancement in coupling between nearby waveguides anywhere in the circuit. While parallel *e-skid* cladding can suppress coupling in routing the rest of the circuit. Here we explore this with the design of a 2D *e-skid* directional coupler with a large gap ($\geq 1.4\mu$m) between the waveguides and large operational bandwidth ($\geq$ 40 nm). We also demonstrate this 2D *e-skid* directional couplers in a complementary metal-oxide semiconductor (CMOS) photonic platform, thereby affirming the manufacturability of *e-skid* components and integration with foundry offerings.

Ultimately, this work is motivated by the idea that we can employ two-dimensional *e-skid* techniques to ensure low crosstalk in the routing phase (with parallel *e-skid*) and, without the use of bends, efficient coupling between waveguides with larger-than-normal gaps (with perpendicular *e-skid*) - fully leveraging both directions of *e-skid*. Consequently, two-dimensional *e-skid* allows for dense integration within the constraints of CMOS manufacturing.

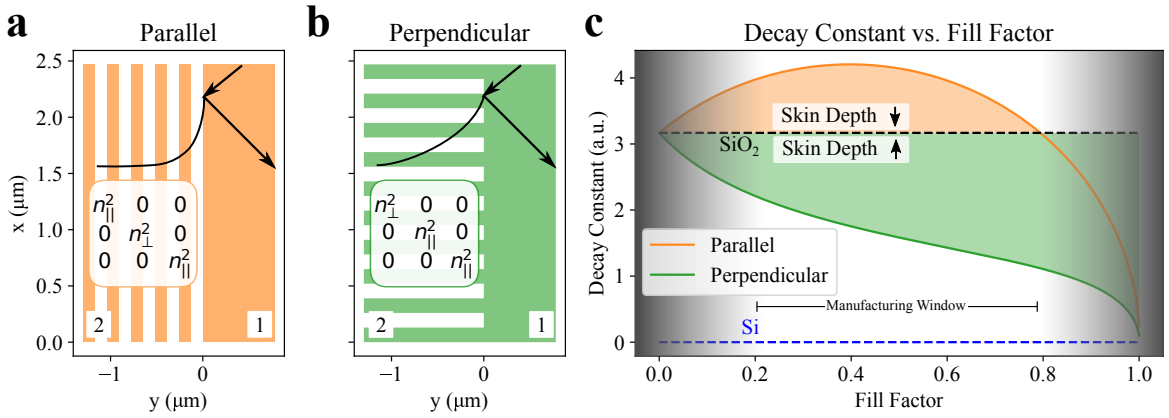

Figure 2.1: (a) Here, we demonstrate the parallel-oriented subwavelength, multi-layer cladding. The anisotropic permittivity tensor is displayed over the cladding, which follows the Rytov relations for each direction. (b) The perpendicular cladding essentially swaps the xx and yy components of the permittivity tensor in (a). The incident wave is reflected (shown by the two arrows in medium one), and the evanescent wave is strongly decaying in medium two in (a) and weakly decaying in (b). (c) A plot of Eq. 2.3 for the two different cladding strategies. We see that decay increases over SiO_2 for most of the fill factors of the parallel case. In contrast, we see a variable decrease in decay by almost the whole scale between the two materials in the perpendicular cladding. The likely "manufacturing window" over which these features can be fabricated in a CMOS silicon photonics foundry is indicated.

2.2 Theory

2.2.1 *E-skid* in Two Directions

Consider two media with an index of refraction n_1, n_2. When an incoming wave from n_1 meets the boundary at n_2 and the angle is greater than the critical angle, $\theta_i < \theta_c = \sin^{-1}(n_2/n_1)$, an evanescent wave is formed in the second medium. This wave does not carry power across the boundary; it exponentially decays into the second medium [46]. *E-skid* allows us to tune the decay constant of this evanescent wave by introducing sub-wavelength, periodic structures that transform a wave's (specifically, a polarized wave's) momentum [51,52]. These features change the second medium from an isotropic material to an anisotropic metamaterial. The anisotropy here refers to the permittivity values of

the material's dielectric tensor (where we assume that the permittivity can be defined $\epsilon_r = n^2$). For deep subwavelength features, these component values are defined by the Rytov relations [15, 99]:

$$n_{\parallel}^2 = n_1^2\rho + n_2^2(1-\rho), \tag{2.1}$$

$$n_{\perp}^{-2} = n_1^{-2}\rho + n_2^{-2}(1-\rho), \tag{2.2}$$

where n_1, n_2 are the indices of the first and second medium, respectively, and ρ is the fill factor. The parallel component, $n_{\parallel}$, is defined in the direction parallel to the periodic structure's orientation, and the perpendicular component, $n_{\perp}$, is oriented perpendicular to the periodic structure. These relations demonstrate how the second material transforms from isotropic to an anisotropic metamaterial for deep-subwavelength features. However, any subwavelength structures will exhibit anisotropy, albeit without these neat relations. The key result of the *e-skid* derivation leverages this anisotropy for the evanescent wave, which is characterized by the decay constant, β:

$$\beta(\theta_i;\rho) = \frac{1}{\delta(\theta_i)} = k_0\frac{n_{2x}(\rho)}{n_{2y}(\rho)}\sqrt{n_1^2 sin^2(\theta_i) - n_{2y}(\rho)^2}. \tag{2.3}$$

where k_0 is the wavevector and θ_i is the angle of the incident wave to the boundary (we assume paraxial $\theta_i \approx \pi/2$) [51]. The decay constant is now subject to a degree of variable tunability (ρ), allowing for control of the evanescent wave [51, 54].

In Fig. 2.1 (a), we show the dielectric tensor for the *e-skid* structure, where the periodicity of the subwavelength features is parallel to the boundary ($y = 0$). In this orientation, the diagonal components of the second material become $[n_{2x}^2, n_{2y}^2, n_{2z}^2] =$

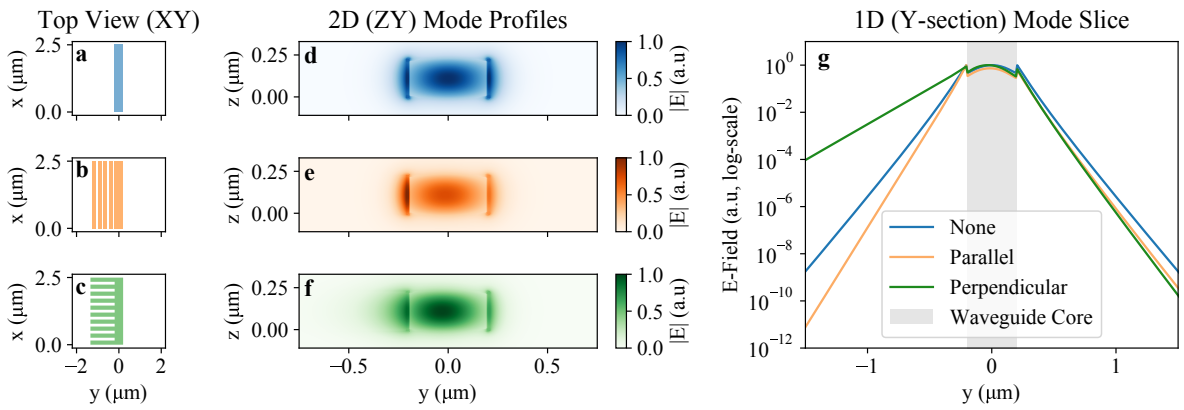

Figure 2.2: A comparison of the different cladding strategies discussed, noting that they have a common cladding on the right-hand side (normal isotropic SiO_2). (a,b,c) Here we have the top-down perspective of the waveguides, which shows the cladding for the none, parallel and perpendicular structures on the left-hand side, respectively. (d,e,f) The ZY plane cross-section mode profiles correspond to the cladding diagrams from (a,b,c), where the width of the waveguides is 400 nm and the height is 220 nm. (g) A center slice through each mode profile demonstrates, on a log-linear scale, the amount of control we can impose on the evanescent wave with these structures. Simulations were done with an anisotropic material following Rytov relations (Eqs. 2.1,2.2), where the core waveguide width was 400 nm, and the fill factor was 0.6 for both orientations.

$[n_{\parallel}^2, n_{\perp}^2, n_{\parallel}^2]$ in accordance with the Rytov relations (Eqs. 2.1, 2.2). This structure will increase the decay constant of the evanescent wave, thereby decreasing the skin depth [54]. Without loss of generality, we recognize that we can rotate the optical axis by rotating the subwavelength features and realize *e-skid* in a *second* direction. Due to the direction dependency outlined by the Rytov relations, when we rotate the periodicity of the features, we effectively swap the xx and yy components of the dielectric tensor of the parallel cladding such that we now see $[n_{2x}^2, n_{2y}^2, n_{2z}^2] = [n_{\perp}^2, n_{\parallel}^2, n_{\parallel}^2]$ (Fig. 2.1 (b)). The values of n_{2x}, n_{2y} in Eq. 2.3 control the decay constant, and by rotating the periodic structure, we can dictate an increase or decrease.

We populated Eq. 2.3 with the new dielectric tensor values outlined in Fig. 2.1 (a,b) such that we show in Fig. 2.1 (c) the full range of decay constant tunability of *e-skid*

in two directions. Figure 2.1 (c) shows clearly that both decreasing and increasing skin depth can be achieved by the parallel features, however applying this to CMOS photonics manufacturing, we generally omit the higher and lower fill factors due to resolution constraints. [106]. We used a material platform consistent with CMOS photonics in (c), such that material one is silicon (Si) and material two is silicon dioxide (SiO_2). However, this is true for any optical material combination as long as $n_1 > n_2$.

2.2.2 *E-skid* in Two Directions in Waveguides

Optical waveguiding is not fully described by the simple electromagnetic wave-at-a-boundary example above. While it lends intuition, we must find the electromagnetic mode of the entire structure to get a clear picture of this effect. We used Ansys-Lumerical's finite difference eigenmode (FDE) solver to simulate the TE modes of three specific types of waveguides to demonstrate *e-skid* in two directions [8]. Figure 2.2 (a) shows a top view of a single-mode strip waveguide, where the propagation is in the $\hat{x}$ direction. Next to the strip waveguide, Fig. 2.2 (d) shows a 2D mode-profile of the fundamental TE propagating mode. We introduce the wave suppressing *e-skid* features on one side of the waveguide in Fig. 2.2 (b) and show the corresponding 2D mode-profile in (e). Finally, we introduce the wave enhancing *e-skid* features in Fig. 2.2 (c) and the corresponding 2D mode-profile in (f). We compiled the cross sections of all three modes in Fig. 2.2 (g) to demonstrate the effect of the features on the evanescent wave of the mode. Figure 2.2 (g) demonstrates, with a log-scale in y, that the parallel features suppress the decaying wave outside of the center of the waveguide and are greatly enhanced by the perpendicular

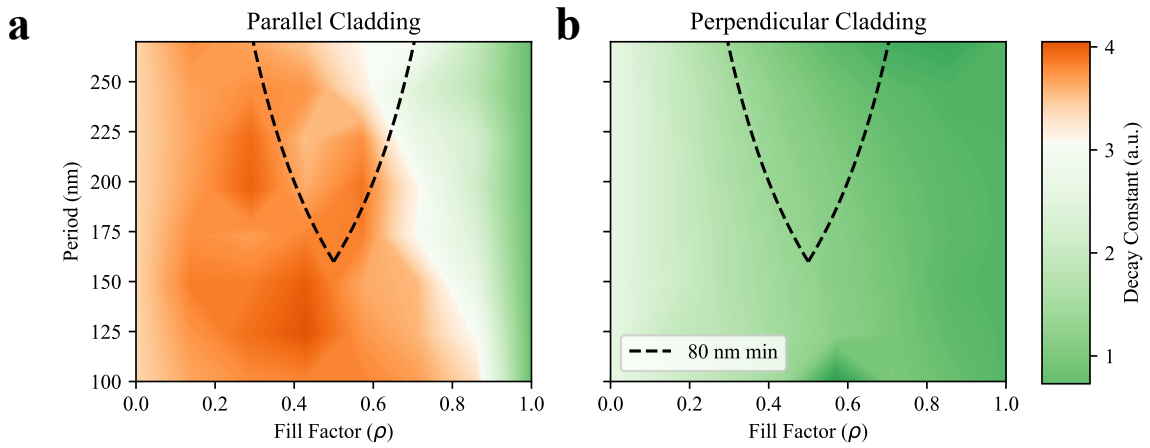

Figure 2.3: Decay constant contour plots extracted via simulation for the (a) parallel and (b) perpendicular cladding features over varying fill factors and periods. We include a trace indicating an 80 nm minimum feature/space cut-off to demonstrate the constraints in subwavelength CMOS photonics. These simulations were run with waveguide widths of 400 nm.

features.

We extend our analysis to include full wave 3D finite-difference time-domain (FDTD) solver with Bloch-periodic boundary conditions [2]. Figure 2.3 displays the decay constant as extracted for varying fill factors and periods of subwavelength features. The decay constant extraction fit for both (a) and (b) had average correlation coefficients of $R^2 = 0.98$. In (a), we show the results of the parallel cladding, and in (b), we show the perpendicular, both of which follow the general shape outlined by the analytical expression in Eq. 2.3 at the paraxial limit, or simply, a guided wave. As the period increases, we notice a shift in the peak behavior towards the lower fill factors, however, this is not significant enough to alter the design parameters. We also note the similarity of Fig. 2.3 and Fig. 2.1 (c), where the parallel features are the only features that increase the decay constant, and the perpendicular features allow for deterministic reduction of the decay constant, thereby enhancing crosstalk. Figure 2.3 stops at 270 nm to allow each simulation to fall below the Bragg limit. We note that the operational window for

subwavelength devices depends on the minimally reliable feature (or hole) size for the manufacturing process. Here, we see at the highest period that only 30-70% fill factors are expected to resolve for an 80 nm minimum.

2.3 2D *E-skid* Directional Coupler Design

We propose and demonstrate a new coupler that leverages *e-skid* in two directions to create coupling in desired regions. In a traditional integrated photonic platform, a directional coupler is created by bending two waveguides close to each other (Fig. 2.4(a)). The waveguides must otherwise be kept far apart in other parts of a circuit in order to avoid unwanted coupling, limiting the circuit density. By using *e-skid* with parallel subwavelength features (Fig. 2.4(b)), that suppress coupling, we overcome this limitation and keep two waveguides within close proximity, with negligible coupling. Furthermore, when the design with *e-skid* needs coupling, we show that the introduction of perpendicular subwavelength features in the coupling region, as are seen in Fig. 2.4(c), will significantly enhance coupling. These features have tunable variables (i) period ($\Lambda_{\perp}$) and (ii) fill factor ($\rho_{\perp}$) which directly tune the amount of coupling experienced. We introduce two-dimensional *e-skid* as a way to leverage the size reduction offered by the parallel features with the addition of the perpendicular features to create practical circuits as seen in Fig. 2.4 (d).

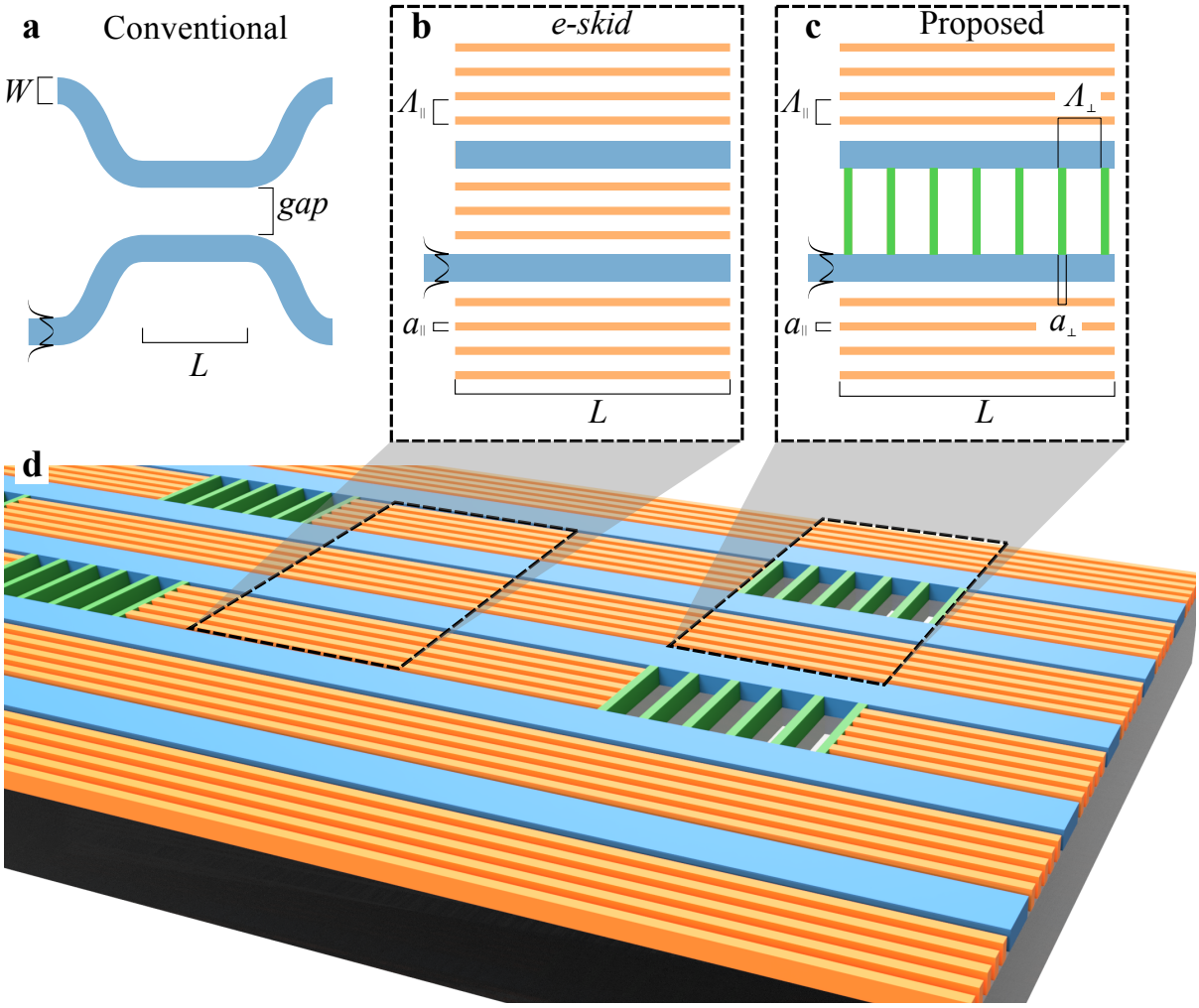

Figure 2.4: (a) A conventional directional coupler with a coupling region characterized by the gap between waveguides and the length of the parallel section. (b) The *e-skid* platform discussed in [54], where the period ($\Lambda_{\parallel}$) and silicon fill ($a_{\parallel}$) characterize the subwavelength features. (c) Our directional coupler leverages the enhancing *e-skid* features in the coupling region, where the features outside the coupling region are the same as (b) and where the period ($\Lambda_{\perp}$) and silicon fill ($a_{\perp}$) characterize the subwavelength features in the coupling region. (d) An example of an integrated photonic circuit implementing two-dimensional *e-skid*. Note the circuit maintains the size reduction of *e-skid* coupled with 2D *e-skid* directional couplers. In the colored version, the colors throughout indicate the photonic waveguides (blue), parallel (orange), and perpendicular (green) *e-skid* features (where all are the same material - namely silicon), and the base is the buried oxide (black). Figure is not drawn to scale.

2.3.1 Coupled Modes for *E-skid*

Even though it carries no power across the boundary, the mode's evanescent wave causes coupling between parallel guides if the overlap between the evanescent waves of supported

modes is large enough [49]. From coupled mode theory [49], we define the power in the bar and cross ports as

$$P_{bar}(L) = P_0 \cos^2(\kappa L) = P_0 \cos^2\left(\frac{\pi}{2}\frac{L}{L_x}\right), \tag{2.4}$$

$$P_{cross}(L) = P_0 \sin^2(\kappa L) = P_0 \sin^2\left(\frac{\pi}{2}\frac{L}{L_x}\right), \tag{2.5}$$

for L as the coupling length, P_0 as the injected power, κ as the coupling coefficient, L_x as the crossover length such that $\kappa = \pi/(2L_x)$, where the bar and cross-refer to the light remaining in the injected waveguide or transitioning to the other waveguide, respectively. The crossover length is defined so that when $L = L_x$, there is a complete power transfer from waveguide one to two.

This approach allows for an intuitive understanding of the device. The crossover length is given as

$$L_x = \frac{\lambda}{2(n_{even,\lambda} - n_{odd,\lambda})}, \tag{2.6}$$

where λ is the free space wavelength and $n_{even,\lambda}, n_{odd,\lambda}$ are the effective indices of the even and odd modes, respectively [19]. The field of the odd mode is antisymmetric across the coupling region and it remains generally unaffected by symmetric features there [42]. However, by introducing features into the coupling region the even mode is affected, thereby enabling dispersion engineering of the directional coupler - specifically, controlling the directional couplers' optical bandwidth. [42]. By crafting L_x, we can dictate how the device performs according to Eqs. 2.4,2.5. Essentially, if we make the slope of L_x as flat as possible over a span of λ, we ensure a useful operating bandwidth

(e.g. a 3 dB coupler) is preserved for that span. Our design is fundamentally different than [42] due to the structural asymmetry, which encourages coupling, and the higher fill factor. These parameters allow us to create a directional coupler with more than an order of magnitude shorter crossover length in comparison, at the penalty of reduced operating bandwidth.

In order to demonstrate the effect of dispersion engineering, we begin by simulating the photonic bandstructure of the directional coupler in a full wave 3D FDTD solver with Bloch-periodic boundary conditions [2]. Figure 2.5 (a,b,c) shows the photonic bandstructure of a traditional, 1D *e-skid* and 2D *e-skid* directional coupler. These directional couplers are fundamentally different from photonic crystals as they are not designed to work in the photonic bandgap, instead, these subwavelength features allow for low loss propagation through the periodic structures below the bandgap [15]. The traditional and *e-skid* couplers exhibit similar bandstructures, but the 2D *e-skid* directional coupler's even mode is approaching the band edge just above 200 THz (1500 nm). Because the even mode is near the band edge, dispersion is increased, which allows for flexibility in tuning the behavior. From the bandstructures, we extract the dispersive effective index of both the fundamental even and odd supermodes (Fig. 2.5 (d,e,f)). The effective indices exhibit a similar characteristic shape to their corresponding bandstructures. The *e-skid* coupler brings the even mode effective index much closer to the odd mode in comparison with the traditional coupler, and the 2D *e-skid* directional coupler has increased the difference between the two effective indices. Figure 2.5 (g,h,i) show the crossover length given the corresponding effective indices. The traditional coupler and *e-skid* behave as expected, with an increase in crossover length for the latter. The 2D *e-skid* directional

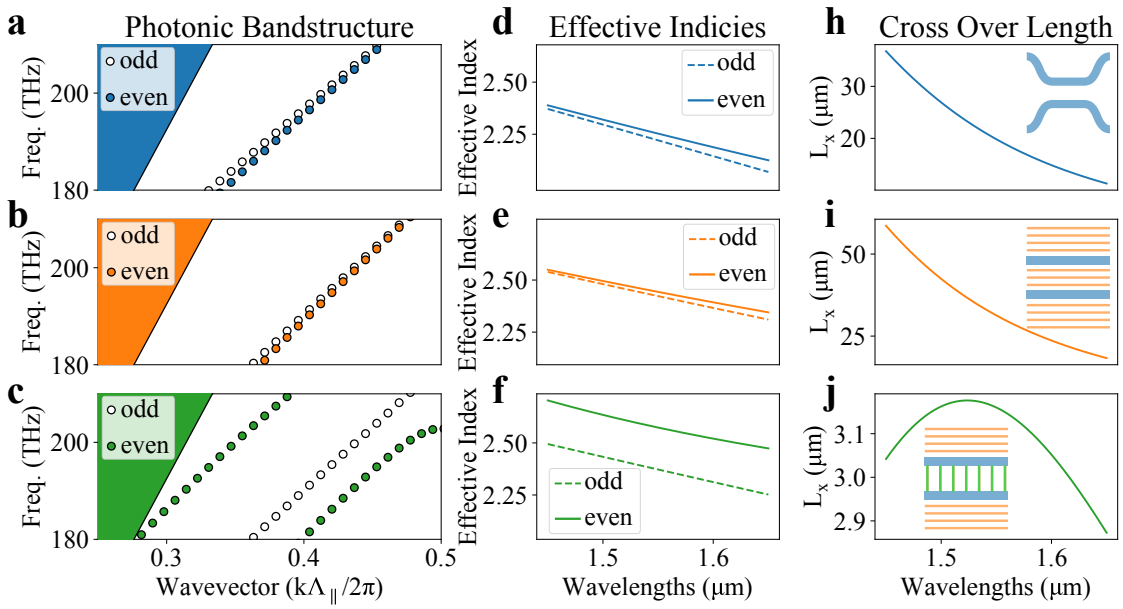

Figure 2.5: (a,b,c) The photonic bandstructures of the conventional, 1D *e-skid* and 2D *e-skid* directional couplers from Fig. 2.4 (a,b,c), respectively. (d,e,f) Extracted effective indices of the bandstructures from (a,b,c), respectively. (g,h,i) The crossover length is calculated from Eq. 2.6. The insets of (g,h,i) show the device diagrams for each type of coupler. These devices were all simulated with the same gap to illustrate the effects on the same scale. In practice, the gap is limited by (i) the fabrication process, (ii) the circuit application, and (iii) the length of the waveguides. Therefore, the gap is often larger than shown here. The *e-skid* design parameters were: gap = 270 nm, $\Lambda_{\parallel} = 50$ nm, $\rho_{\parallel} = 50\%$, $\Lambda_{\perp} = 270$ nm, $\rho_{\perp} = 50\%$, and $W = 400$ nm.

coupler exhibits a dramatically reduced crossover length, and in relation to dispersion engineering, a completely different shape. It is important to note that the 2D *e-skid* directional coupler does support a higher-order mode (Fig. 2.5 (c)), however, the coupling efficiency extracted from a modal overlap integral between the fundamental and the first higher-order mode is $< 8\%$ over the wavelength span for our designs, according to our 3D FDTD simulations [2]. While 8% is not insignificant, tweaking our parameters (specifically $\rho_{\perp}$ and $\Lambda_{\perp}$) can reduce this coupling efficiency into higher-order modes, increasing device performance [42]. For this experiment, the primary design choices were dictated from a perspective of manufacturability. In the future, small decreases to $\rho_{\perp}$ and $\Lambda_{\perp}$ will

result in higher performing couplers verified by our 3D FDTD simulations [2] and prior work [42]. Additionally, careful consideration of tapering, which is not investigated in this work, can also mitigate the excitation of higher-order modes [15, 43].

The 2D *e-skid* directional coupler is fundamentally different from a multimode interferometer (MMI). MMIs offer compact power splitting via a self-imaging effect that manifests from the excitation of higher-ordered modes, creating an interference pattern that repeats at a length specific to the physical parameters [44, 134]. The 2D *e-skid* directional coupler specifically operates as a *directional coupler*, where the characteristics of the device arise from the coupled supermodes supported by the two waveguides [49]. The perpendicular, metamaterial features act to control the decay constant of the two modes' evanescent tails by shaping the material between the waveguides, enabling 2D *e-skid* and dispersion engineering, as demonstrated by the simulation results in Fig. 2.3.

2.3.2 Device Design

We investigate the effect of different parameter variations of the 2D *e-skid* directional coupler. Figure 2.6 shows the results of varying the fill factor, period, and the gap between waveguides. These parameter variations indicate the substantial tunability offered by two-directional *e-skid*. First, we selected the operating gap between the two waveguides to be 1.44 μm in order to stay consistent with the parallel *e-skid* features ($\rho_{\parallel} = 60\%, \Lambda_{\parallel} = 225$ nm, 6 layers deep results in a 1.44 μm gap). We designed these

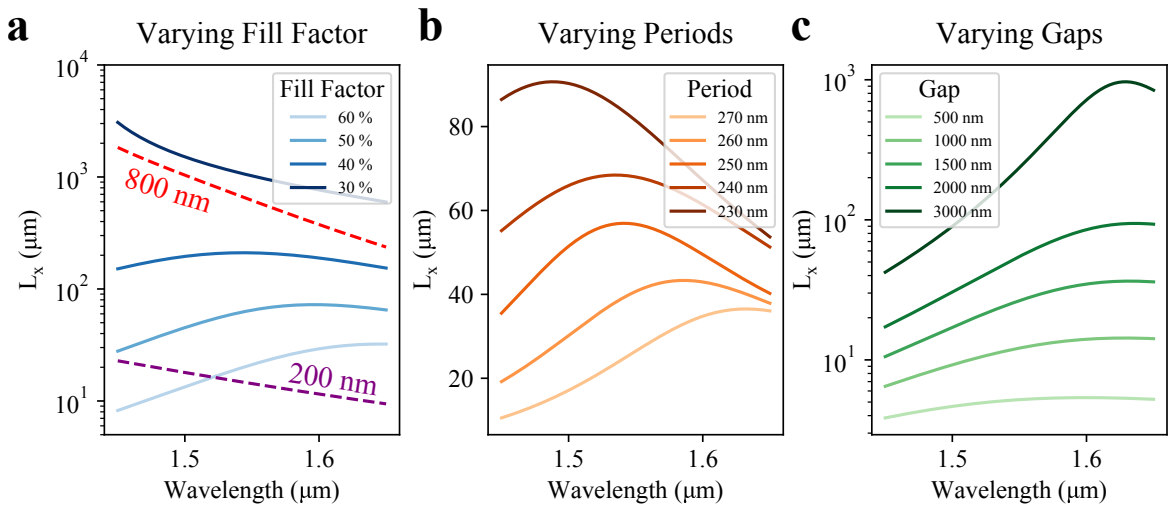

Figure 2.6: Dispersive plots representing the cross-over length for different varying parameters. (a) A fill factor sweep with period and gap fixed at 270 nm and 1500 nm, respectively. The reference lines indicate the cross-over lengths of traditional directional couplers with gaps of 200 nm and 800 nm. (b) A period sweep with fill factor and gap fixed at 60% and 1500 nm, respectively. (c) A coupling gap sweep with period and fill factor fixed at 270 nm and 60%, respectively.

devices for manufacturing with the American Institute of Manufacturing (AIM) Photonics CMOS foundry Multi-Project Wafer (MPW) offering. For a photolithographic process like this one, we must take into account the limitations of the processing, like feature size. For example, many prior work designs with features smaller than 60 nm would not resolve with CMOS processing compared to electron beam lithography. We chose to design our devices with $\Lambda_{\perp} = 275$ nm to remain beneath the Bragg limit but maintain high manufacturing quality. The parallel features were designed with $\Lambda_{\parallel} = 225$ nm. It should be noted that the parallel cladding structures are less challenging for a lithographic system because they are lines, not holes [106]. We targeted $\rho_{\perp} = 60\%$ for the majority of our devices because we assumed that the features would be over-etched, a common practice in SOI fabrication so that the fill factor would decrease [13].

2.4 2D *E-skid* Directional Coupler Device Measurements & Parameter Extraction

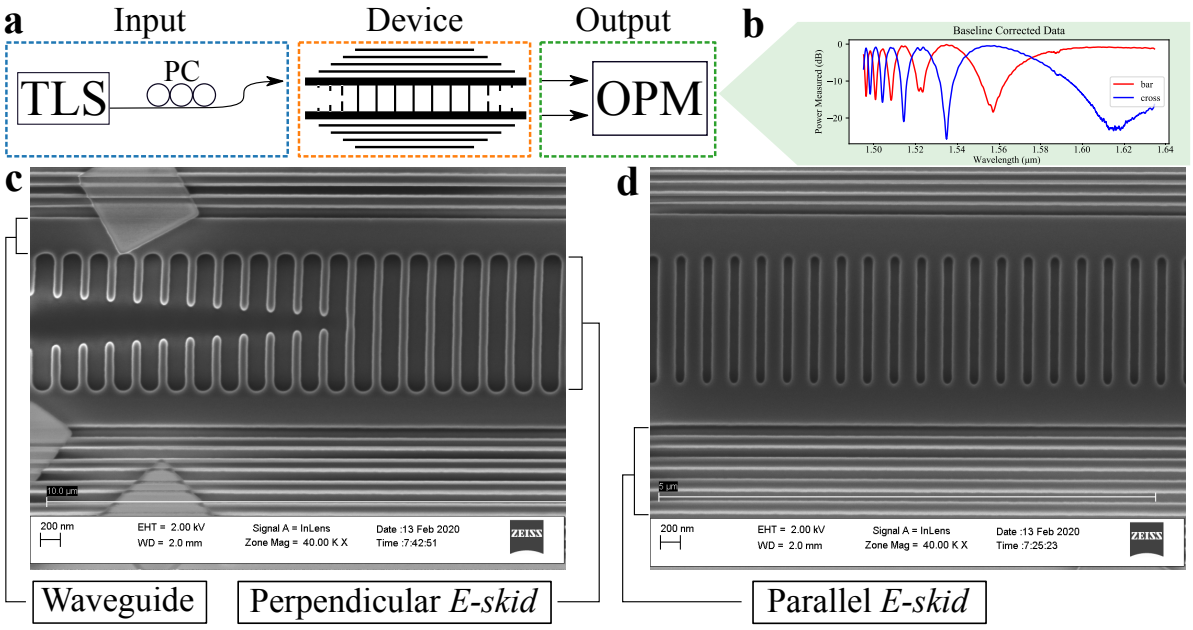

Figure 2.7: (a) Experimental setup, a tunable laser source (TLS) is connected to the device via a polarization controller (PC). The outputs of the device are then connected to an optical power meter (OPM). (b) An example spectrum from the measured data. (c,d) A scanning electron microscope (SEM) image of two different 2D *e-skid* directional couplers fabricated by AIM photonics, focused on the taper from the strip waveguides used to couple to optical fibers (c) and in the center of the device (d). The objects are debris from the oxide-release etch for the SEM.

2.4.1 Experiment

Our experimental setup is shown in Fig. 2.7 (a). We placed the chip on a mount in between two 3-axis stages with a bare fiber on either side for coupling in and out. For the input, we connected the fiber to a tunable laser source (TLS), and a polarization controller (PC) to ensure TE polarization and we measured the output signal with an optical power meter (OPM). The fibers were edge coupled to the chip, which routed the

light through strip/wire waveguides to the devices. To transition from the strip waveguide mode to the *e-skid*'s, we slowly introduced the parallel and perpendicular claddings as seen in Fig. 2.7 (c) and depicted in the schematic in Fig. 2.7 (a). We were careful to design simple tapers because when periodic, asymmetric features are introduced there is a chance for radiative losses [15]. We measured total device loss at ≤ 2 dB per device, which can be improved with longer tapers or carefully designed width-varying tapers to ensure a smooth modal transition.

We collected the transmission spectra (an example is shown in Fig. 2.7 (b)). The spectrum shows a characteristic "chirp"-like behavior, which is due to the dispersive nature of the cross-over length. This is seen in Fig. 2.6 where, based on the perpendicular *e-skid* parameters of the device, the cross-over length can exhibit a significant change with wavelength. This will result in a quickly oscillating output (given by Eqs. (4) and (5)). With that said, using the experimental measurements in conjunction with the previously described models we were able to extract the parametric dependence of the coupler designs.

2.4.2 Parameter Extraction Method

After extracting the transmission spectra from our fabricated devices, we used a dispersive model for extracting the cross-over length from the data. Because inverse sine functions are multi-valued, we can not obtain L_x directly from Eqs. 2.4 or 2.5. We instead employed the behavioral model for characterizing directional couplers [137]. For the cross-over length, we are interested in the wavelength dependence, so we prepared the

data by filtering the noise and normalizing the bar and cross measurements. We used a polynomial expansion of the coupling coefficient coupled with a non-linear least squares (NLS) optimization algorithm to find the best fit for L_x [2, 127, 129]. Because many of our devices exhibit a strong "chirp-like" behavior (Fig. 2.7 (b)), we used a third-order polynomial expression for L_x to determine the best fit, such that

$$L_x(\lambda) = L_{x,0} + L_{x,1}\lambda + L_{x,2}\lambda^2 + L_{x,3}\lambda^3, \tag{2.7}$$

where the curve is characterized by fitting parameters $L_{x,0}, L_{x,1}, L_{x,2}, L_{x,3}$ pertaining to the wavelength, λ. The NLS optimization aimed to minimize the difference between the measured and theoretical spectra by adjusting the fit parameters in Eq. 2.7.

2.4.3 Experimental Results

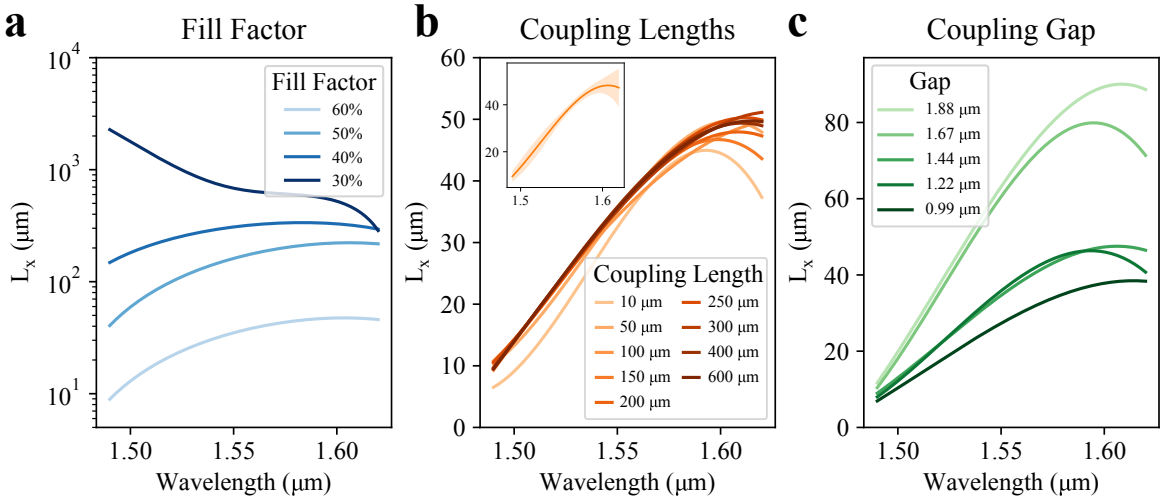

Figure 2.8: (a) Experimental extraction of L_x for the varying fill factors. (b) Experimental extraction of L_x for the varying coupling lengths, here, we expect that the values will be similar. (c) Experimental extraction of L_x for the varying coupling gaps, indicating an *average* change of L_x, but less conclusive over the range.

In Fig. 2.8 (a), we show the device's dependence on fill factor variation. Fill factors up to 60% were successfully fabricated in the CMOS process and will be reported below, while fabrication-specific optimization needs to go into higher fill factor devices. The fill factor variations resemble those from the simulations (Fig. 2.6 (a)), where higher fill factors increased L_x. In Fig. 2.8 (b) we investigated coupler-length variation for a fixed fill factor (60%) and gap (1.44 μm). We fit nine different directional couplers with the exact same parameters changing only the coupling length. According to Eqs. 2.4 and 2.5, the length is independent of the coupling coefficient, κ and therefore these couplers should exhibit identical L_x measurements. There are manufacturing variations, measurement errors, and fitting errors that reveal themselves in Fig. 2.8 (b). The inset shows a 95 % confidence interval for these nine couplers' L_x extraction and showcases expected similar behavior for all of these couplers. This truly highlights the utility of these devices, as they are able to couple fully in ≤ 50 μm, even with a large gap of 1.44 μm. Finally, Fig. 2.8 (c) shows the L_x extraction for varying gaps. Even though there is a qualitative match between the simulations and experimental results, manufacturing variations account for the quantitative differences.

2.5 Future Work

In this work, we performed a comprehensive study of the parameter space of our proposed directional coupler using two-dimensional *e-skid*. In the future, it will be desirable to realize devices with particular performance characteristics. From our experimental data (Fig. 2.8), we extracted that we are able to realize a 2D *e-skid* directional coupler that

achieves 100% coupling (100/0 splitting ratio) with a coupling length of $L = 50$ μm as seen in Fig 2.9 (a), using $\rho_{\perp} = 60\%$, $\Lambda_{\perp} = 275$ nm and coupling gap of 1.44 μm. We can take this to design a 50/50 directional coupler. Figure 2.9 (b) shows the theoretical transmission spectrum of this device. We set the coupling length $L = \text{avg}(L_x)/2 = 23.5$ μm, and we see broadband behaviour of nearly 40 nm. Additionally, we can slightly vary parameters to tune the device to a more desirable center wavelength. For example, by reducing the period of the 2D *e-skid* directional coupler to $\Lambda_{\perp} = 255$ nm, $\rho_{\perp} = 50\%$, and coupling gap of 1.44 μm, the device's operating band can now be centered closer to 1.55 μm and achieves an even larger operating bandwidth of > 40 nm (Fig. 2.9 (c)), which is sufficient for many applications.

2.6 Conclusion

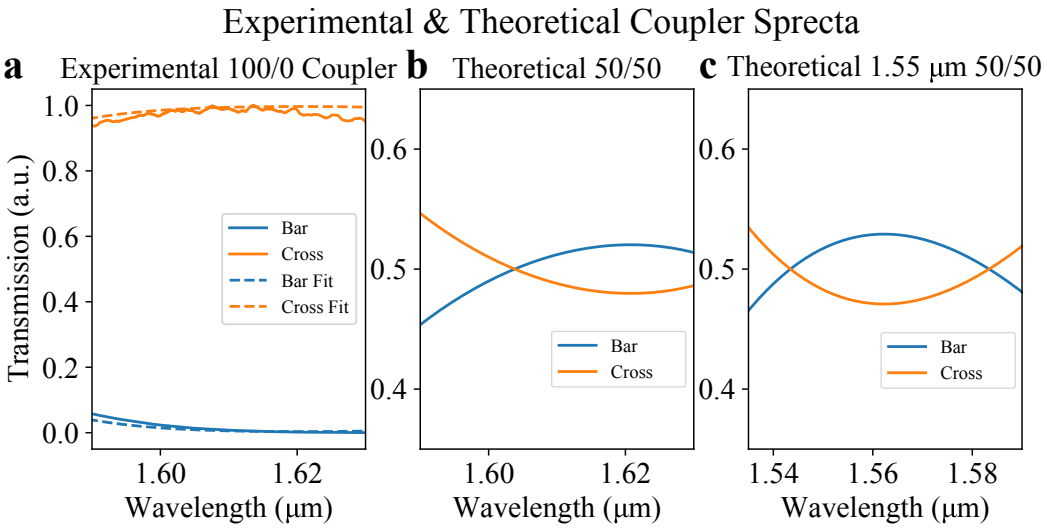

Figure 2.9: (a) Experimental and Fitting results for an L_x Coupler (100/0) splitting ratio. (a) Based on the experimental results in (a), we propose a 50/50 splitter operating from 1.58 to 1.62 μm. (b) Based on simulation results from Fig. 2.6 (a), we propose a 50/50 splitter operating from 1.54 to 1.58 μm.

We introduced the deterministic, targeted control of the evanescent wave in the TE mode of strip waveguides by employing *e-skid* features in two directions. We designed and demonstrated 2D *e-skid* directional couplers on a CMOS photonic chip fabricated by AIM Photonics. All of the results were compared to simulations by extracting design parameters, using a NLS optimization technique coupled with a behavioral model of the directional coupler [137]. With the parameter extraction, we show experimentally that *e-skid* waveguides, and the 2D *e-skid* directional coupler in particular, are possible to realize in a CMOS platform. Moving forward, we can design full two-dimensional *e-skid* circuits that achieve high densities by both suppressing and enhancing coupling at will, while operating with large bandwidths.

Appendix A: Verifying Mode Conversion Efficiency of 1D *e-skid* for CMOS Photonics

We verify the conversion efficiency between strip waveguides and *e-skid* waveguides by comparing the loss coefficients at the resonances of a racetrack resonator [45, 70]. We designed three strip waveguide racetrack resonators, Fig. 2.10 (a), with varying gaps, and then the exact same racetrack resonators in which we add two *e-skid* features into the ring, Fig. 2.10 (c). The intent of this experiment is to quantify the additional loss created by these features, and thereby measure the mode conversion efficiency between the strip and *e-skid* waveguides [54]. A simple ring resonator (shown in Fig. 2.10 (a)) can be parameterized by two coefficients, τ, the self-coupling coefficient that indicates

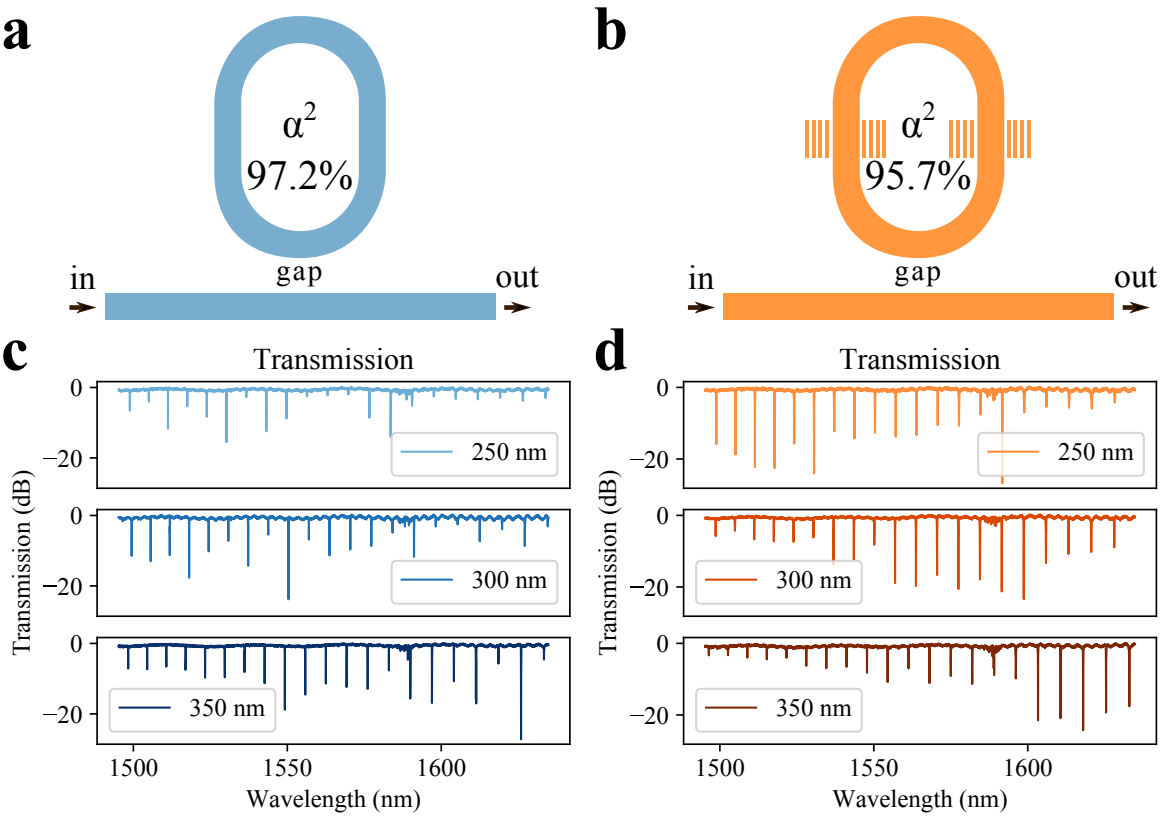

Figure 2.10: (a) A strip waveguide racetrack ring resonator. We measured an average loss into the ring of 97.2% at the resonant peaks. (b) The transmission spectra for three equivalent strip waveguide racetrack ring resonators with varying coupling gaps of 250, 300 and 350 nm. (c) A strip waveguide racetrack ring resonator with two parallel cladding features for comparison with (a). The loss into the ring was measured at 95.7% on average. (d) Transmission spectra for the three rings of the same varying gaps as (b) with the addition of the features.

how much light goes through the coupler, and α, the loss coefficient which indicates how much light is lost into the ring. We extract α and τ according to the method described

by [70], such that

$$\mathcal{F} \equiv \frac{\Delta\lambda_{\mathrm{FSR}}}{\Delta\lambda_{\mathrm{FWHM}}}, \tag{2.8}$$

$$\mathcal{E} \equiv \frac{T_{\mathrm{MAX}}}{T_{\mathrm{MIN}}}, \tag{2.9}$$

$$A = \frac{\cos(\pi/\mathcal{F})}{1+\sin(\pi/\mathcal{F})}, \tag{2.10}$$

$$B = 1 - \left(1 - \frac{\cos(\pi/\mathcal{F})}{1+\cos(\pi/\mathcal{F})}\right)\frac{1}{\mathcal{E}}, \tag{2.11}$$

$$(\alpha,\tau) = \left[\frac{A}{B}\right]^{1/2} \pm \left[\frac{A}{B} - A\right]^{1/2}. \tag{2.12}$$

The finesse, $\mathcal{F}$, is defined as the ratio between the free spectral range, $\Delta\lambda_{\mathrm{FSR}}$, and the full width at half maximum, $\Delta\lambda_{\mathrm{FWHM}}$, of each resonance. The extinction ratio, $\mathcal{E}$, is defined as the ratio between the transmission maximum, T_{MAX}, off-resonance and the minimum, T_{MIN}, at each resonance. We can decouple α, τ in Eq. 2.12 using the method further discussed in [70]. When we determine α, we know that α^2 indicates the percentage lost into the ring, which allows us to compare these two different resonators. This resulted in average values of $\alpha^2 = 97.2\%$ and $\alpha^2 = 95.7\%$ for the two ring types (Fig. 2.10 (a,c)), respectively. This leads to a mode conversion efficiency of 99.6%.

Chapter 3

High ER Microring Modulator

3.1 Motivation

The development of silicon photonic resonant modulators has been driven by the need for compact, high-speed modulation on chip [122, 133]. These modulators offer several advantages for communication systems, particularly for applications such as wavelength-division-multiplexing (WDM) and low-power consumption. Resonant modulators are particularly suited for these applications because of their ability to efficiently transmit multiple wavelengths of light, making them a useful tool for the creation of dense wavelength-division-multiplexing (DWDM) systems that are used to increase the capacity of fiber optic communication systems.

Resonant modulators prioritize operating bandwidth and power consumption over extinction ratio (ER), which is the ratio of the intensity of light passing through the modulator to the intensity of light that is blocked by the modulator. This is due to the inherent photon-limited frequency bandwidth, which is a fundamental limit on the ability of a system to detect and process light [38]. High-speed resonant modulators are designed

to operate at high frequencies and with low power consumption, making them ideal for applications such as optical interconnects and high-speed communication systems.

However, there are also applications for high ER resonant modulators, particularly in the areas of quantum communication and optical memory [66]. In these applications, the priority is on the resonator ER as a filter or switch. Resonant modulators can be used as optical switches to control the flow of light in a quantum communication system or to store information in an optical memory device.

3.2 Theory

MRR theory is covered in Sec. 1.4.4. However, here we expand on the relevant features here relating to the design of electro-optic MRRs. From an abstract perspective, the MRR is two components: a waveguide coupler and a waveguide loop back on itself (as shown in Fig. 1.5 b.). We can consider these two regions into two separate equations, starting with the coupler equation as

$$\kappa^2 + \tau^2 = 1, \tag{3.1}$$

for κ as the cross coupling and τ as the through coupling of the coupler. We also have the loss in the cavity region as

$$A^2 = e^{-\alpha/L}, \tag{3.2}$$

with α as the propagation loss in per unit length, and the length. Often the length is described in microns. We then can describe the coupling condition hinges on the

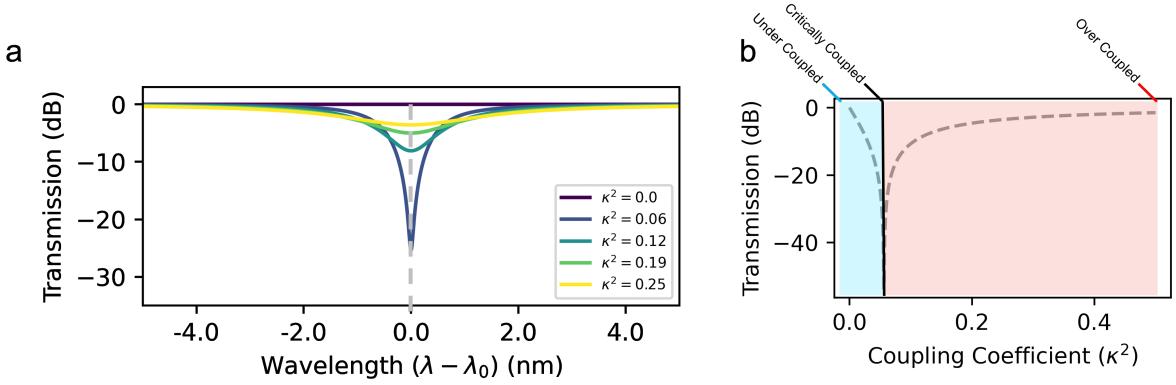

Figure 3.1: a) The transmission spectrum of the MRR as we change the κ^2 from 0 to 0.25 for a fixed $A^2 = 0.94$. b) The corresponding plot of a) where we have fixed the wavelength at $\lambda = \lambda_0$ to show how the coupling coefficient changes the condition of the resonator.

relationship between the cross-coupling coefficient (τ) and the loss in the cavity (A). We now have three categories of the coupling condition of MRRs, which have similarities to how we describe spring motion in classical mechanics: critical, over, and under. Critically coupled is when $\tau = A$, over-coupled is when $\tau < A$ and under-coupled is when $\tau > A$. We can describe this relationship using the cross-coupling coefficient κ too, which results in the following standard expressions of the resonant conditions:

$$1 - \kappa^2 = A^2 \mid \text{Critically Coupled},$$

$$1 - \kappa^2 < A^2 \mid \text{Over Coupled},$$

$$1 - \kappa^2 > A^2 \mid \text{Under Coupled}.$$

The primary considerations are the coupling coefficient of the coupler and the loss within the cavity.

For example, we can show what happens if we have a cavity fixed to $A^2 = 0.94$, where the length of the cavity is arbitrary. With such a cavity, we can vary the coupler by

changing the value of $\kappa^2 \in [0, 0.25]$. In Fig. 3.1 a), we can see how this scenario is realized for the spectral behavior of the ring resonator. Here, we begin with no cross-coupling into the cavity, which implies that the guided mode will pass at all wavelengths for this coupler. However, as we increase the κ coefficient of the coupler, we see that for the given wavelength that matches the resonant condition, i.e., from eq. . And finally, we begin to over-couple into the resonant cavity, and less light is "trapped" in the cavity. We show this in the dB scale, and in this case "critical-coupling" is seen near 25 dB extinction, or 3×10^{-1} % of the input signal is passed through (implying 99.7 % of the input magnitude is coupled into the cavity). Fig. 3.1 b) shows the regions for under-coupling, critical-coupling, and over-coupling for the fixed wavelength at resonance ($\lambda_0 - \lambda = 0$).

In practical terms, we are able to manipulate κ in two different coupler styles. If we have a point-coupler for the MRR, which is the case where the ring is brought to some small distance from a straight waveguide, we simply control the κ, τ values by increasing or decreasing the gap of the coupler. Other couplers, known as "pulley" couplers or racetrack couplers, are designed to control the interaction length of the coupling region. In this type of coupler, both the interaction length and the coupling gap are degrees of freedom to design the κ, τ values. Primarily, interaction couplers are used to try to control the spectral characteristics of the coupling or to suppress higher-ordered modes in the case of micro-disk resonators []. The coupler design is primarily a fixed parameter in the design phase, and after fabrication, the value of κ, τ does not change.

Similarly, we can show in Fig. 3.2 the case where the coupler of the MRR is fixed, and we change the loss conditions of the resonator. In this example, we imagine the coupler has the cross-coupling coefficient $\kappa^2 = 0.05$. Here the loss in the cavity can be described

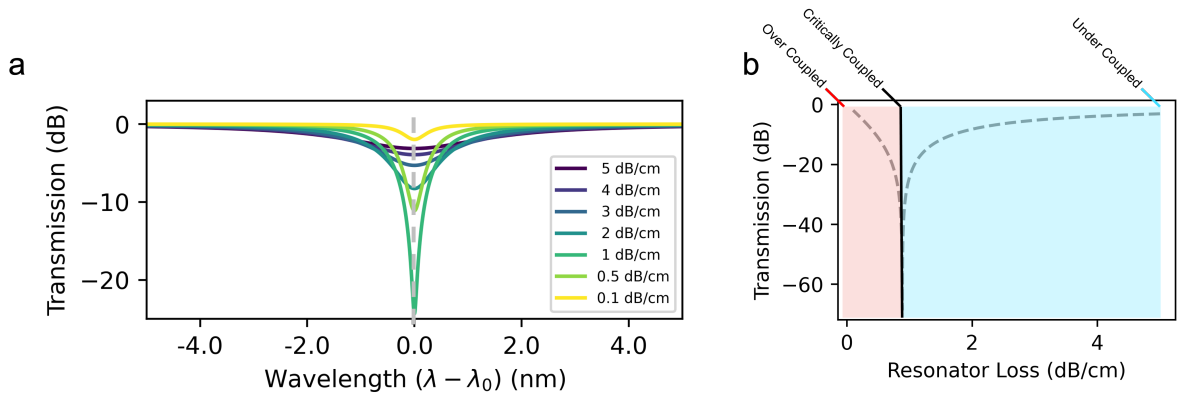

Figure 3.2: a) The transmission spectrum of the MRR as we change the A^2 by changing the propagation loss coefficient from 0.1 dB/cm to 5 dB/cm. b) The corresponding plot of a) where we have fixed the wavelength at $\lambda = \lambda_0$ to show how the loss in the cavity (as it relates to this propagation loss) will change the condition of the resonator.

by the loss of the waveguide. If the cavity has a radius of $5\mu m$, then this implies we can calculate the round-trip loss using the propagation loss α to describe our cavity loss as $A^2 = exp(-\alpha 2\pi 5\mu\text{m})$. We vary the value of $A^2 \in [0.1, 5]$ dB/cm. Fig. 3.2 a) shows the spectral behavior of this MRR again. At first, we couple too strongly into the MRR when the loss in the cavity is high. As the loss decreases in the MRR, the condition changes to critical and finally to under-coupled.

As mentioned in this example, we can control the resonator's loss by varying the cavity's loss. Unlike the coupler, this is often exploited to create MRR-based electro-optic devices. To practically implement this, we need to utilize a method for imparting the loss change. Many options are available; for filters or low-speed filtering, a thermo-optic or MEMs phase shifter is embedded into the cavity. Electro-optic diodes are embedded into the cavity for high-speed switching data transfer, where the junction design becomes the primary consideration. We will explore this in the following section.

A final theoretical discussion point for designing a high ER MRR is understanding

the ER and the quality factor (Q). The ER is the ratio of the minima and maxima of the MRR spectral (or driven) response. The Q of the ring is a measure of the width of the resonance, indicating how wide or narrow the resonance operation is on the signal's wavelength. The Q of a ring is described as

$$Q = \frac{\lambda_{res}}{\mathrm{FWHM}_{\lambda_{res}}} = \frac{f_{res}}{\mathrm{FWHM}_{f_{res}}}, \tag{3.3}$$

where $\mathrm{FWHM}_{\lambda_{res}}$ is the Full Width at Half Maximum of a given resonance in either term of wavelength or optical frequency. In Ref. [122], we find a relationship between the optical Q and the modulation (or drive) bandwidth of the resonator, described as

$$f_{3dB}^{opt} = \sqrt{\sqrt{2} - 1} \cdot f_{\mathrm{FWHM}} \tag{3.4}$$

An example resonator is shown in Fig. 3.3, where we relate the optical modulation limited bandwidth to the Q factor using the same theoretical ring as above at critical coupling. We indicate that the upper left portion of this log-log curve describes much of the work employing resonators for information processing [97, 121, 122]. Additionally, we describe the slightly lower left of this region for quantum and low signal information processing, which we highlight as the motivation for this particular high ER modulator. For quantum systems, high-quality filtering on the chip is often challenging to achieve for quantum systems yet necessary. In addition, this can be utilized in optical memory or quantum optical clocking [30, 34, 126]

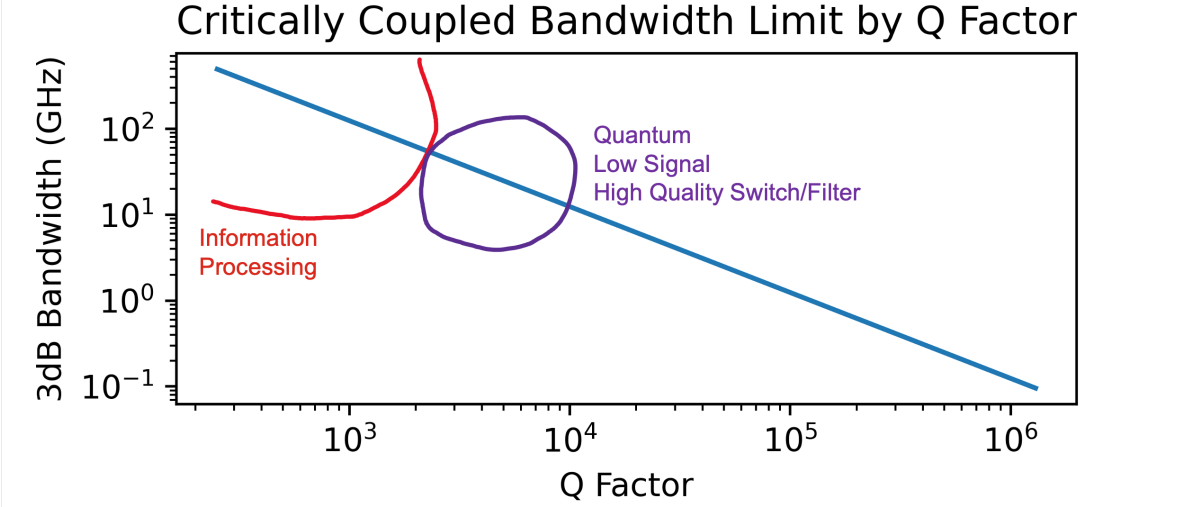

Figure 3.3: A log-log plot of the relationship between Q factor and optical modulation limited bandwidth of a critically coupled ring at 1.55 μm. We also indicate two rough regions where we often find devices designed for traditional or classical information processing and for quantum or low signal processing.

3.3 Design & Simulations

There are multiple considerations and approaches to designing an MRR like the one we have described. We chose an interior ridge MRR design as the starting point for this design. This design enforces the single mode criterion (which can be difficult when using a microdisk, for example). We also see that the point coupler is more predictable than the pulley coupler, allowing us to tailor the coupling conditions. However, there are many ways to create this type of resonator – what follows are the design steps to achieve our device.

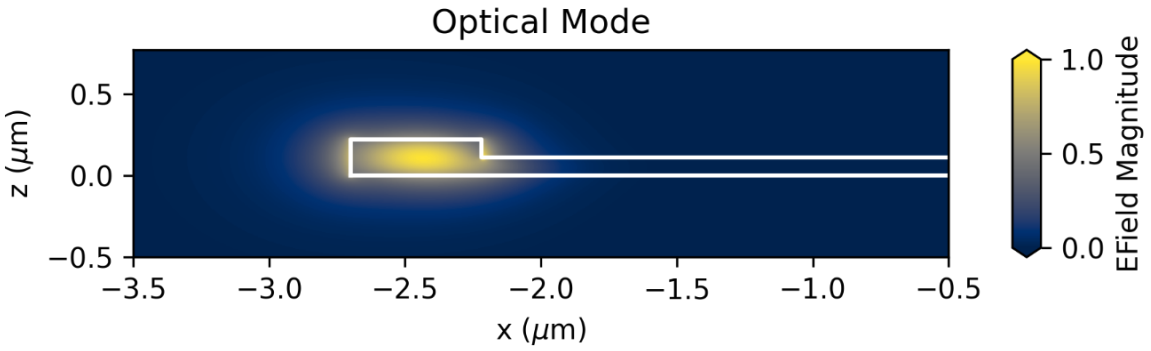

Figure 3.4: The optical mode of the interior ridge optical resonator. We impose the bend radius of 2.75μm on the bend to simulate the effect inside the resonator.

3.3.1 Simulate the Optical Mode & Determine the Bend Radius

We simulate the optical mode with Ansys-Lumerical's FDE tool [3]. We describe the waveguide cross-section as "interior ridge," which describes a waveguide with an asymmetric rib structure similar to Fig 1.4 a). The only difference is that the thin ridge would only extend in one direction; for our device specifically, this is the radially inward direction; however, for the simulation, it is arbitrary which direction. The simulation helps us to find the supported modes for the structure we are interested in within the variations, such as the core waveguide width and the bend radius. Fig. 3.4 shows the final supported mode for this design and overlays the physical structure. Here, we impose a 2.5 μm bend radius on the waveguide and ensure it can continue supporting the mode.

Ultimately, we turn to a full-wave simulation with FDTD to simulate the loss of the bent waveguide [2]. This step is required since FDTD can capture the radiation loss of the structure. This simulation has a bearing on the final radius we ultimately choose, as it accurately estimates the contribution of cavity loss we can expect from the bent waveguide. In our case, the cavity loss of the inner ridge modulator is under 1 dB after

the radius increases to 2.75 μm. For passive MRR design, this would be more important. However, we expect the loss in the cavity to be dominated by doping, which we will discuss after the next step.

3.3.2 Simulate the Coupling Gap

We simulate the coupling gap of the resonator as a point-coupler in FDTD [2]. This means that we can move the waveguide further away to decrease the coupling coefficient. To simulate, that is precisely what we do. We start as close as we can physically space the silicon feature ($\sim$ 100 nm is typical), and we iteratively move the waveguide further away from the cavity while recording the transmission into the cavity as the κ. We see the result of this sweep in Fig. 3.5 a), along with a small sketch indicating the point coupler's gap increase and simulation method in b).

3.3.3 Simulate the Junction Design

Junction design is a rich and deep topic of discussion. These references specifically provide fantastic overviews of modulator junction design in silicon photonics, including discussions of high-speed and low-power modulation [57], recent advances in silicon modulator technology [28], and resonant modulator design [48,67,122]. In this subsection, we summarize diode manufacturing and then discuss our design choices and methods related to this device.

The n- and p-type dopants are small impurities implanted into the silicon, often phosphorus or arsenic for n-type and boron for p-type, which change the electron-hole

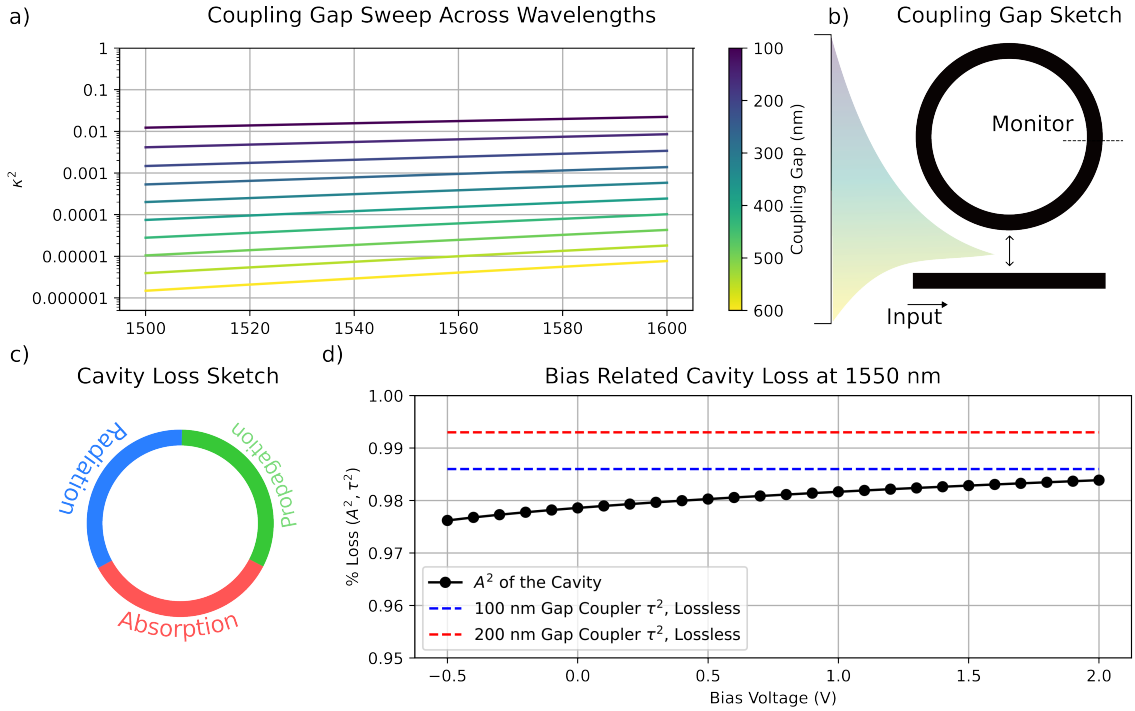

Figure 3.5: a) Here we see the coefficient of κ^2 for the point coupler of the inner ridge modulator design. This coefficient has a large dependence on the coupling gap, showing that on one side, we achieve over 1% coupling with a 100 nm, and on the other, we see 1×10^{-6}% coupling at a gap of 600 nm, effectively no coupling here. b) A sketch describing the method and design concept of the coupling gap. We shift the input waveguide away from the resonant cavity to achieve a new κ^2. c) A sketch describing the primary cavity loss factors: Absorption, Propagation, and Radiation. It is important to note that absorption from doping is the primary loss agent for this cavity. d) The results of the optimized device cavity loss concerning *reverse-biased* – meaning the sign here is flipped – voltage using CHARGE and FDTD simulations to arrive at A^2 [1, 2]. We also impose the coupler τ^2, which is extracted from the simulations in a) to illustrate how we can select the gap of this coupler.

relationship of the "intrinsic" silicon. Silicon is therefore considered a semi-conductor, as it is ideally a near insulator in its pure form but can be doped to a near conductor in highly concentrated doses. Of course, this implies, too, that we can use n- and p-types of dopants to create diodes that have opposite polarity from one side of the center to the other – allowing current to flow in only one direction. The interface between the two materials forms a depletion region, which is a region with few mobile charge carriers. When a forward bias voltage is applied to the diode, the depletion region

narrows, allowing current to flow. When a reverse bias voltage is applied, the depletion region widens, blocking current flow. From an optical perspective, these carriers are related to the silicon's refraction index by the free-carrier plasma dispersion effect.

The free-carrier plasma dispersion effect in silicon occurs when the concentration of free charge carriers, such as electrons and holes, is high enough to affect the material's refractive index at high frequencies. This effect is described by the Soref equation, which considers the free carrier concentration, wavelength, and material properties [83]. The empirical relationship for the refractive index (real and imaginary) of silicon at $\lambda = 1550$ nm is described as:

$$\Delta n = -5.4 \times 10^{-22}(\Delta N)^{1.011} - 1.53 \times 10^{-18}(\Delta P)^{0.838}, \tag{3.5}$$

$$\Delta \alpha = 10 \log_{10} e^{-8.88\times10^{-21}(\Delta N)^{1.167} - 5.84\times10^{-20}(\Delta P)^{1.109}}, \tag{3.6}$$

for the carrier concentrations $\Delta P, \Delta N$. We see here that the carrier concentration *nearly* changes the index of refraction linearly, where in comparison, the absorption coefficient (imaginary part of refractive index) varies logarithmically with concentration. This relationship places the onus on careful junction design near the optical mode to find the right balance between refractive index change (what we want for a phase shift) and loss (what we want to minimize).

We begin our junction design with the available doping profile set by the fabrication process. In this development wafer with AIM Photonics, we had access to doping profiles that enable a vertical junction in a rib/ridge waveguide. To achieve this type of junction, we had access to 6 specific dopings: three p-type dopings of increasing concentration and

three n-type dopings of increasing concentration, called P/N, P+/N+, and P++/N++.

For this process to allow a vertical junction, N is targeted for the top of the full-height silicon (nominally 220 nm in thickness), and the N+ doping acts as a series resistor linking the full-height silicon with the half-height ridge region, as shown in Fig 3.6 a). We then can use contact doping, N++, to connect our metal to the bond pads at the surface. Conversely, the p-type dopant sits on the bottom of the full-height silicon, completing the vertical junction (P-N) with the higher n-type. The p-type bottom also has a series resistor in P+, which stays in the lower half of the waveguide and is primarily used to link this with the contact. Finally, P++ creates the contact for the anode side up to the bond pads.

While Fig 3.6 a) shows the cross-section of the doping, we can describe the extrapolated "spoke" style of the inner ridge modulator more clearly. To create the vertical junction as described, with full coverage of the cavity, we employ an interleaving method [121]. This method means that we have some regions where there are n-type contacts, and somewhere there are p-type contacts, which alternate around the cavity's radius. At the point of reaching the core waveguide, the light dopings of P and N are extended along the core to link with the adjacent contact regions. This contact structure allows for a single vertical junction with multiple contacts interleaved along the cavity, all while staying geometrically within the design rules provided by the foundry. Fig. 3.6 c) shows a clearer image of this interleaved pattern.

With the entire cavity covered by this PN diode, we can look closely at the waveguide's core. Recall the optical mode from Fig. 3.4, which depicted the e-field of the supported, fundamental mode primarily in the core of this waveguide. This intuitively implies that

the most important and sensitive region of the waveguides material properties belongs in this core area. It is, therefore, that junction design aims to create the optimal effect where the light is. We see in Fig 3.6 b) from our CHARGE simulation that the diode is created in the waveguide core between the N and P types [1]. As we increase the voltage to the cathode, thereby increasing the reverse bias, we see the depletion region also increase. We import this carrier profile from CHARGE into the FDE solver to numerically extract the change of our mode's effective index while populating the silicon model with the Eqs. 3.5, 3.6 [1, 3, 83]. We calculate the loss change using the imaginary part of the refractive index based on the results of the CHARGE simulation, propagation loss from process information, and radiation loss from MODE. However, the carrier absorption loss is the primary agent for loss change. We evaluate this for each bias point and plot the results of the cavity loss A^2 in Fig. 3.5 d).

3.3.4 Put It All Together

We finally put the design together, drawing the interleaved doping pattern in the cavity and selecting the correct coupling gap. For the experimental setup, we sweep the coupling gap between 100 and 200 nm, accounting for manufacturing variations in the search for our ideal device. Fig. 3.6 c) shows the ring's completed design, which we will test in the next section.

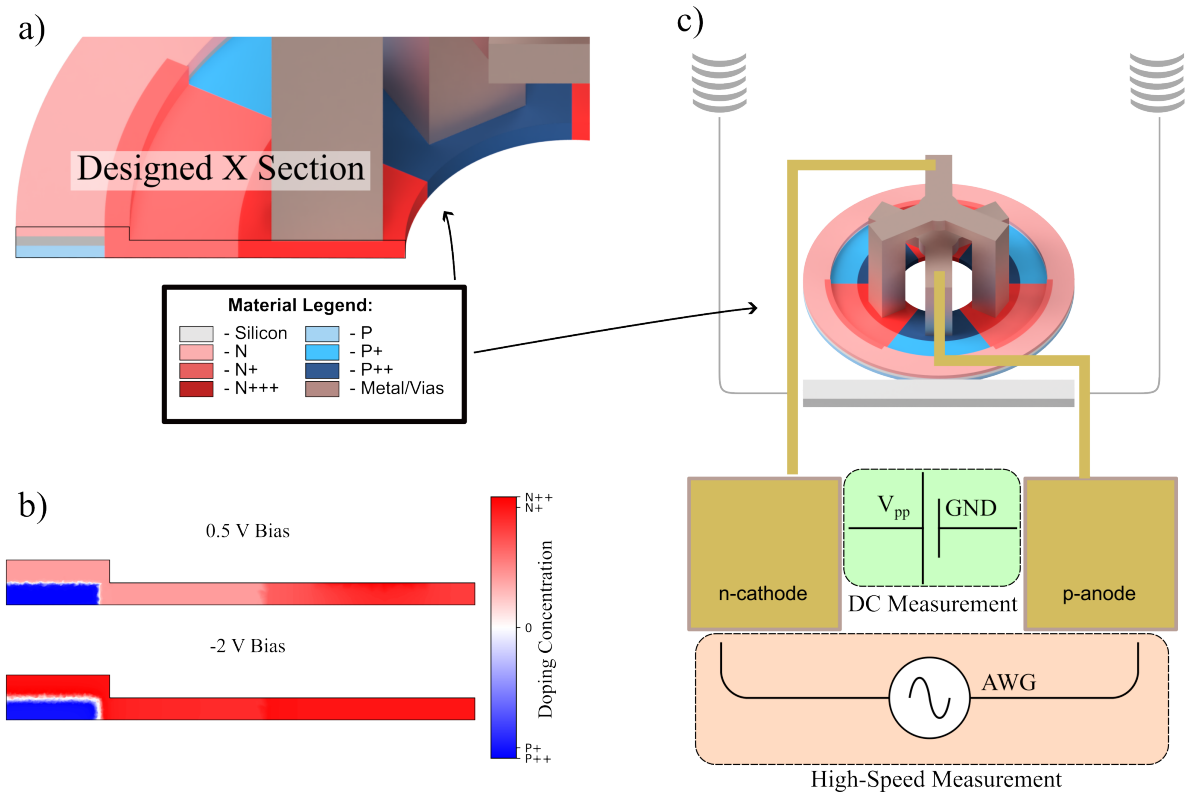

Figure 3.6: a) The nominal junction design, where we see how the different dopings form the junction. b) Two CHARGE simulations that show how the junction depletion region changes with increased reverse bias [1]. c) The final device design, circuit, and methods for DC measurement and High-Speed Measurements used for device characterization.

3.4 Experimental Set Up

3.4.1 DC Characterization

In this study, we utilized a Keithley 2400 source measure unit (SMU) to perform DC characterization of our circuit. This technique involves iteratively biasing the circuit from 0.5 V to -2 V (as shown in Fig. 3.7 a)) to shift the resonance fully. The Keithley 2400 SMU provides a precise means of sourcing accurate voltage and measuring the current of our device.

Our DC characterization measurements showed that our circuit has a strong "on-off" modulation depth of 30 dB over the voltage range tested (as shown in Fig. 3.7

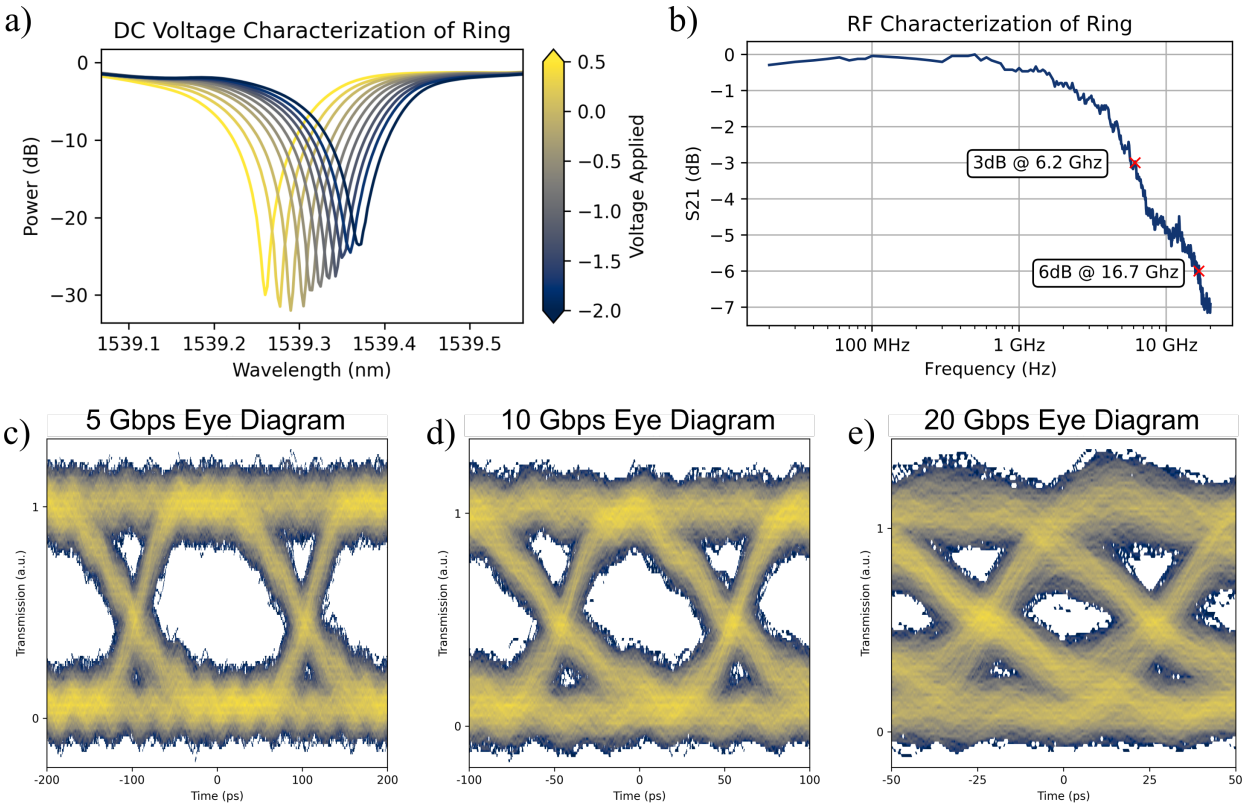

Figure 3.7: a) The optical spectrum response concerning a change in voltage bias. b) An RF S21 measurement using a sine-wave voltage input and frequency sweep from 20 MHz to 20 GHz. The $f_{3dB} = 6.2$ GHz c - e) Eye diagrams corresponding to 5, 10, and 20 Gbps input signals, generated by an AWG using a PRBS $2^9 - 1$

(a)). This result indicates that our circuit can effectively modulate the light signal. Additionally, we performed multiple measurements over a random selection of dies across the 300mm wafer. The extinction ratio (ER) was 27 ± 2.5 dB across the selected die over the bias range. This indicates that our circuit can effectively suppress the undesired signal, allowing for a higher signal-to-noise ratio. In addition to the ER measurement, we measured the resonance shift at -2 V bias across the wafer, which was found to be 100 pm $\pm$10 pm.

3.4.2 High Speed Characterization

We substituted the SMU with a Keysight M8195A AWG for high-speed measurements and directed the optical output to a 33GHz oscilloscope (Keysight UXR UXR0334A with N7004A). We applied a sine-wave voltage input to characterize the response and swept the frequency from 20 MHz to 20 GHz. Figure 3.7 b) illustrates the frequency sweep results, demonstrating that our S21 measured response decays to 3 dB at 6.2 GHz.

The high-ER modulator offers an advantage, as it achieves a significant modulation depth even at frequencies as high as 16.7GHz, which is evident from the analysis of three measured eye diagrams. We configured the AWG to produce NRZ, amplified it to $2V_{pp}$ using a Thorlabs MX40A Benchtop Modulator Driver, and plotted the results for a 5, 10, and 20 Gbit$\cdot$s^{-1} measured eye-diagram with a PRBS $2^9 - 1$ signal in Fig. 3.7 (c-e). The reported ER for each measurement is $17.6, 14.4$, and 13.9 dB, with higher results expected in the future as we used a less-critically coupled resonance for the dynamic measurements, owing to a sub-optimal grating coupler peak transmission. Furthermore, when we increased V_{pp} to $4V$, we obtained an even higher ER > 25 dB for speeds up to 10 Gbps.

3.5 Results & Discussion

We design a high ER microdisk modulator fabricated with AIM Photonics. This modulator employs a vertical p-n junction, enabling a small radius and large electro-optic effect. We measured a bias agnostic 27 dB ± 2 dB ER across the wafer. The high-speed

characterization demonstrated a slightly lower ER at 20 Gbps, limited by the transmission peak of the grating coupler. This device is an important step toward high-speed silicon photonic quantum information systems.

Chapter 4

Thermal Isolation of Phase Shifters

4.1 Introduction

Thermo-optic devices are necessary for many silicon photonics switching, filtering, and programming tasks. Unfortunately, thermal components typically consume tens of milliwatts per element and are susceptible to considerable intra-chip parasitic crosstalk. For these reasons, thermal devices do not meet the performance requirements inherent in large-scale circuits, such as microring arrays or programmable circuits [12]. However, the efficiency of thermal devices can be dramatically increased by selectively removing the silicon substrate under the heater, effectively eliminating the path of lowest thermal resistance and thus isolating the device. Drawbacks of previous demonstrations of efficient undercut thermal phase shifters in the silicon-on-insulator (SOI) platform include the use of a backside etch for substrate removal [23] and fabrication in a low-volume e-beam platform [114].

In the following sections, we design, present, and validate a thermally isolated phase-shifting cell fabricated in a commercial 300 mm CMOS foundry, which exhibits low-crosstalk and low-power operation. While minimal post-processing steps were used for the final substrate removal, this approach can be extended to wafer-scale processing. This enables future high-throughput fabrication of circuits containing thousands of phase shifters with minimal crosstalk and energy consumption.

4.2 Design and Simulation

Thermal phase shifters are a well-understood and oft-implemented unit cell for many applications on PICs. Therefore, the methodology for design and implementation is well understood. We provide a short tutorial on our thermo-optic phase shifter (TOPS) design. The phase shift is induced by the thermo-optic effect, which causes the refractive index of the waveguide material to change with temperature. We can introduce the heating element as silicon rails near the waveguide core or else as a resistive metal generally above the waveguide core (embedded in the cladding, which is typically glass). These two geometries represent the standard available method in a CMOS photonics process. In our design process, we have a variety of silicon rails available, as seen in Fig. 4.1

This work represents the effort to integrate thermal isolation into the process at the wafer scale. This process creates an air gap between the heater and the waveguide, which can significantly improve the efficiency of the thermal phase shifter. One of the main benefits of thermal undercut is the reduction of the thermal capacitance, which allows for faster and more precise modulation of the optical phase. Additionally, the air gap acts

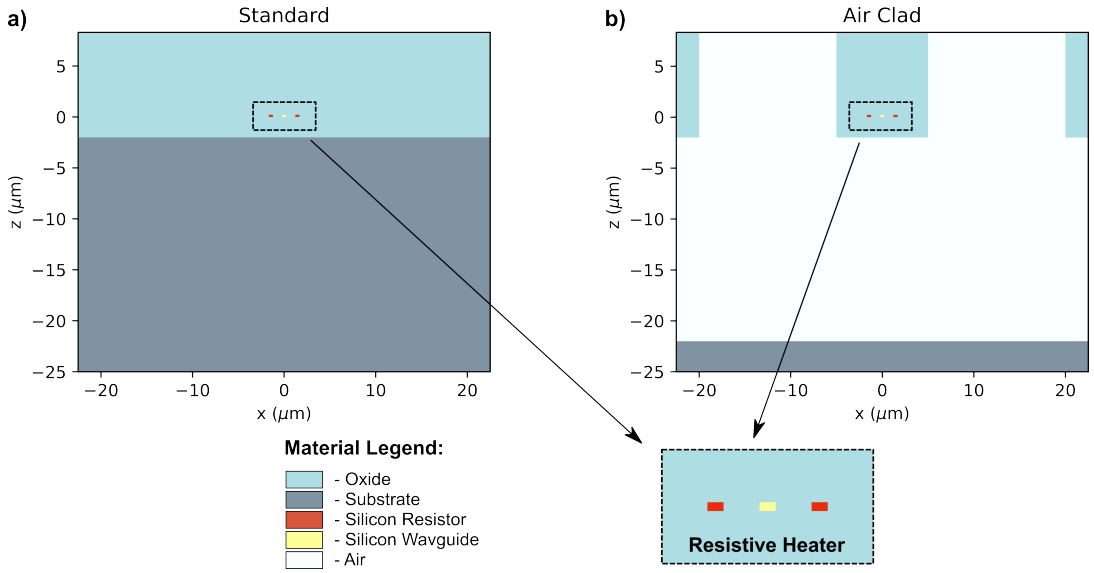

Figure 4.1: a) An x-section of the material stack for the standard thermal geometry of our phase shifters. This includes a silicon handle, buried and upper oxides, and a resistive heater defined by silicon and doped silicon resistors. b) The air-clad x-section shows a geometry wherein the handle and two pockets have been removed to create isolation conditions. However, the actual thermal elements are not changed.

as a thermal insulator, which helps to confine the heat generated by the heater to the waveguide region, resulting in a more uniform temperature distribution and lower power consumption. Thermal undercut also improves the device's thermal resistance, enabling it to operate at higher temperatures without thermal breakdown. Fig 4.3 shows the stark difference in thermal concentration for the same input power (i.e., 5 mW) in the heater elements.

We use Lumerical's HEAT solver to design and confirm the thermal isolation effect of this heater. Our design estimates a clean undercut, as seen in Fig 4.3 b), which isolates the guiding portion of the light and thermal effects from the silicon substrate, which can act as a heat spreader. If not removed, as shown in a), the substrate will remove heat from the waveguide core and deliver heat to adjacent or nearby devices on the PIC as it spreads it from the heat source to any heat sink. For this device, we chose to focus on the

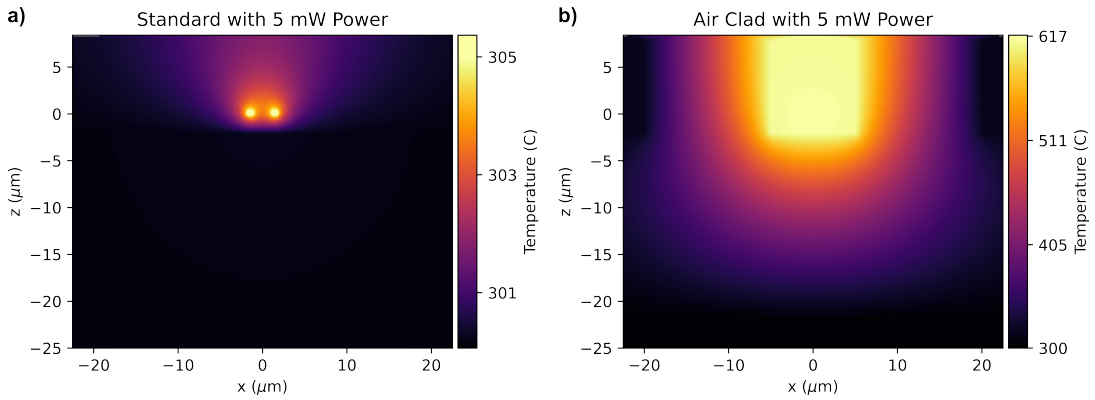

Figure 4.2: a) A thermal simulation of the standard heater design Fig. 4.1 a) with a five mW input power into the resistors, we see a peak temperature increase of 5° C. b) A thermal simulation of the air-clad heater design Fig. 4.1 a) with a five mW input power into the resistors, we see a relatively massive peak temperature increase of 317° C. Additionally, the heat fully occupies the thermally isolated region, demonstrating the lack of a thermal escape route.

process implementation over the optimal geometry of the thermal tuner. For example, the resistive heater has silicon rails on either side of the waveguide. However, more efficient results can be achieved with a single silicon rail or a thin silicon slab between the rail and waveguide []. This design is intended to have compatibility with low-signal optical circuits (such as neuromorphic or quantum applications). Therefore we prioritize optical loss and cross-talk over absolute thermal efficiency improvements.

Even though the design of this is perhaps not entirely new, we confirm the results in Fig 4.3 for forward action at the foundry level as we develop the wafer-scale process, using XeF_2 for the final substrate release.

4.3 Fabrication and Experimental Results

Integrating a thermal undercut into the process for silicon photonics using xenon difluoride (XeF_2) involves several steps. The thermal undercut region is defined by patterning

Figure 4.3: A before and after microscope image of the ICP RIE oxide removal and XeF2 silicon removal etches, we see on the right-hand side that the silicon handle is removed below the device,

the SiO2 layer using lithography and etching. The device is then placed in a XeF_2 vapor chamber at a temperature above the boiling point of XeF_2, which is approximately 50°C []. XeF_2 is a highly reactive gas that selectively etches away the glass in the thermal undercut region. After a sufficient amount of time (several hours), the device is removed from the XeF_2 chamber and rinsed with a solvent such as an isopropyl alcohol to remove any residue. Finally, the device is annealed at a high temperature (e.g., 1000°C) to remove any remaining glass and smooth out the thermal undercut region [].

The thermal undercut created by this process can suspend the silicon photonic device above the buried oxide layer, reducing optical losses and improving device performance. This process creates a well-defined and controlled thermal undercut, an essential step in fabricating high-performance silicon photonic devices. XeF2 enables selective etching of the SiO2 layer while leaving the silicon layer intact. Moreover, this process is compatible with standard lithography and etching techniques, which makes it a straightforward and cost-effective method for creating thermal undercuts in silicon photonics [].

4.3.1 Extracting P_π and Frequency Roll-Off

To optically address the MZI, we used standard SMF28 to couple light onto the chip. The thermal phase shifters are routed to metal pads individually addressed with a DC

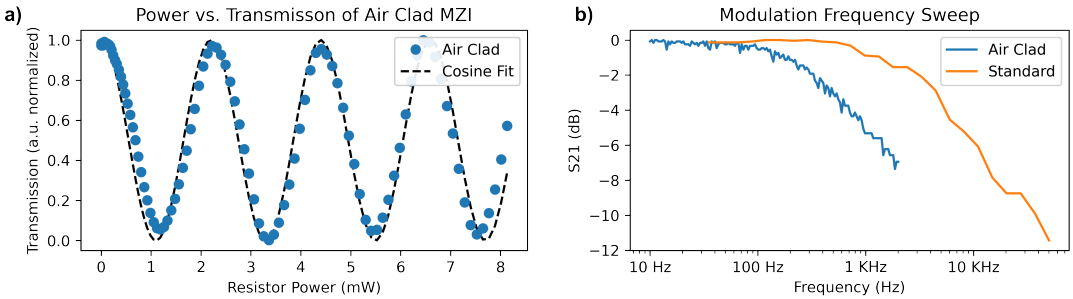

Figure 4.4: a) We extract the air clad P_π of the thermal undercut device and fit a cosine. We see a $P_\pi = 1.2$ mW. b) We measure the S21 response of the standard and air-clad devices and see that the air-clad has a much earlier roll-off of $f_{3dB} = 445$ Hz, compared to the standard device at $f_{3dB} = 4.5$ KHz.

probe. The resistors are designed with a resistance of 2.3 kΩ. We then apply a voltage bias and track the optical transmission through the fully air-clad, thermal-isolated MZI at a wavelength of 1550 nm. As the voltage bias increases, the transmission changes with a sine-squared relationship concerning power, as shown in Fig. 4.4 a).

The extracted P_π, which is the power required for a phase shift of π, is 1.2 mW, with a corresponding voltage of $V_\pi = 1.65$ V. This value of P_π is an essential parameter in characterizing the MZI performance. It represents the point at which the MZI produces a phase shift of π and is used to compare the performance of different MZIs. A lower P_π implies that the MZI requires less power to produce the same phase shift, which is desirable for applications that require low power consumption.

Fig. 4.4 b) shows the results of a modulation frequency sweep of the MZI, which characterizes the device's ability to modulate an optical signal at high frequencies. The standard cladding design exhibits a roll-off frequency of $f_{3dB} = 4.5$ KHz, whereas the air-clad MZI shows a significantly lower roll-off frequency of $f_{3dB} = 445$ Hz. The roll-off frequency is when the modulation response drops to half its maximum value. The

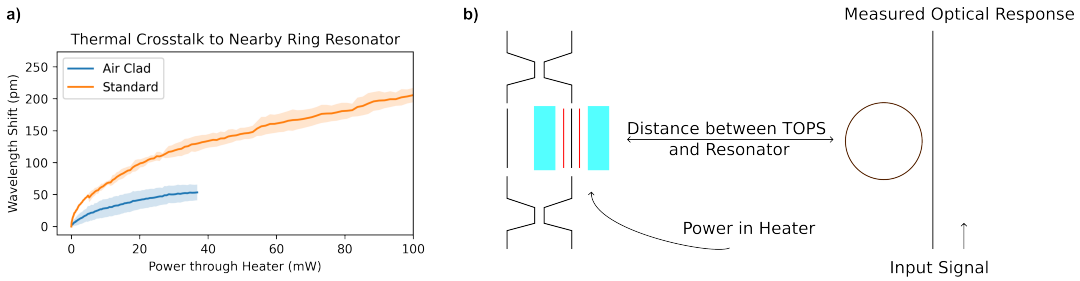

Figure 4.5: We measure the difference of thermal cross-talk between the tested phase shifter and a nearby highly sensitive ring resonator. We see that the resonance change of the ring resonator is affected much more by the power in the heater of the standard clad phase shifter (about a 2.5x improvement). Consider that the operating point of the air-clad device is around 1.1 mW and that of the standard device is 30 mW, and we see nearly a 30x improvement on the *operating* thermal cross-talk.

difference in the roll-off frequency between the two designs highlights the impact of the thermal conductivity of the cladding material and the silicon substrate (i.e., heat spreader) on the MZI's performance.

4.3.2 Thermal Effect on Nearby Resonator

The isolation of a thermal phase shifter can be determined by tracking the resonance peak wavelength of a nearby ring resonator, which is a highly temperature-sensitive device. An increasing range of voltages is applied to the thermal phase shifter while the ring resonator is optically probed. The resonance shift is then investigated as a function of heat in the resistor. This procedure is shown in Fig. 4.5 a).

The standard device is observed to shift the resonance more than the air-clad device for the same power, demonstrating the heaters' thermal isolation. The dramatic performance increase of the air-clad device is also noted, along with the difference in resonance shift at the operating power. The standard machine has a $P_\pi = 30$ mW and induces a 170 pm resonance shift. In contrast, the air-clad device has $P_\pi = 1.2$ mW and only

causes a resonance shift of 20 pm. This substantial difference in resonance shift between the two devices is attributed to the superior thermal isolation of the air-clad device.

4.4 Conclusions and Future Works

This study presents a thermally isolated phase shifter based on a resistive heater design that utilizes the thermo-optic effect to induce a phase shift in the optical waveguide. The device's performance is compared with a standard device cross-section, and the results show a significant improvement in heat generation and thermal isolation for the thermally isolated design. The device was fabricated on a 300 mm commercial foundry, and low-power consumption ($P_\pi = 1.2$ mW) and low thermal crosstalk were demonstrated.

Future work could focus on optimizing the performance of the thermally isolated phase shifter, including further reducing power consumption and thermal crosstalk, as well as improving the modulation speed and accuracy. Improving process scalability by performing this XeF_2 substrate etching in-foundry is an ongoing work with our research partners at AIM Photonics. Additionally, investigating the potential for integrating the thermally isolated phase shifter with other photonic components, such as modulators and detectors, could develop more complex and functional photonic circuits. Moreover, exploring the scalability of the thermally isolated design to larger-scale photonic circuits could enable the development of more advanced and integrated photonic systems.

Chapter 5

Wavelength Diverse Integrated Photonic Linear Neuron

5.1 Introduction

Current neuromorphic photonic architectures fall broadly into two main categories: wavelength coherent and incoherent [101]. Wavelength coherent designs rely on careful placement and tuning of Mach-Zehnder interferometers (MZI), such that optical interference provides multiplication and summation operations. Carefully designed MZI mesh circuits can implement any real-valued matrix on-chip by reconfiguring phase shifters according to singular value decomposition [104, 141]. The coherent optical linear neuron (COLN) architecture implements vector-vector multiplication and summation by leveraging MZIs as inputs and weights for multiplication and coherent interference as the summation operator [76, 123]. Wavelength incoherent designs rely on multi-wavelength operation, often employing a broadband photodetector to perform summation, but through this method, the phase of the signal is lost to optical-electrical conversion. For example, in

the "Broadcast & Weight" protocol, multiple wavelength channels are launched as inputs along the same waveguide, which allows for parallel weighting with micro-ring resonator weight banks [117]. A balanced photodetector pair sums the vector-vector multiplication; however, multiple micro-ring weight banks can operate in parallel for vector-matrix operation, as demonstrated in [116].

In this work, we present a novel architecture that leverages both coherent *and* incoherent aspects to create a massively scalable PNN-specific linear operator. This architecture requires fewer components, provides orders-of-magnitude footprint reduction, and consumes substantially less power than comparable designs. As an initial proof-of-principle, we experimentally demonstrate simple addition and subtraction on-chip. Additionally, we implement and demonstrate a neural network task on-chip designed to perform logic gate operations (AND, OR, and XOR). Due to the compact footprint, low energy consumption, and scalability of the demonstrated linear neuron, these results pave the way toward large-scale PNNs with hundreds to thousands of neurons on a single silicon chip.

5.2 Proposed Architecture

5.2.1 Background

We begin our design derivation with inspiration from the COLN [76]. The operating principle of the COLN employs four key stages: Fan-Out, Input, Weighting, and Fan-In, as shown in Figure 5.1(a). A continuous-wave (CW) optical carrier signal is split at the Fan-Out into N copies. Here, the Input and Weight stage is represented by amplitude

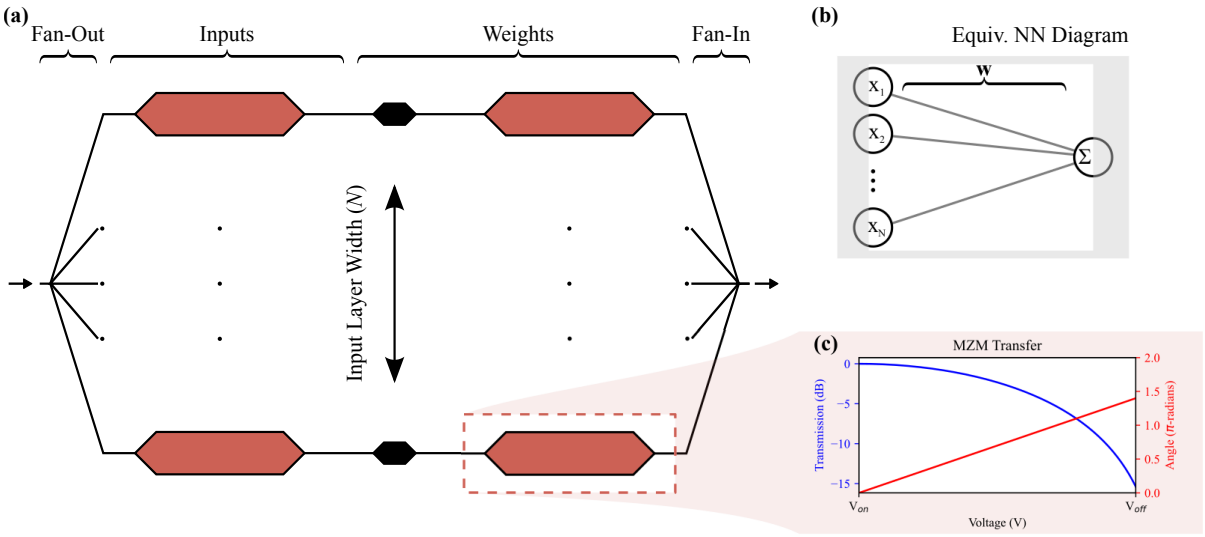

Figure 5.1: (a) The COLN architecture. The carrier signal is fanned out, where it meets an MZM that imparts the input and then meets another MZM that imparts the magnitude of the weight, with the phase shifter controlling the sign of the weight. Each bus is then recombined at the Fan-In stage, representing the vector-vector multiplication of the COLN. (b) An equivalent NN diagram for this circuit. The unshaded region represents the multiplication and summation performed by the COLN, ignoring the activation function which is external to this architecture. (c) A typical transfer function for an MZM. The extinction ratio (ER) and phase can vary with design.

modulators, which are implemented in hardware as Mach-Zehnder Modulators (MZM) and impose the inputs and weights onto the CW carrier signal. This signal is recombined from N paths down to 1 path, performing a summation of the N input/weight products at the Fan-In through constructive interference. Figure 5.1(b) demonstrates the equivalent neural network (NN) diagram for this circuit, which performs the linear summation of the inputs multiplied by the synaptic weights, preparing the carrier signal for the activation stage. This optical circuit and the NN both reduce to the general form:

$$out = \sum_{n}^{N} \mathbf{w}_n \mathbf{x}_n, \tag{5.1}$$

Where, for the optical circuit, *out* is the electric field output, N as the number of layers in the width (or Fan-Out), $\mathbf{w}$ is the weight vector, and $\mathbf{x}$ is the input vector [76]. In this formulation, $\mathbf{w}$ and $\mathbf{x}$ are represented by the electro-optic transfer function of the MZM in each layer similar to that seen in Figure 5.1(c). A vital feature of the COLN circuit is that the output is in a state which, with minimal additional passive circuitry, is fully compatible both with activation based on optical-electrical (OE), electrical processing, and electrical-optical (EO) conversion [136], direct optical-electrical-optical (OEO) activation [115] or all-optical, non-linear activation [55, 77].

In a fiber-optic-based system, phase shifters and splitters are readily available off-the-shelf [76]. However, when attempting to scale these down for an integrated photonic system, the size of MZM translates into a practical barrier. Current PN diode, free-carrier MZM devices each occupy a physical space of > 1 mm^2 in the on-chip area, severely limiting the scalability of this architecture. The first on-chip demonstration of the COLN in a silicon photonic chip utilizes a 25 mm^2 chip to implement a 4 input/weight linear neuron, implementing the weights with thermal phase shifters, which are dramatically more minor at the cost of update speed [39, 75].

5.2.2 WDIPLN Introduction

Our circuit architecture, which we subsequently refer to as Wavelength Diverse Integrated Photonic Linear Neuron (WDIPLN), begins as a COLN with naïve replacement of MZMs with micro-resonant devices (MRD), either micro-ring resonators or micro-disks, as shown in Figure 5.2(a). With MRDs, we immediately benefit from the decrease in footprint.

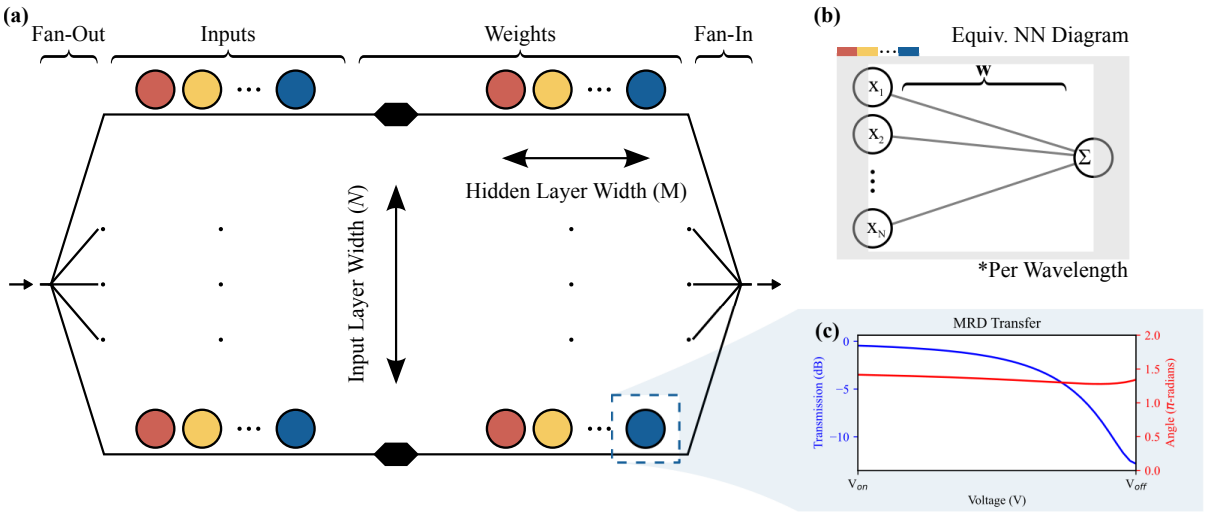

Figure 5.2: (a) The naïve WDIPLN architecture. The carrier signal is fanned out, where it meets an MRD that imparts the input and then meets another MRD that imparts the magnitude of the weight, with the phase shifter controlling the sign of the weight. Each bus is then recombined at the Fan-In stage, representing the vector-vector multiplication of the WDIPLN. (b) An equivalent NN diagram for this circuit. The unshaded region represents the multiplication and summation performed by the WDIPLN, ignoring the activation function which is external to this architecture. The tabs in the upper left-hand side represent the WDIPLN's ability to represent M many-to-one networks shown here. (c) A typical transfer function for an MRD. The ER and phase can vary with design; however, in the "slightly under coupled" regime, we will see a relatively flat phase response.

Single-channel MRDs are commonly $< 100\ \mu\text{m}^2$ in size [122]. In addition, careful design of the MRD allows for a desirable adaptation of the previous electro-optic transfer function, displayed in Figure 5.2(c). The MRD acts as an amplitude filter in the network's input and weight stage for each bus in the layer width. MRDs are sensitive, but if the coupling condition is designed to be slightly under-coupled, it is possible to mitigate some of the phase sensitivity (for example, a depletion-biased MRD would remain under-coupled during operation). This enables the coherent summation at the Fan-In stage to remain relatively stable for each channel.

Naïve WDIPLN introduces a feature beyond the COLN, namely wavelength diversity. For sufficiently small radii MRD (or, conversely, sufficiently large free-spectral-range (FSR)), we can place additional MRDs at new resonant wavelengths (or along specific channels, similar to wavelength-division-multiplexing (WDM) in optical communications). For example, in silicon photonics, a sufficiently large channel spacing has been demonstrated with a 2.5 μm radius microdisk modulator, with > 55 μm, full width at half maximum of ~ 0.2 nm, and high-quality factor ($> 8,000$), which indicates high channel isolation [122].

Recently, a similar WDM scheme was proposed using MZMs, but the WDIPLN implementation with ring resonators has the advantage that the wavelength multiplexing/demultiplexing is done intrinsically using the single device in a compact footprint [123]. We show a schematic for the naïve WDIPLN in Figure 5.2(a). This enables wavelength-diverse operation of the coherent circuit *at each wavelength channel* such that we have an incoherent circuit that acts like a coherent operator at each channel (λ).

5.2.3 WDIPLN Detailed Derivation

Here, we pause to work through a derivation of a simple case for the WDIPLN, as shown in Fig. 5.3, which has two rings in the input column, and two rings in the output column. First, we set up the transfer matrix based on Fig. 5.3.

$$\begin{bmatrix} E_3 \\ E_4 \end{bmatrix} = \underbrace{\frac{1}{2}\begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix}}_{\text{50:50 splitter}} \underbrace{\begin{bmatrix} \frac{t_3 - Ae^{i\phi_3}}{1 - At_1e^{i\phi_3}} & 0 \\ 0 & \frac{t_4 - Ae^{i\phi_4}}{1 - At_1e^{i\phi_4}} \end{bmatrix}}_{\text{rings 3 \& 4}} \underbrace{\begin{bmatrix} e^{i\theta_1} & 0 \\ 0 & e^{i\theta_2} \end{bmatrix}}_{\text{phase shifters}} \underbrace{\begin{bmatrix} \frac{t_1 - Ae^{i\phi_1}}{1 - At_1e^{i\phi_1}} & 0 \\ 0 & \frac{t_2 - Ae^{i\phi_2}}{1 - At_1e^{i\phi_2}} \end{bmatrix}}_{\text{rings 1 \& 2}} \underbrace{\begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix}}_{\text{50:50 splitter}} \underbrace{\begin{bmatrix} E_1 \\ E_2 \end{bmatrix}}_{\text{inputs}} \tag{5.2}$$

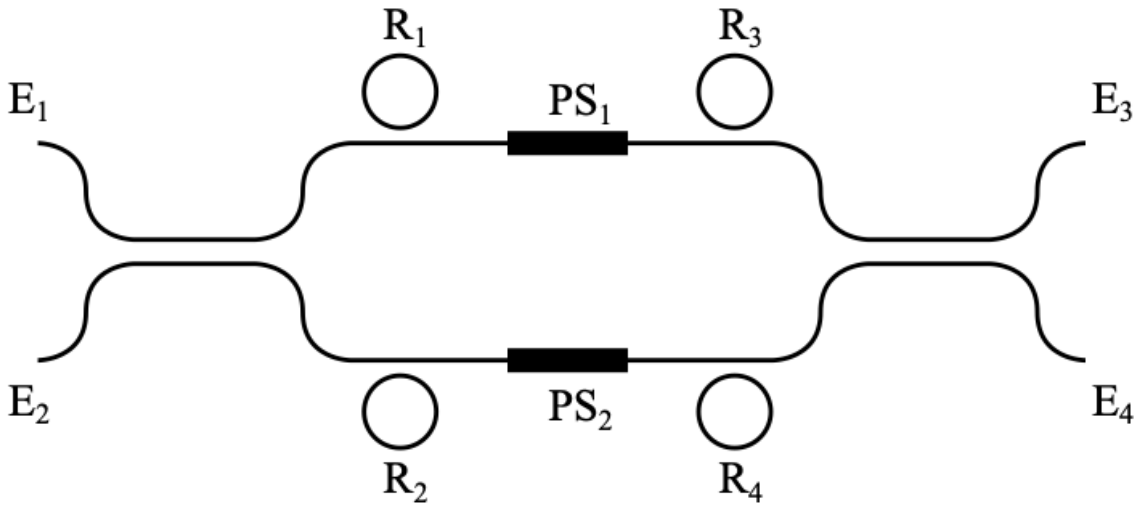

Figure 5.3: Inputs, E_1, E_2 feed into a directional coupler, which splits the light equally into paths containing a ring, phase shifter, and a second ring (either R_1, PS_1, R_3 or R_2, PS_3, R_4), then recombines the light in a second directional coupler before reaching the outputs, E_3, E_4. This circuit represents the simplest formulation of the proposed WDIPLN.

Instead of working this through with the specific ring values, we can make this more intuitive by utilizing a substitution for each ring component:

$$R_1 = \frac{t_1 - Ae^{i\phi_1}}{1 - At_1e^{i\phi_1}},\ R_2 = \frac{t_2 - Ae^{i\phi_2}}{1 - At_1e^{i\phi_2}},\ R_3 = \frac{t_3 - Ae^{i\phi_3}}{1 - At_1e^{i\phi_3}},\ R_4 = \frac{t_4 - Ae^{i\phi_4}}{1 - At_1e^{i\phi_4}}. \tag{5.3}$$

Here, we point out that utilizing the ring resonator instead of an MZI creates benefits and caveats. We've noted the benefits (reduced footprint, WDM) but want to address the caveats. Specifically, the phase detuning is extremely sensitive to the ring coefficients. The voltage required to apply the weights results from the ring resonator's inherent quality. Finding the appropriate voltage range for weight application versus ring quality will be part of the experimental endeavor. However, within a well-designed MRR modulator, the quality factor should not be so high that we cannot reliably reconfigure the weights or so low that we cannot leverage the channel multiplexing inherent in the design. If

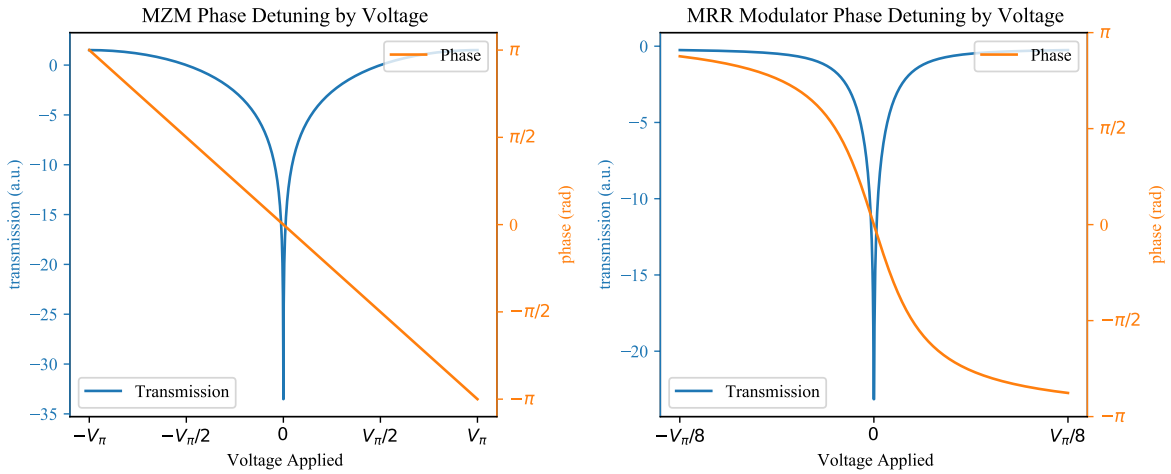

Figure 5.4: a) The phase detuning by the voltage of an MZM. We see transmission in blue and phase argument in orange. b) The phase detuning by the voltage of an MRR modulator. We see transmission in blue and phase argument in orange. While these two are different, the transmission characteristics are quite similar, and the phase arguments are always decreasing – even if there is a sigmoidal shape to the MRR. Both provide a mechanism for applying the varied range of weights via a voltage input – all of which will have transmissions that peak at a given voltage (i.e., Voltage = 0) and decrease in phase. Additionally, V_π is generally much higher for MRR modulators, such that the actual voltage applied for these two plots will be similar

designed well, the phase detuning of an MRR modulator can mimic the weight application mechanism available in the phase detuning of an MZI for a given voltage applied – shown in Fig. 5.4 Returning to the derivation, with that substitution, we can rewrite the transfer matrix as:

$$\begin{bmatrix} E_3 \\ E_4 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} E_1(R_1R_3e^{i\phi_1} - R_2R_4e^{i\phi_2}) + E_2(R_1R_3e^{i\phi_1} - R_2R_4e^{i\phi_2}) \\ iE_1(R_1R_3e^{i\phi_1} + R_2R_4e^{i\phi_2}) + iE_2(R_1R_3e^{i\phi_1} - R_2R_4e^{i\phi_2}) \end{bmatrix} \begin{bmatrix} E_1 \\ E_2 \end{bmatrix} \tag{5.4}$$

And, let's assume we only input in port 1; we further reduce to:

$$\begin{bmatrix} E_3 \\ E_4 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} E_1(R_1R_3e^{i\phi_1} - R_2R_4e^{i\phi_2}) \\ iE_1(R_1R_3e^{i\phi_1} + R_2R_4e^{i\phi_2}) \end{bmatrix} \begin{bmatrix} E_1 \\ E_2 \end{bmatrix} \tag{5.5}$$

And finally, let's isolate the output at port 3,

$$E_3 = \frac{E_1}{2}(R_1 R_3 e^{i\phi_1} - R_2 R_4 e^{i\phi_2}) \tag{5.6}$$

$$E_3 = \frac{E_1}{2}(R_1 R_3 \pm R_2 R_4) \tag{5.7}$$

We see a strong correlation between this form and the initially proposed COLN:

$$E_{out} = \frac{E_1}{2}[w_1 x_1 e^{i\phi_1} + w_2 x_2 e^{i\phi_2}] \tag{5.8}$$

$$E_{out} = \frac{E_1}{2}[w_1 x_1 \pm w_2 x_2] \tag{5.9}$$

For a better understanding of the operation conditions, we can present this form,

$$E_3 = \frac{E_1}{2}(R_1(\lambda) R_3(\lambda) \pm R_2(\lambda) R_4(\lambda)) \tag{5.10}$$

The strength of this approach is the ability a) to scale back on size and b) to enable multiplexing. It is trivial to extend the above derivation to the case where N rings replace each ring, or we wish to operate with N wavelengths. However, we would be better served to define the first column of rings (before the phase shifter) as the data inputs, i.e., $x_1, x_2 \ldots, x_N$, and the second column of rings as the weights, i.e., $w_1, w_2, \ldots, w_N$.

This would update our image to be represented by Fig. 5.5 and turn our equation into

$$E_3 = \frac{E_1}{2}(R_1(\lambda) R_3(\lambda) \ldots R_{N-1}(\lambda) \pm R_2(\lambda) R_4(\lambda) \ldots R_N(\lambda)), \tag{5.11}$$

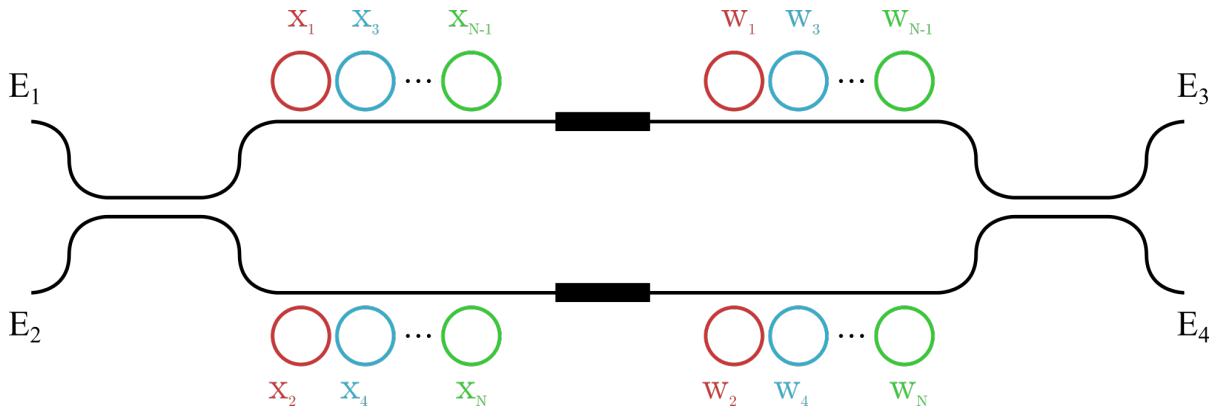

Figure 5.5: A multiplexed version of the WDIPLN, which shows that inputs ($x_{1,2,N}$)can be brought into the first column of rings at any wavelength, and then weights ($w_{1,2,N}$) can be applied in the second column of rings at any wavelength.

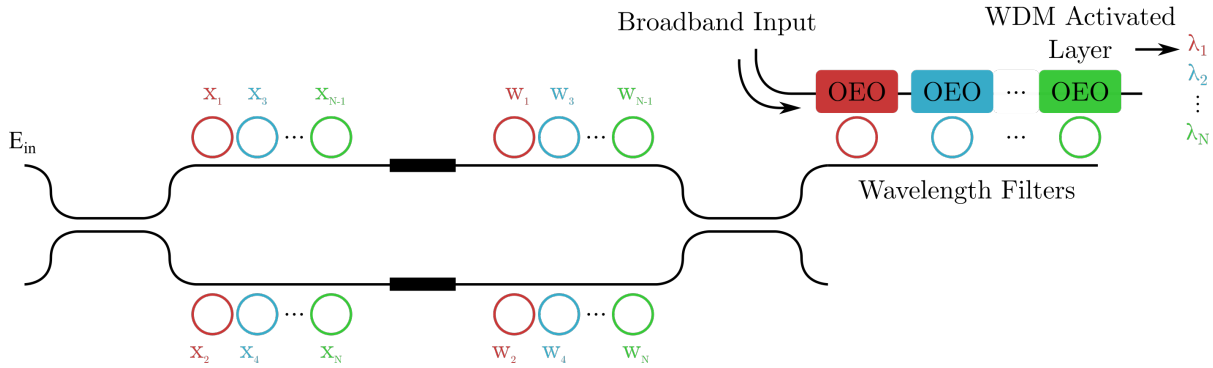

Figure 5.6: The WDIPLN can be fed into a bank of nonlinear operators, allowing this circuit to act as multiple many-to-one (or one many-to-many) linear networks.

However, a more useful way to think of this is that, at each wavelength, we have the simplified version, such that:

$$\text{at } \lambda_{1-2} : E_{out} = \frac{E_{in}}{2}(x_1 w_1 \pm x_2 w_2)$$

$$\text{at } \lambda_{3-4} : E_{out} = \frac{E_{in}}{2}(x_3 w_3 \pm x_4 w_4)$$

$$\dots$$

$$\text{at } \lambda_{N-1-N} : E_{out} = \frac{E_{in}}{2}(x_{N_1} w_{N-1} \pm x_N w_N)$$

In addition, we can show more directional couplers, i.e., more nested MZIs, scales in a similar way to the COLN, wherein we can now add a multiplexing degree of freedom to the number of weights degree of freedom. For the COLN, we see this (for $L = 2m$ layers) as:

$$E_{out} = \frac{1}{2^{2m+1}} E_{in} \sum_{j=1}^{2^{2m+1}} x_j w_j e^{i\phi_j}$$

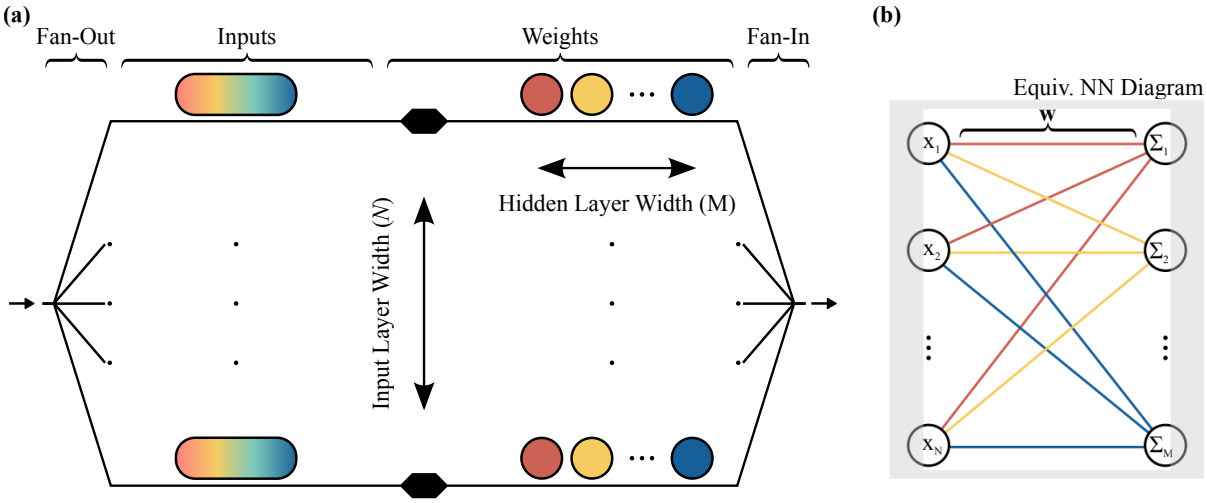

Figure 5.7: (a) The fully formed WDIPLN architecture. The carrier signal is fanned out, meeting a large MRD that simultaneously imparts the input to each channel. Each channel's input then meets a small MRD, which imposes the magnitude of the weight, with the phase shifter controlling the sign of the weight. Each bus is then recombined at the Fan-In stage, representing the vector-matrix multiplication of the WDIPLN. (b) An equivalent NN diagram for this circuit. The unshaded region represents the multiplication and summation performed by the WDIPLN, ignoring the activation function external to this architecture. This architecture fully enables the linear stage of the MLP layer.

5.2.4 Full WDIPLN Formalism

Finally, we can rewrite the WDIPLN formalism as an extension of the COLN at each channel, simplified and presented as an adaptation of Eq. 5.1 to

$$out_\lambda = \sum_n^N \mathbf{w}_{\lambda,n} \mathbf{x}_{\lambda,n}, \tag{5.12}$$

We can claim Equation 5.12 for WDIPLN design with sufficiently isolated channels, where the interaction of each resonance occurs only in and around that wavelength. As discussed, this is achieved using sufficiently small MRDs with large FSRs. Using isolated wavelength channels enables the resonant-based circuit to act as any m individual COLN circuit for $m \in M$, where M here represents the number of wavelength channels in operation *and* the number of COLN circuits represented by a single WDIPLN circuit. Therefore, this design not only leverages smaller footprint devices but enables further footprint reduction of a given system of linear neurons from the limit of one optical circuit per perceptron to one optical circuit per fully connected multi-layer perceptron (*MLP*) by using the input/weight pairs at each wavelength as shown in Figure 5.2(b).

In the naïve case, we have simply swapped each MZM with an MRD, such that the circuit requires $2NM$ MRDs, for layer width N and wavelength channels (or layer depth) M. This case is the same as using M COLN circuits, equivalent only to abstracted device count complexity. Thus, an actual network with COLNs would require M copies of the circuit resulting in a larger implementation. However, we recognize that the input stage MRDs are redundant. Therefore, to realize *full* WDIPLN, we replace all M input MRDs for a given layer with a carefully designed, larger radius MRD with an FSR that matches the channel spacing, as outlined in Figure 5.7(a). This adjustment lowers the total device count to $N(M+1)$, which represents a dramatic improvement in device count scalability in addition to the physical size. As seen in Figure 5.7(b), the WDIPLN circuit in this form represents the linear weighting and summation stage of a fully connected *MLP*. A caveat of the fully-realized WDIPLN architecture is the inability to realize phase shifting per weight. In this way, we must limit the weight attribution

Architecture (variation)	COLN (nominal)	COLN (w/ thermal MZI)	WDIPLN (naïve)	WDIPLN (nominal)
Input Element	MZM	MZM	Small MRD	Large MRD
Approx. Element Size (mm^2)	0.8	0.8	10^{-4}	10^{-2}
Weight Element	MZM	Thermal MZI	Small MRD	Small MRD
Approx. Element Size (mm^2)	0.8	10^{-1}	10^{-4}	10^{-4}
Scaling Rule	$2NM$	$2NM$	$2NM$	$N(M+1)$
Physical Size (mm^2)	Component Size Only (no routing considered)			
$N=8, M=1$	12.8	7.2	1.6×10^{-3}	8.08×10^{-2}
$N=8, M=8$	102.4	57.6	1.28×10^{-2}	8.64×10^{-2}
Electrical I/O	Consider 4 Electrical I/O per element (2 EO, 2 Thermal)			
$N=8, M=1$	64	64	64	64
$N=8, M=8$	512	512	512	288
Power Consumption (mW)	Same Configuration (2 EO, 2 Thermal)			
EO Power - ($P_{On/Off}$)				
$N=8, M=1$	2.72	1.36	1.6	0.48
$N=8, M=8$	21.76	10.88	12.8	3.84
Thermal Power - (P_{π})				
$N=8, M=1$	89.6	89.6	89.6	25.2
$N=8, M=8$	716.8	716.8	716.8	201.6

Table 5.1: Table comparing the primary differences between the COLN and WDIPLN architectures. We consider the nominal COLN, a COLN where weights are done with thermal MZIs, the naïve WDIPLN and nominal WDIPLN. For physical size calculations, we only consider the footprint of individual elements and recognize that additional optical and electrical routing will be necessary for each design. The physical size and power consumption values were estimated from available foundry PDKs in [32, 40, 64].

to $[-1, 0]$ or $[0, 1]$. The phase shifters in this architecture are here as a global change to all weights simultaneously, serving both as path balancing and switching between subtractive or additive behavior. Table 5.1 summarizes the key differences between the COLN and WDIPLN, including scaling rules, physical size, electrical I/O, and power consumption estimates from available foundry process design kits (PDKs) [32, 40, 64]. We note that the physical size is calculated purely as fundamental device area. However, both the COLN and WDIPLN require additional optical and electrical routing to connect to the outside world. This routing footprint was omitted from the total since it is highly implementation-specific and depends heavily on packaging constraints (i.e., edge coupled versus grating coupled, flip-chip bonded or wire-bonded, etc.). However, is is clear that

the extremely small device footprint of the WDIPLN architectures leaves ample room for additional circuitry, such that optical couplers and electrical bond pads will dominate the final device area. For example, 288 electrical pads (60 μm $\times$ 60 μm) on a 16×18 grid at 150 μm spacing take up 6.48 mm^2 in on-chip area. If we add optical couplers (400 μm $\times$ 20 μm) on the edge of the chip at a large pitch for ease of packaging, the total on-chip area is < 7 mm^2 for the WDIPLN of $N = 8, M = 8$. In addition, we note the power consumption values are also estimated from available information in [40], which utilize thermal isolation methods to achieve a $P_\pi \sim 2.8$ mW. Importantly, these values illustrate the improved power budget of the WDIPLN architecture – regardless of technology implementation.

5.2.5 A Note on Bias

We have not discussed the bias term in any detail here, as it is not a strictly necessary component in this circuit. However, the bias term can be a crucially enabling feature of a given neural network, and it has a ready-made physical implementation in the context of an optical circuit. But, if we place a splitter before we reach the WDIPLN as described above, and add in a phase shifter, and amplitude filter, then we have reformed the expression in Eq. 5.12,

$$out_\lambda = \sum_n^N \mathbf{w}_{\lambda,n}\mathbf{x}_{\lambda,n} + \mathbf{b}_\lambda, \tag{5.13}$$

for a bias term $\mathbf{b}$. Here, we show that the standard notation for the MLP and optical circuit is again equivalent with the addition of the bias. The bias is neither necessary nor sufficient, but it can be quite useful. The bias acts in two key ways for the neural

network architecture we are interested in exploring here.

First, the optical signal of the WDIPLN circuit is enhanced by the bias to showcase negative values in the detected output, which is performed by a photodetector. Photodetectors detect the magnitude squared (i.e., the intensity of the electric field) such that we lose the phase. While the benefits of this detection method abound, it presents this particular circuit with an odd consequence – we cannot differentiate between addition and subtraction (we go into some more detail below on demonstration details). When we add a bias, optically speaking, we provide a reference for a "ground" signal. This allows us to reflect negative and positive values above and below this ground. As the phase of the bias is changed from 0 (in phase) to pi (out of phase), we go from fully constructive interference to fully deconstructive interference, or addition and subtraction, respectively. In addition, we can modify the amplitude of this bias value. With this addition, we can probe the circuit and set the operative point, which is useful in a complicated circuit considering fabrication variations.

Secondly, including the bias is mathematically advantageous for the linear neuron as a physical embodiment of the MLP. In the absence of a bias, the network is restricted to operating in the first quadrant of the real value domain, with the operation being confined primarily to either addition or subtraction. This considerably restricts the operative scope of the network, necessitating an increase in network depth and width and preprocessing of the data. As an aside, optically compatible preprocessing methods such as Fourier transforms, spatial light modulation and spatial filtering can be performed passively after configuring, reducing the resources involved in this step. However, for MLPs, generally speaking, a bias will improve the operating scope of the network while

also increasing the resources required.

5.3 Experimental Demonstration of Simple Addition

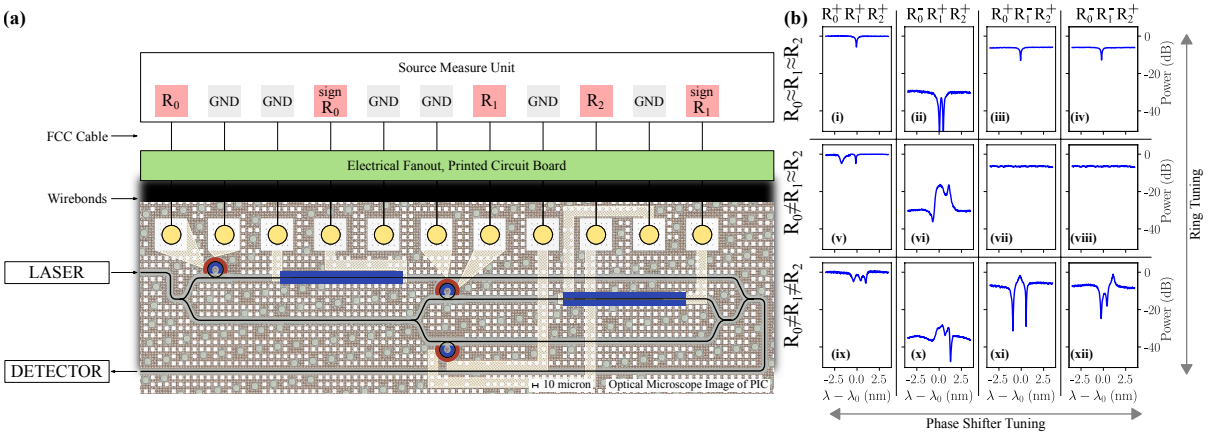

Figure 5.8: (a) Experimental setup. We optically couple the laser and detector to the photonic integrated circuit (PIC) using SMF28 fiber. The PIC consists of a simplified WDIPLN circuit with a bias line, for a total of 3 MRDs (R_0, R_1, R_2) and two-phase shifters $(\text{sign}(R_0), \text{sign}(R_1)$. All of the splitters/combiners are designed for an equal splitting ratio. The electrical traces connect the device terminals to the bond pads, which are wire bonded from the PIC to an electrical fanout and then connected to a printed circuit board. The printed circuit board is routed to a source measure unit (SMU) through a flat conductor cable (FCC). The SMU enables electro-optic control of each device. (b) Demonstration of addition and subtraction. The columns here represent the tuning of the phase shifters in (a), and the rows represent the tuning of the MRDs. The four columns show the following behavior: $+R_0 + R_1 + R_2$, $-R_0 + R_1 + R_2$, $+R_0 - R_1 + R_2$, $-R_0 - R_1 + R_2$. The three rows demonstrate states where $R_0 \sim R_1 \sim R_2$, $R_0 \neq R_1 \sim R_2$, $R_0 \neq R_1 \neq R_2$. The small wavelength range is kept to provide a "big-picture" of the circuit behavior. However, we evaluate each operation at $\lambda = \lambda_0$.

As an initial demonstration, we show the ability of WDIPLN to add and subtract in a simplified form. This represents a fundamental building block of the architecture. While tuning the MRDs ultimately imposes the input/weight onto the circuit, each of the N buses in the circuit are additionally reconfigurable via a phase shifter. The phase shifter rotates the phase of a given bus so that we are able to express the sign coefficient of the weights as either -1 or 1. The summation operation occurs as optical interference,

which will "add" (constructively interfere) when the phase is the same and "subtract" (destructively interfere) when the phase is opposite. We consider the optical experiment in Figure 5.8(a) for an initial demonstration. We designed a WDIPLN circuit with an added bias. There are five active circuit elements: the bias ring, R_0, bias phase shifter, PS_0, top arm ring R_1, top arm phase shifter PS_1 and bottom arm ring R_2. Each element is electrically connected via a wire bond to an electrical fanout, which is, in turn, wire-bonded to a printed circuit board (PCB) and connected to a source measure unit (SMU). The circuit is optically coupled from two single-mode fibers (SMF28) embedded in a titled fiber array to two grating couplers on either side of the circuit. A laser and detector pair is connected to the other ends of the respective SMF28 fibers. The MRD we employ is a carrier injection-based PIN micro-ring resonator. We employ a PIN for proof of principle due to the large wavelength change. However, future systems will utilize application-specific, optimized MRDs operating in carrier depletion mode. The photonic integrated circuit (PIC) was fabricated in American Institute for Manufacturing (AIM) Photonics' 300 mm silicon photonics process [32].

We demonstrate a variety of circuit configurations as summarized in Figure 5.8(b). We achieve these configurations by tuning the phase shifters to 0 or π (columns) and setting the rings at various resonant wavelengths (rows), for which we indicate the state of each ring by the label of R_0, R_1, or R_2. The bias path is slightly shorter than the internal WDILPN, which means we can utilize the phase shifter and the natural wavelength-dependent interference to extract the set-point behavior. Sub-figures (*i, v, ix*) show three states of addition with phase shifters tuned to 0 where the rings are all aligned to a single resonance ($R_0 = R_1 = R_2$, *i*), R_1 is tuned off ($R_1 \neq R_0 = R_2$, *v*), and all

the rings are at different resonances ($R_0 \neq R_1 \neq R_2$, *ix*). These states manifest as one, two, or three resonance peaks. Sub-figures (*ii, vi, x*) demonstrate the corresponding states where the bias phase is at π rather than 0. Due to manufacturing variations, we observe behavior that deviates from the designed point. In (*ii*), we expect no signal to pass through; however, due to path and y-branch imbalance, the circuit floor is ~ -30 dB—Additionally, small differences in the ring are seen in the transmission. Therefore, the maximum value in (*ii*) is ~ -33 dB at λ_0, where the resonant wavelength sits, which indicates a slight imbalance but overall high suppression of any difference in the rings. Sub-figure (*vi*) shows two peaks above the minimum, as $R_0 \neq R_1 = R_2$, and (*x*) shows three distinguishable peaks as $R_0 \neq R_1 \neq R_2$. The peaks overlap such that distinguishing individual peaks is difficult. Sub-figures (*iii, iv*) demonstrate a single resonance, similar to (*i*); however, note the maximum power is rough -6 dB since no light passes through in the lower path and each y-branch contributes additional loss of 3 dB. Sub-figures (*vii, viii*) show detuned R_0, and $R_1 = R_2$, where these rings fully cancel out each others' magnitudes (subtraction). Sub-figures (*xi, xii*) demonstrate $R_0 \neq R_1 \neq R_2$ for the subtraction operator, which is seen by the distinct peak difference above and below the bias line. Finally, R_1, R_2 each flip signs between (*xi, xii*) due to the phase control of the bias line.

5.4 Experimental Demonstration of Logic Gates

We designed a second circuit to demonstrate a learning task. This design reflects the WDIPLN architecture for $N = 2, M = 1$, as seen in Figure 5.9(a). This circuit places

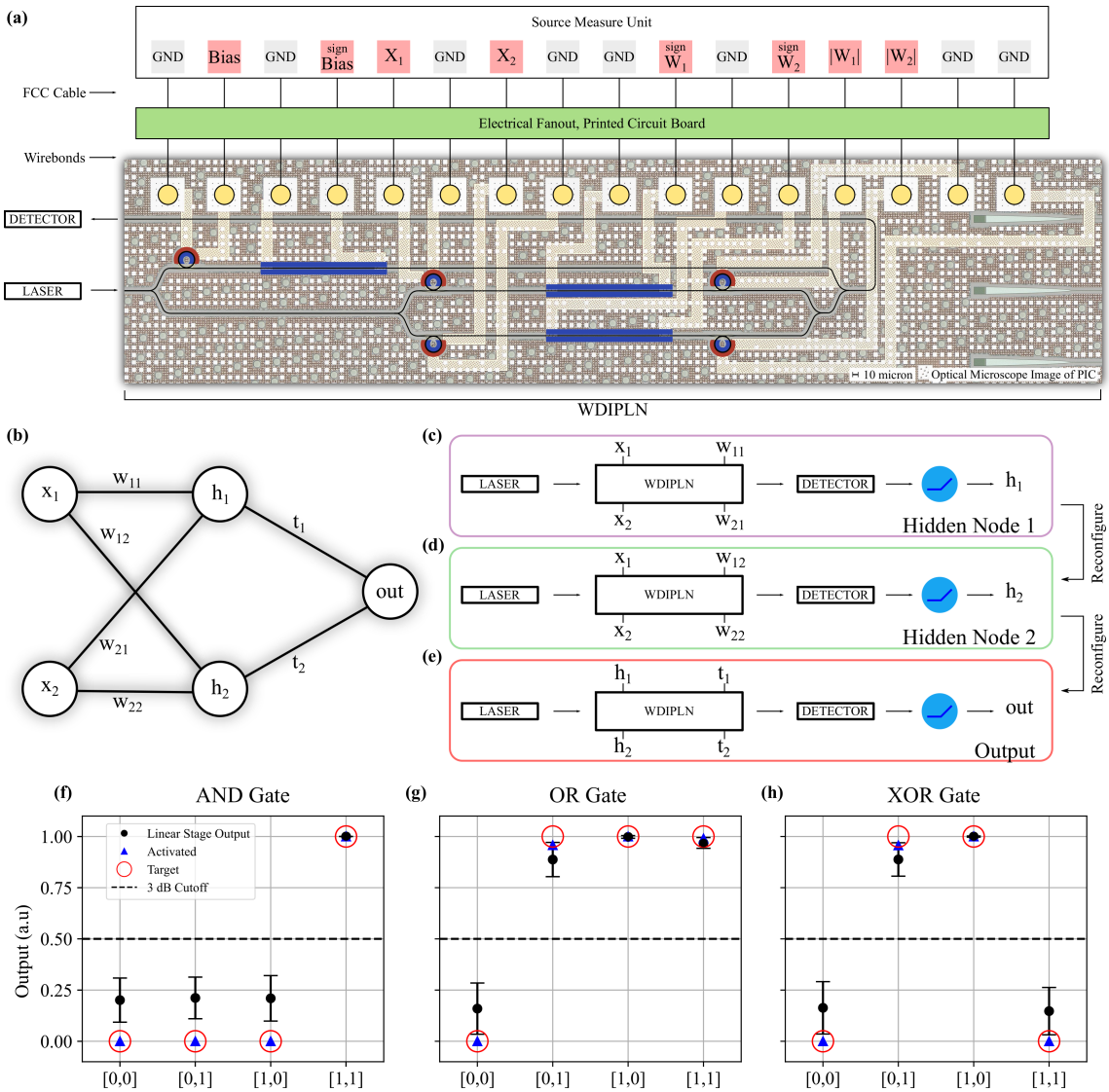

Figure 5.9: Photonic Neural Network experimental demonstration. (a) Experimental set-up, similar to Figure 5.8. We optically couple the laser and detector to the PIC. The PIC comprises a WDIPLN circuit with $N = 2, M = 1$, and a bias line for 5 MRDs and 3 phase shifters. All splitters/combiners are designed for equal splitting ratio.The electrical wiring goes out the bond pads, which are wire bonded from the PIC to an electrical fanout and then connected to a printed circuit board. The printed circuit board is routed to a source measure unit (SMU) through a flat conductor cable (FCC). The SMU enables electro-optic control of each device. (b) The selected architecture for the simple neural network we implement for this task. Inputs (x_1, x_2) are connected by a weight matrix, $\mathbf{w}$, to a hidden layer with nodes h_1, h_2. From here, a simple weight vector, $\mathbf{t}$, connects the hidden layer to the output. (c–e) The configure-recycle process for the circuit in (a). We send in the input pairs $[0,0], [0,1], [1,0], [1,1]$ as x_1, x_2 for the first two stages and subsequently send in h_1, h_2 in the final stage. We configure the weights of the WDIPLN to match that in (b) step-by-step from pairs $[w_{11}, w_{21}]$, $[w_{12}, w_{22}]$, and $[t_1, t_2]$ for each stage of the network, respectively. Each configure-to-measurement cycle takes approximately 1 second, which is dominated by the read/writes speeds of the SMUs. (f–h) Results of the experimental demonstration. The AND, OR, and XOR gates correctly predict the outputs with accuracy $96.8\%, 99\%$, and 98.5%, respectively.

both input and weight rings along the inner buses, along with a bias branch that also contains a ring. We wire-bond out to an electrical carrier, connecting to programmable SMU channels to electrically control each element. We optically couple the laser and detector through fiber to grating couplers on the input and output waveguides.

We selected a simple network architecture that can reconfigure the weights to represent AND, OR, and XOR gates, shown in Figure 5.9(b). We trained these networks on a CPU using PyTorch [87]. We employ the rectified linear unit (ReLU) for the activation function for all stages. We can create a simple map from the trained model to our physical chip by reconfiguring the chip in between the three linear stages, leading to the three non-input nodes h_1, h_2, and *out*. For the ring resonators, we define a value of "0" as $V_{off} = 1.4$ V and "1" as $V_{on} = 1.2$ V. This ensures we are operating in the slightly undercoupled regime for both "0" and "1". For this demonstration, the V_{off} and V_{on} values are globally set for all rings. In addition, the bias ring is set to V_{off} and the bias phase shifter is set to $\pi/2$, specifically for visibility of the full range of $[-1, 1]$ through direct detection as opposed to inference [76].

The process of reconfiguration and implementation of the two hidden nodes and the output is shown in Figure 5.9(c–e). For the two hidden nodes, h_1 and h_2, we configure the weights of the optical circuit, namely $|W_1|$ and $|W_2|$, according to the pairs of $[w_{11}, w_{21}]$ and $[w_{12}, w_{22}]$, respectively. In addition, if the sign of the weight is negative, we adjust the corresponding phase shifter, namely $\text{sign}(W_1)$ and $\text{sign}(W_2)$, from 0 to π. Once we configure the chip for h_1, we feed the four input pairs $[0, 0], [0, 1], [1, 0], [1, 1]$ into X_1, X_2. We measure the output from the detector for each input, apply the (ReLU) activation function used in training with a "3 dB cutoff", and record the result. After measuring

the two hidden nodes, we reconfigure the circuit for the final stage, according to the final weight stage from training t_1, t_2. Finally, we feed the four recorded pairs of h_1, h_2 into the circuit as inputs, measure the output, apply the activation function, and record the result.

According to the training-to-implementation scheme described above and in Figure 5.9(c–e), we demonstrate the circuit's performance for three different 2-bit logic gates: AND, OR, and XOR. In training, the accuracy of this simple network reaches $\sim 100\%$. The resulting outputs of the three gates are shown in Figure 5.9(f–h). The black dots are the square root (i.e., magnitude) of the output values measured by the detector at the laser probe wavelength, which in this experiment was globally set to $\lambda_0 = 1,526$ nm. The error bars represent the standard output deviation for a 100 pm window around $\lambda_0 \pm 50$ pm. The blue triangles show the circuit outputs after activation, and the red circles show the target for the gates. The AND, OR, and XOR gates achieve predictive accuracy of $96.8\%, 99\%$, and 98.5%, respectively.

5.5 Experimental Demonstration of Four Logic Gates

Having demonstrated the one-logic-gate-at-a-time circuit, we wanted to do a larger demonstration of the WDIPLN, which exhibits the wavelength capability of the architecture. Therefore, we proposed a circuit design that exhibits a WDIPLN configuration of $N = 4, M = 4$, allowing for us to have 4 electro-optic inputs and behavior at four distinct wavelength outputs. The purpose of this design is specifically to demonstrate the wavelength capability, and as such the task assigned was to implement the final stage

of a simple neural network architecture, as before, which represents the final 4×4 layer of our algorithm.

5.5.1 Design and Set Up

The neural network architecture for this design can be seen in Fig. 5.10 a). We are training logic gates again so that the inputs at X_1, X_2 are the same $[0, 0, 1, 1]$ and $[0, 1, 0, 1]$ as before. In this case, we again train the network offline using PyTorch, and this time we only evaluate the final stage of the network as seen in a) [87]. The previous demonstration validated the full network for one logic gate, and we are specifically focused here are demonstrating the wavelength ability of this circuit.

Following the WDIPLN architecture for the case of $N = 4, M = 4$, we see that the circuit begins with an optical fanout stage, copying a single input carrier signal onto 4 equal paths Fig. 5.10 b). From here, each path encounters a racetrack MRD, a phase shifter, and four microdisk modulators before fanning in to the single waveguide/bus output. In order to limit the thermal cross-talk of this high number of sensitive devices, we implemented the thermal isolation techniques described in **??**.

We see the racetrack MRD in c), the design conception here was to create a fairly straightforward PN diode based resonator with an integrated heater. This device has a round-trip-length equal to 403.65 μm, which was targeted to provide an FSR near 1.5 nm. This FSR is chosen because it is sufficiently large for channel seperation and spacing, while also leaving the possibility of integrating with a comb-source open for future implementations. The PN diode is created using standard lateral doping available

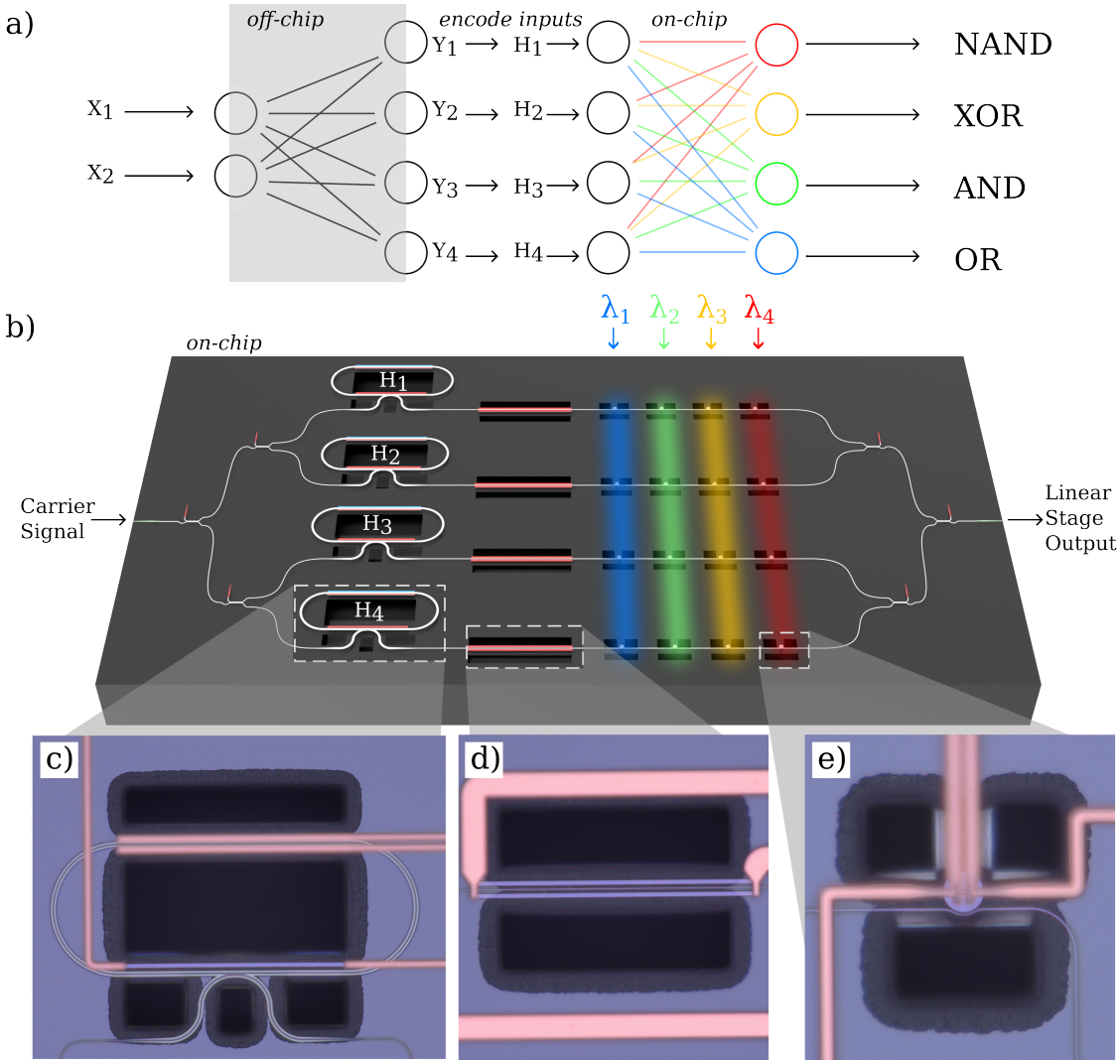

Figure 5.10: Set up diagram for four gates at four wavelengths demonstration. a) The NN architecture used for training and defining the gates at the output colors. We see by peaking at b) that these gates correspond to the specific resonances as follows: OR $\rightarrow \lambda_0$, AND $\rightarrow \lambda_1$, XOR $\rightarrow \lambda_2$, and NAND $\rightarrow \lambda_3$. b) The experimental overview and setup. Similarly to the single wavelength demonstration, we see that we have a nested interference circuit with multiple rings that follow the WDIPLN architecture. A carrier signal is passed into the circuit, which is fanned out onto four identical paths. Each path contains an input MRD, denoted as H_1, H_2, H_3, H_4 which corresponds to the input encoding from a). Next, we pass through the weights along the bus, each set to a different wavelength but corresponding to the other rings in that *column*. After this, the signals are fanned back in to a single bus and passed to the output, carrying the linear stage output signal at each channel. c - e) Insets showing the input MRD, phase shifter and weight MRDs, which have the isolation trenches

from the AIM photonics MPW PDK, and the resistive heater is created with a p-type doped silicon rail near the waveguide inside the cavity. We define the trench windows to

surround the electro-optic devices and suspend the ring in an air cladding, thus solidifying it as a thermally isolated device.

Next, we see that the phase shifter in d) has the same design and thermal isolation as the device we described in the previous chapter, as we upcycled this design. Finally, we implement a micro-disk modulator as the single channel weight transducers in this circuit [79]. This decision was down to process compatibility, since the project funding this wafer had a specific doping structure available to it. Therefore, our collaborators at Columbia University let us borrow the design they worked on for use in this circuit. The operating principles of micro-disk modulators are the same as ring resonators, there are just different compromises being made in order to achieve the desired effect, such as needing to consider higher ordered modes, coupler design. footprint and Q-factor [122]. We also were careful to try to space these microdisk modulator's resonances around 1.5 nm using our best-guess approach. According to simulations from FDTD and MODE, we observed a linear relationship between the radius of the microdisk and the resonant wavelength, with a slope of 265 nm of radius change for each 1 nm of wavelength shift, leading us to increase the radius by about 400 nm for each wavelength of interest [2, 3]. However, this relationship was based on doping simulations, which we had not refined at this time. After fabrication, the true relationship was slightly tighter, such that the average spacing of the wavelengths was closer to 1 nm. Fortunately, as MRDs are sensitive to begin with, we planned to use the integrated heaters to configure this circuit.

Initially, we use packging techniques described in Chapter 5 to flip-chip package this PIC onto a system that allows us to address all of the electro-optic elements in our design. Our specific package can be seen in

The operating principle of this circuit is very similar to the previous demonstration. First, we characterize the transfer function of the ring resonator, allowing us to impose whatever weight/input we choose into the network. This procedure is on a test device, and the outcome is shown in Fig. 5.11, however we will leave the discussion of this until the next section as it requires a deeper explanation. We then train the neural network offline, we then take the values of the final weight layer and store them. We are careful to make sure the isomorphism between the offline and online network is preserved, therefore in the final layer we *do not train a bias*, as we do not have a bias available to us. Architecturally, this is an inhibitive obstacle, but for this demonstration it is sufficient. We also then extract the hidden node values, $[H_1, H_2, H_3, H_4]$ in Fig. 5.10 a) and store them for each input question fed into $[X_1, X_2]$.

We then impose the values of the weights onto the circuit and run through the 4 input states, recording a value at each of the 4 wavelengths in Fig. 5.10 b) as the output of the linear stage. We apply the final activation of the output layer and compare the results to the target gates at each wavelength.

5.5.2 Thermal Effects of the Depletion Based, Undercut Ring Modulator

Here it is important to pause and investigate the resonance shift of the ring modulator under differing conditions. Regularly, or in the standard photonics material stack case, the resonance shift in depletion mode modulation is characterized by a square root relationship. This relationship is quite intuitive since we see that initially, the intrinsic,

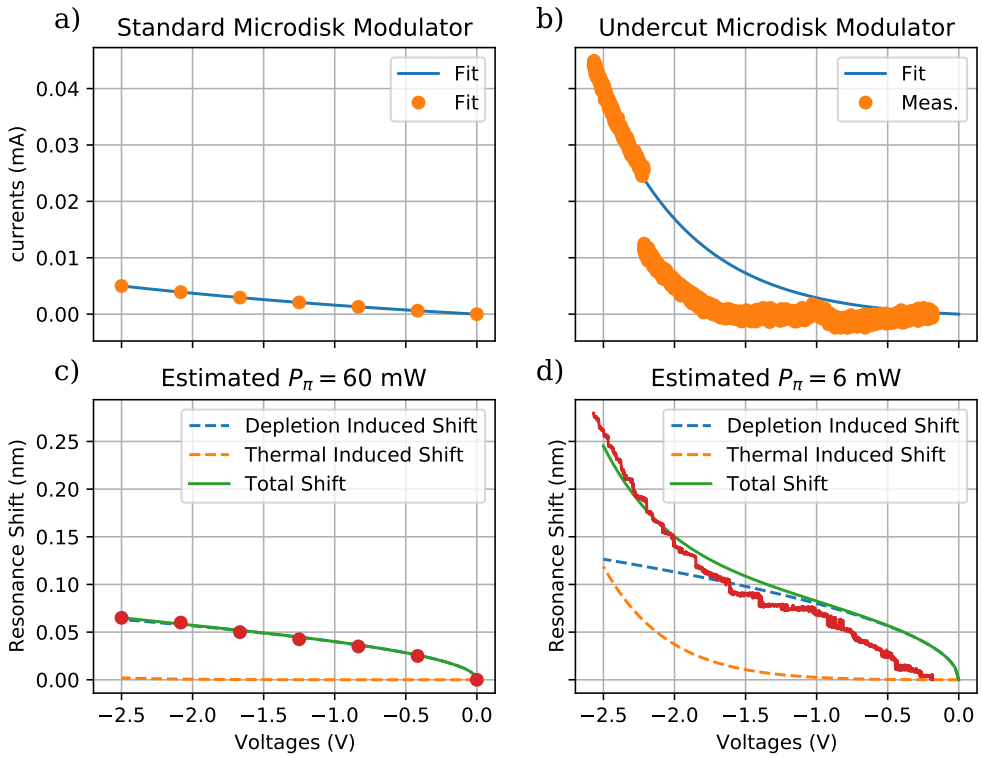

Figure 5.11: a) The IV curve of a standard stack microdisk modulator. b) The IV curve, finely measured, of a microdisk modulator that has the undercut etching. c) The resonance change with respect to voltage in the reverse bias of the standard stack device. We estimate this heater geometry to have a $P_\pi = 60mW$ c) The resonance change with respect to voltage in the reverse bias of the undercut device. We estimate this heater geometry to have a $P_\pi = 6mW$
. Notice how the thermal effect takes over suddenly around -1.5 V, as the thermal sensitivity has been increased due to the air isolation.

or depletion, width change greatly overlaps the contained mode. However, as the bias increases, the effect of pulling the carriers away occurs further away from the optical mode, so the effect lessens with greater reverse bias. We generally approximate this as a square-root-like function. Specifically, we measured a ring modulator in reverse bias using a Keithley 2400 SMU and saw that the reverse bias current stays low for such a small diode, while the effect of the resonance shift extracted exhibits this square root shape Fig. 5.11 a, c).

We estimate the power required for a 60 mW/π phase shift from a HEAT simulation for the standard photonic stack and microring cross-section. This is a fairly typical thermal efficiency for this type of ring structure [1, 21, 121]. We describe the resonance shift in DC operation as a function of the diode behavior and the thermal behavior of the ring modulator from the diode drive. This approach implies that if the power dissipated in reverse bias increases enough to change the temperature, the resonance shift would confluence these two effects. Generally, the temperature generated by a small reverse bias current is not detectable, as seen in Fig. 5.11 c).

In the case where we perform the thermal undercut process described in an earlier chapter, we see a few interesting new things happen. Firstly, the device is better thermally insulated, and the thermal generation potential is much higher – the heat generated stays near the heat source and is not sunk away by the silicon handle as readily as before. This changes the diode behavior, slightly increasing the reverse bias current due to thermal effects, observing the measurements in b). They were performed with our experimental SMU, the Qontrol, which is used for its cost-effectiveness on high channel count while sacrificing a little bit of accuracy compared with the Keithley 2400 SMU. This difference explains some of the non-idealities in the current measured from this sweep in b). Additionally, the sample size of our data is small (i.e., $N = 1$), so this difference can also be described in the process variation and post-process effects; more data collection is the next step to learning about the changes of the diode. Secondly, this increase in reverse bias current and improved thermal efficiency of the geometry creates a more thermally sensitive device. Here, as a thermally addressed ring resonator, this is a desirable outcome. But, we can see in Fig. 5.11 d) that the resonance shape now

exhibits a sum of the square-root diode behavior and the square thermal behavior of the current. We can describe this generally using the following equations

$$|\lambda_{\text{PN Diode}} - \lambda_0| = \sqrt{A^2 V}, \tag{5.14}$$

$$|\lambda_{\text{Parasitic Thermal}} - \lambda_0| = BV^2, \tag{5.15}$$

$$|\lambda_{TOTAL} - \lambda_0| = |\lambda_{\text{PN Diode}} - \lambda_0| + |\lambda_{\text{Parasitic Thermal}} - \lambda_0|, \tag{5.16}$$

$$= A\sqrt{V} + BV^2, \tag{5.17}$$

where A, B are fit coefficients for the two curves, and V is voltage. We see that from Eq. (5.17), the final shift is described as this sum, which implies that for large voltages or a large B, the thermal effect takes over. At large voltages, diodes become increasingly difficult to model as they are prone to breakdown, which causes the diode to become damaged. However, B can be increased by changing the thermal system in which the diode exists, as we are demonstrating here.

With an estimated 6 mW/π thermal efficiency from the geometry, we see that the thermal power dissipation has an overwhelmingly larger effect on the resonance shift of the device. We expect that at high drive modulation speeds, the thermal effect would begin to average as it lost its ability to keep up, leading to an averaged temperature and the diode behavior as the only mechanism for modulation. This hypothesis represents future work for further study of ring modulators with thermal isolation undercut structures. In addition, the high thermal sensitivity decreases the threshold for introducing optical bistability as the self-heating from the resonator will be enhanced, which is an additional future work from this device.

For this experiment, however, we are interested in using these diode rings as the weight attributes where they remain fixed during operation. Diodes were selected for their ability to operate nearly thermally, and in an extremely thermally sensitive environment, this choice is critical. Even with the small thermally induced shift we see in these devices, so long as we are able to capture this in the calibration stage, we are able to operate with this slight non-ideality, as the thermal cross-talk is highly suppressed.

5.5.3 Configuration of Circuit

Before we can run the experiment, we need to configure the circuit. We discuss the packaging of this experiment in the next chapter in more detail, but for now, we can visualize that the chip is packaged electronically and is addressable optically. For the SMU, as previously mentioned, we use the Qontrol system, which has a series of single-channel SMUs that can be connected to our circuit using flex connectors. This circuit requires 96 connections to run, and half of these are ground. However, in order to avoid ground-looping, we chose not to tie grounds off the chip, therefore, each connection is tied to an active channel on the Qontrol, allowing us to explicitly set each bias point by setting the voltage. We mapped the traces coming from the PIC, all the way through the two PCBs and electronic interposer used to the channels of the qontrol. Once packaged, we verified the electronic connections by performing IV sweeps of each device. We include this configuration step in Fig. 5.12

In the design of this circuit, on each of the 4 busses containing the inputs, phase shifter, and weights, we have a small 1% power tap connected to a photodiode. The

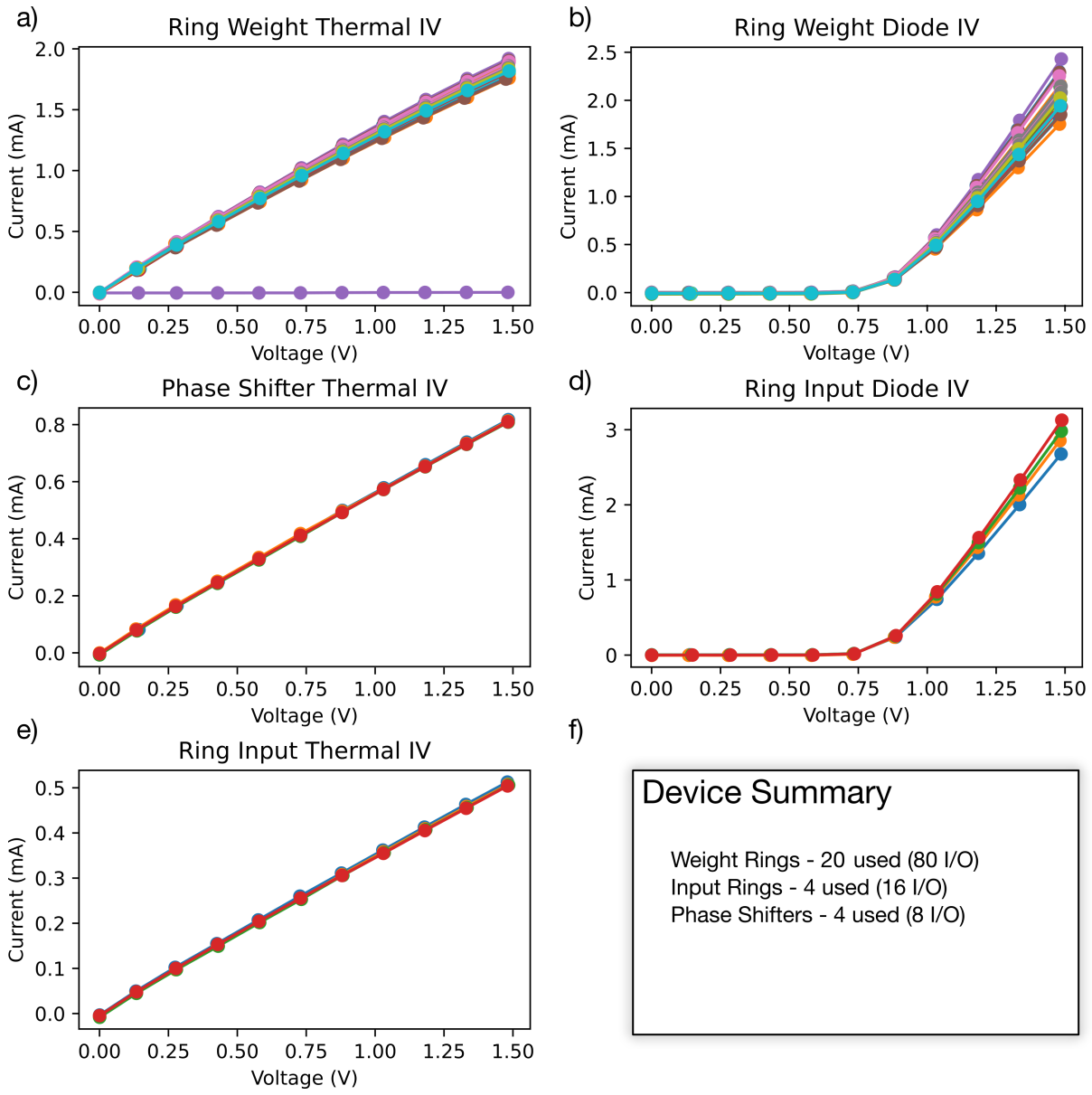

Figure 5.12: Configuration curves of the packaged devices to demonstrate connectedness. a) The resistor IV curves of the weight MRDs, showing the single open connection in the fourth column (3rd row) and average resistance 1 $k\omega$. b) The diode IV curves of the weight MRDs. c) The resistor IV curves of the phase shifter elements, with average resistance 1.6 $k\omega$. d) The diode IV curves of the input MRDs. e) The resistor IV curves of the input MRDs, with average resistance 2.5 $k\omega$. f) A short device summary, showing the count of each category of device and total I/O counts.

intention of this was to improve the configuration of the circuit by individually monitoring each bus. This approach allows us to have a cleaner viewpoint of each device, without the interference that comes from the fan-in stage. However, as this wafer run was a

developmental run to achieve the undercut processing, the foundry chose to short loop the wafer processing by eliminating the germanium epitaxy. This decision leaves us without photodiodes. We are given one option for circuit configuration, spectral inspection, and "brute force." We follow the algorithmic approach outlined in the following pythonic pseudo-code Listing 1.

To the keen observer, we have left out some specific function definitions and general imports, somewhat because this is pseudo-code, but primarily because this is intended as instructive rather than operating code. Implementing this requires more work in the way of equipment interfacing and data collection. Here, we focus on the algorithm for configuration, which if we follow the psuedo-code, we see that it is an iterative endeavor. The circuit begins in a random-state, which is an assumed unideal configuration. We begin by sweeping over the desired devices, fitting the way the resonance curve changes with respect to the voltage bias, and mapping the desired wavelength back to the appropriate voltage. In this way, we fit and set each device during and throughout the configuration. In doing so, the configuration allows for new thermal gradients to be accounted for as we repeat this process a few times, where in the final iterations we are making only fine adjustments.

Fig. 5.13 shows a few parts of this procedure. First, in a) we see the initial device response across some wavelengths. this is a confusing optical spectrum, but not to fear, our circuit exists inside this mess if we just coax it out. In b) and c), we show an early and later configuration sweep. Notice how the early configuration sweep has more spectral confusion, such that the resonance lines are many and varied. However, in the second sweep in c) all of the devices are nearly perfectly aligned, and if we look extremely

```
"""
Configure the Circuit
"""
# list phase shifter devices
ps_devices = ["ps0", "ps1", "ps2", "ps3"]
# list out the devices for the input rings
input_devices = ["h0", "h1", "h2", "h3"]
# generator function to get the weight device names
# define util function to define the weight ring names
def weight_devices(wavelength_index):
    return [f"wl_{k}{wavelength_index}" for k in range(4)]
# get the 4 wavelengths of rings
w0 = weight_devices(0)
w1 = weight_devices(1)
w2 = weight_devices(2)
w3 = weight_devices(3)

# set the desired wavelengths
wl_targets = [y0, y1, y2, y3]

# set the voltage range, here we say from 0V to 2V with 0.1 steps
voltages = arange(0,2,0.1)
repeat_count = 5 # set a repeat counter
for j in repeat_count: # iterate over the procedure repeat_count times
    # iterate over the device lists
    for dev_list in zip(ps_devices, input_devices, w0, w1, w2, w3):
        for i, dev in enumerate(dev_list):
            res = wl_v_sweep(dev, voltages) # wavelength and voltage sweep
            # fit the resonance shift versus voltage
            fit = extract_transfer(res) # expect V^2
            # store the fit
            save(fit)
            # set the wavelength
            # if a phase shifter, maximize this circuit value
            # ensuring that we are in constructive interference state.
            if dev.contains("ps"):
                set_device_at_max(dev)
            # if this device is an input resonator
            # set the wavelengths of all devices to the first target
            # since they have the same FSR
            elif dev.contains("h"):
                set_device_wl(dev, wl_target(0))
            # otherwise, set each column of rings to the new target
            # so that we line them up
            else:
                set_device_wl(dev, wl_target(i))
```

Listing 1: Pythonic Psuedo-Code for Circuit Configuration

closely we can see that there is one faint resonance shifting from near $1,541.2$ nm at 0 V and increasing quadratically with higher voltage. Again, we model this transfer as

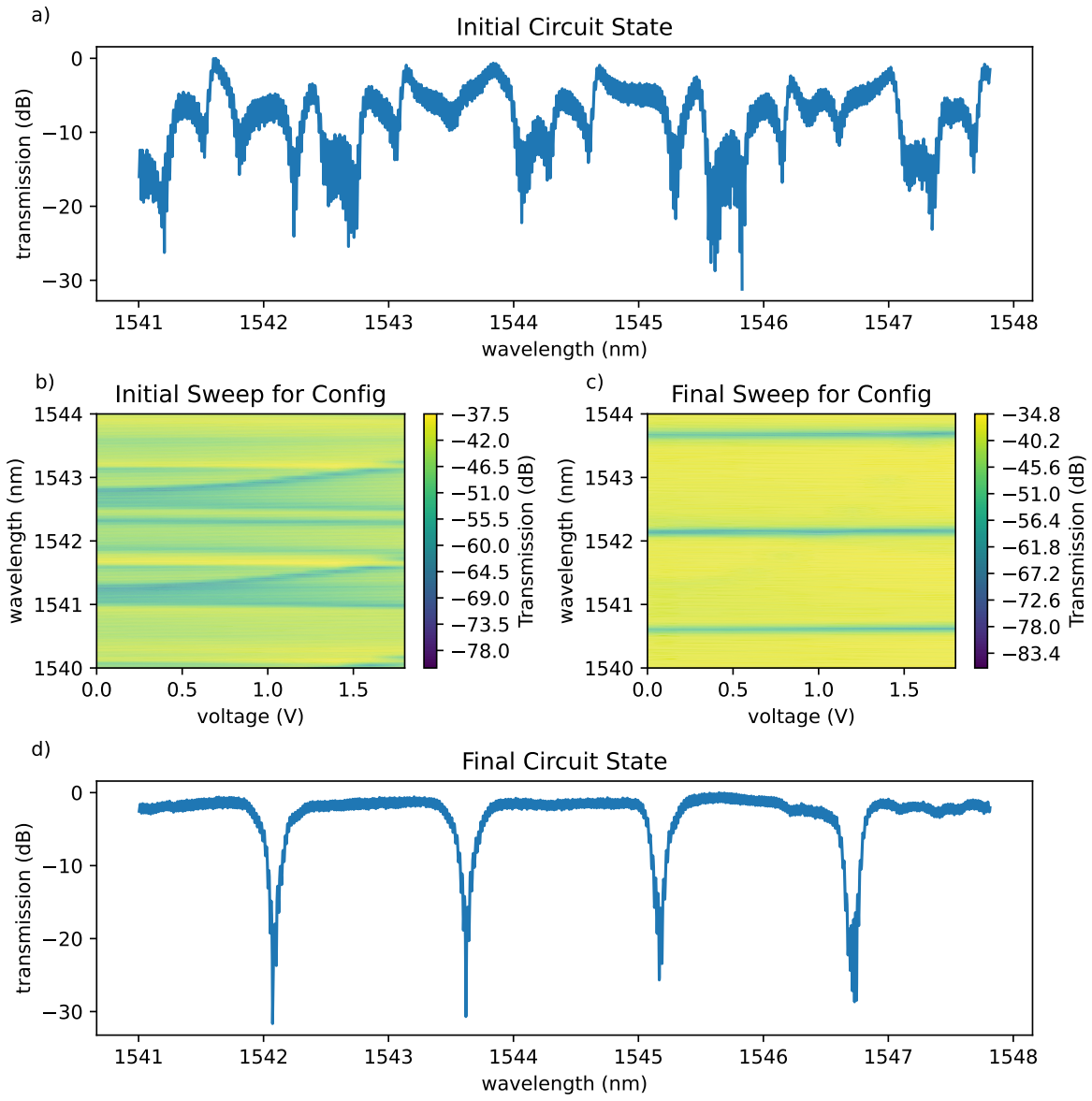

Figure 5.13: a) An initial wavelength sweep of the circuit, showing randomly distributed devices that seem out of phase. b) A first configuration sweep where the wavelength and voltages are swept to determine how a particauly MRD is shifting. We use this in the alignment procedure. c) A final configuration sweep, where the configuration has corrected the state of the circuit, and we see a single, faint resonance shift. d) The final circuit state before test, all the appropriate resonances are aligned at four distinct wavelengths.

a quadratic, where the fit of this faint peak shift is stored. Finally, in d) we can see a fairly optimal optical spectrum for our device, and we know that the four input rings are aligned to one another and that each column of weight rings is aligned to successive

resonances on the input rings. We are now configured and ready to begin the experiment.

At this point, it is important to discuss one practical matter which is specific to the device we tested in Fig. 5.13. There were a total of 8 MRDs in the weights, as we had initially designed this circuit to accommodate up to 8 outputs. However, due to packaging difficulties, device inconsistency, and sensitivities, not all of the connections were made on the fourth column of rings. Therefore, as part of the calibration, we shift the fourth column out of the experimental region and use the fifth column in its place to perform the measurements. The disconnected MRD can be seen spectrally in 5.13 d) near $1,546.5$ nm and other unused rings beyond the final resonances as small teeth. We suspect that these missed connections may impede the accuracy of the measurements.

5.5.4 Results

Using the imposition scheme for the weights and inputs, we proceed to perform the experiment. The experimental procedure itself is very straightforward. With the configuration of the devices in place using the thermal responses, we impose the weights onto the diode response of the weight MRDs in columns 1, 2, 3, and 5. We then iterate through each input set corresponding to each input question of the logic dataset, by retrieving the stored values from the offline training and imposing them onto the input modulators with an inverse mapping. This mapping is created using a fit of the curve similar to that described in Fig. 5.11. Once we set the input, we sweep the laser across the four resonances, as seen in Fig. 5.13 d), and record the outputs. We use a wavelength determined from the initial configuration state to probe the output spectra. These wavelengths are defined as

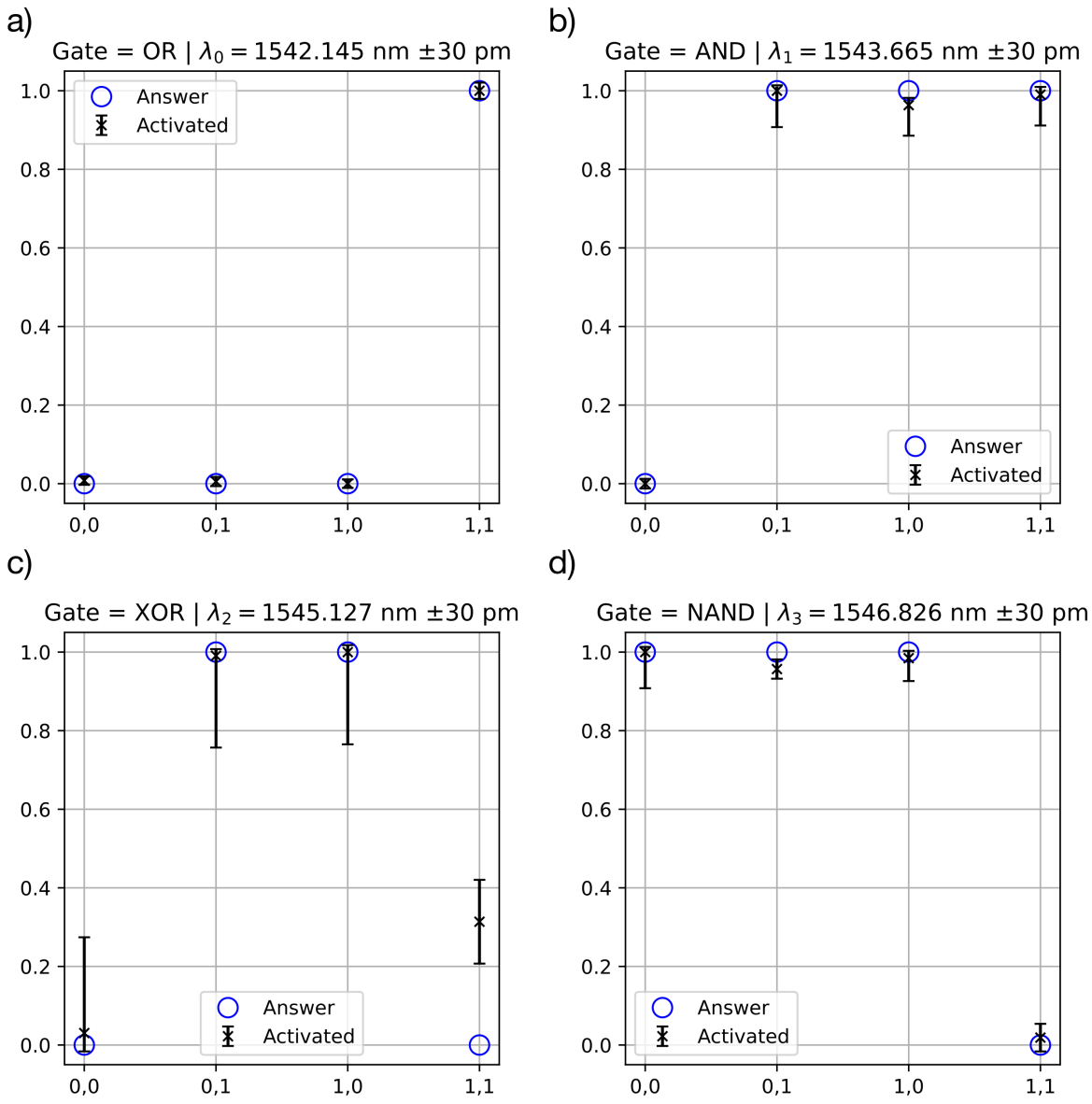

Figure 5.14: The results of four gates at four wavelengths. We have some variance around the probe wavelength of ±30 pm. a) OR gate, which we demonstrate to an accuracy of 99.87%±0.1% at $\lambda_0 = 1,542.145$ nm. b) AND gate, which we demonstrate to an accuracy of$99.05\% \pm 0.78\%$ at $\lambda_1 = 1,543.665$ nm. c) XOR gate, which we demonstrate to an accuracy of$98.05\% \pm 1.6\%$ at $\lambda_2 = 1,545.127$ nm. d) NAND gate, which we demonstrate to an accuracy of$99.73\% \pm 0.38\%$ at $\lambda_3 = 1,546.826$ nm.

$\lambda_0 = 1,542.145$ nm, $\lambda_1 = 1,543.665$ nm, $\lambda_2 = 1,545.127$ nm, $\lambda_3 = 1,546.826$ nm. We take a 30 pm window around these wavelengths to arrive at the result. Fig. 5.14 displays the final results for each gate. For the OR gate, we are at the lowest wavelength and achieve

an accuracy to the gate of $99.87\% \pm 0.1\%$, after we apply a logistic activation function. The logistic activation function is centered at 0.5 and is the same as we used in training. The amplitude of the measured value is normalized, similar to the single gate procedure followed above. In b) we show the AND gate result, targeted at the second wavelength, to an accuracy of $99.05\% \pm 0.78$. In c) we show the XOR gate result, targeted at the second wavelength, to an accuracy of $98.05\% \pm 1.6$. Here, we recognize the variance is much higher than the other gates and we believe this is due to phase interference from the disconnected MRD. Finally, in c) we show the XOR gate result, targeted at the second wavelength, to an accuracy of $99.73\% \pm 0.38$.

5.6 Conclusion

Silicon photonics holds promise for implementing large-scale PNNs directly on-chip with substantially improved performance compared to electronic implementations. In this work, we propose and demonstrate a novel PNN architecture, WDIPLN, which exhibits highly favorable characteristics compared to previous literature. The WDIPLN derives from the COLN architecture detailed in [76], enabling similar vector-vector multiplication and summation using an optical carrier signal with the additional benefits of wavelength parallelism, drastically reduced device footprint, and significantly lower energy consumption. The majority of these beneficial properties come from exchanging large, power-hungry MZMs with compact MRDs. While we can trivially exchange the fundamental elements (MRDs for MZMs) in a one-to-one manner to enable multi-channel operation, we show that by replacing the input MRDs with a single, multi-channel MRD

with an FSR equal to the weighting channel spacing we can reduce the total device count significantly. As a first proof-of-principle, we experimentally demonstrated simple addition and subtraction between rings in a WDIPLN circuit. We then show a WDIPLN configured with $N = 2$ and $M = 1$, which we configure and recycle to perform 2-bit logic gates (AND, OR, and XOR). We observe the implementation accuracy for each gate to be $96.8\%, 99\%,$, and 98.5%, respectively. We next demonstrated a WDIPLN configured with $N = 2$ and $M = 4$, where we simultaneously trained all four logic gates using the wavelength multiplexing of the design. This effort requires more careful configuration, however, it is more readily scalable and compatible with sources like frequency combs. The wavelength channels of this demonstration were defined at $\lambda_0 = 1,542.145$ nm, $\lambda_1 = 1,543.665$ nm, $\lambda_2 = 1,545.127$ nm, $\lambda_3 = 1,546.826$ nm. This circuit achieved the implementation accuracy of $99.87\%, 99.05\%, 98.05\%$, and 99.73%, respectively. Furthermore, the natural wavelength parallelism of the proposed WDIPLN architecture can be exploited using chip-based Kerr frequency comb sources [36] for massive scaling in the frequency domain, similar to recent demonstrations in the silicon photonics platform for high bandwidth data communications [97]. These demonstrations open new opportunities in massively parallel silicon photonic PNNs and pave the way to large-scale systems in the thousand-neuron regime on a single chip with minimal energy consumption.

5.7 Future Works and Considerations

5.7.1 Phase Shifting

There are many avenues to explore to improve this architecture and its implementation. Namely, a huge inhibitor of this circuit architecture is a lack of a phase shifter specific to wavelength. Our phase shifters operate similarly across the wavelengths we are using since they are not resonant phase shifters. A new avenue to enhance architectural robustness would be to investigate a resonant phase shifter. This effect could be achieved by introducing a copy of the weight MRD, but with the coupling greatly suppressed. The intrinsic Q factor for this ring would be identical, leaving only a phase shift across what would be the resonant wavelength. If the amplitude of this resonance is suppressed while we achieve a phase response, this may be an excellent candidate for wavelength-specific phase shifting. However, in the case that this path proves fruitless, the current architecture works well within the context of similar sign weights, and upon a weight change we would need to serialize the operations as we adjust the phase. A future, polished circuit implementation could use a high-speed phase shifter in place of the thermo-optic phase shifters in this demonstration, which can increase the throughput speed.

5.7.2 Improved Device Design

Both the input and weight MRDs were inherited or naïve designs for this project. Considerable circuit improvement is possible through device optimization. In particular, the weight MRDs can be replaced by the inner ridge modulator discussed in Chapter 3.

5.7.3 Backpropagation

In this circuit, we implemented a simple transduction of the real values trained offline onto the voltages in a very naïve way. We assumed all rings act the same and that we could impose the weights from the real values to voltages from calibration structures. This assumption obviously holds to an extent, but improvement can be achieved, not to mention reduced calibration, if we implement an on-chip learning algorithm. We have a proposal for such an algorithm and present it as future work for this architecture. We have

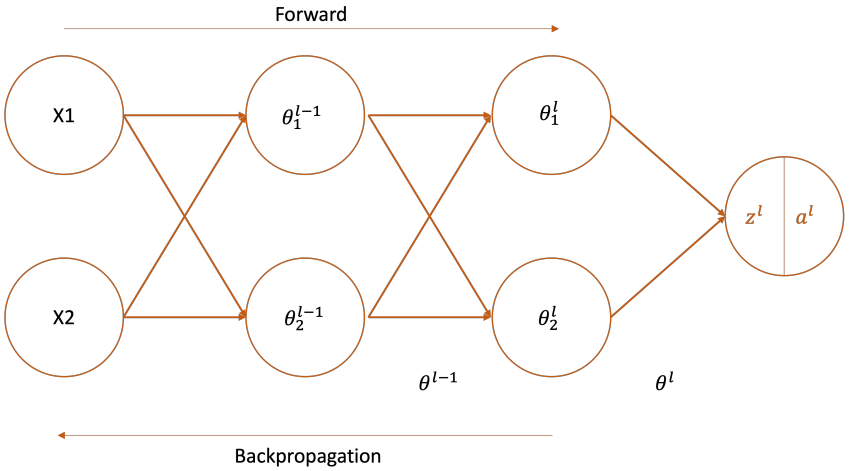

Figure 5.15: A simple schematic for reference of the gradient descent algorithm.

developed a simple gradient descent algorithm to handle the on-chip backpropagation for the WDIPLN circuit. The complication is that we need to derive the gradient descent with respect to the ring resonator's transfer function according to the bias voltage, not just an arbitrary real-valued number. This requires an extra step in the derivative. We can begin with some definitions. We can define the full system as $a^l = g(z^l)$, where $z^l = \theta^l a^{l-1}$. Specifically, if l is the layer and $theta$ indicates the weight matrix associated with the respective layer, and if g is the activation function. Here, we define an error

function, such that

$$E = \frac{(a^l - y)^2}{2}$$

where we then can proceed to a derivative of the error function with respect to the voltage of each ring resonator in the weight matrix.

$$\frac{\partial E}{\partial V^l} = \frac{\partial E}{\partial a^l}\frac{\partial a^l}{\partial z^l}\frac{\partial z^l}{\partial \theta^l}\frac{\partial \theta^l}{\partial V}.$$

Now we define each component of this derivative:

$$\begin{aligned}
\frac{\partial E}{\partial a^l} &= (y - a^l)(-1) = (a^l - y), \\
\frac{\partial a^l}{\partial z^l} &= g'(z^l), \\
\frac{\partial z^l}{\partial \theta^l} &= a^{l-1}, \\
\frac{\partial \theta^l}{\partial V} &= f'(V).
\end{aligned}$$

Finally, we can combine these expression so that we describe the voltage dependent derivative of the error function:

$$\frac{\partial E}{\partial V^l} = [(a^l - y) \odot g'(z^l)][(a^{l-1})^T \odot f'(V)].$$

This derivation of the gradient descent can allow us to automatically configure or train the network on-chip.

Chapter 6

Packaging

As a Ph.D. student, we designed a packaging platform for optical and electrical co-packaged chips using in-house techniques in our lab and nanofabrication facility at RIT. Here, I will describe the relevant details for the packaging platform I developed and maintained ownership of. The full scope of our group's capabilities is beyond what will be described here.

The packaging platform we created can be considered a necessity out of the constraints of design coupled with the restraints of capabilities for different components. For us to have a co-packaged optical electronic device, we need to have multiple electrical I/Os that are orthogonal (do not interrupt) the optical I/O. This idea leads to design rules, design criteria, and the innovation of new solutions.

6.1 Electronic Interposer

The electronic interposer (or electrical fan-out) is a simple yet beneficial component for packaging wire-bonded chips. The primary rationale for this part is to take the wire bond

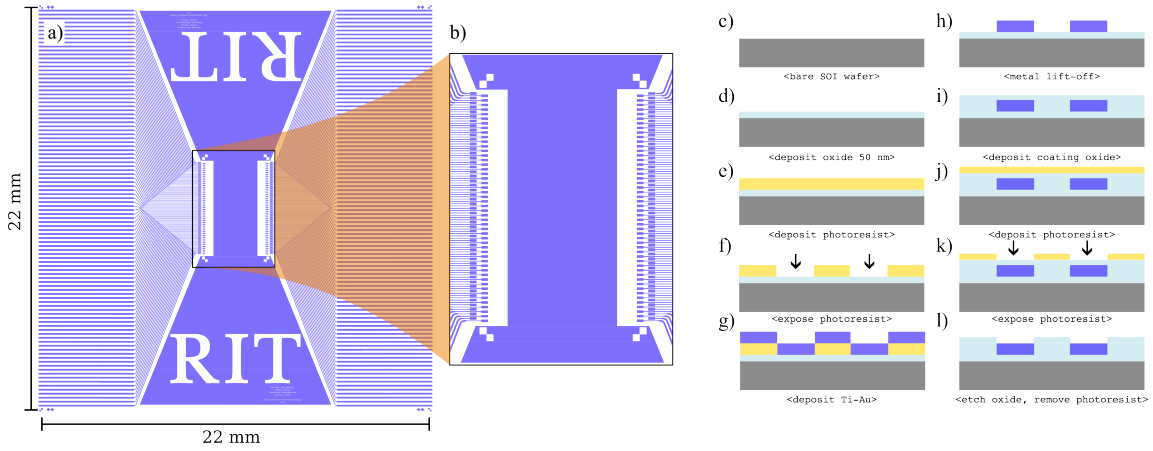

Figure 6.1: a) The fanout mask design, which sticks to a 22 mm × 22 mm size, and we see in b) the close up of the bond region. c) We begin with a bare, high-resistivity silicon wafer. d) We grow a thin layer of oxide on the wafer using PECVD. e) We spin-coat resist onto the wafer. f) We pattern and expose to define the metal. g) We evaporate the metallic Ti-Au onto the wafer using electron beam evaporation. h) We perform a lift-off process by dissolving the photoresist, which removes the metal on-top of the photoresist remaining. i) We use TEOS to deposit more oxide, to bury the metal and create passivation. j) We spin-coat more photoresist. k) We pattern and expose this photoresist using a mask that defines the openings. l) Finally, we etch the oxide and remove the photoresist, leaving a wafer with passivated metal and bond point openings.

pads on the circuit chip and route them out to a larger pitch. The motivation is two-fold: making a cheaper overall solution and improving quality/yield likelihood. It can cost- and capability-prohibitive to have a printed circuit board (PCB) fabrication facility attempt to reach the feature size required for the pitch of our PICs, which can often hover around $1-200$ μm. This translates (in PCB language) to under 1 mil (1/1000th of an inch), which is very difficult to achieve in this non-photolithographic process. PCBs typically have minimum resolutions of around $4-5$ mils, with exceptions near 2 mils, are often more costly. Therefore, we design an electronic interposer to interface between the small pitch available on our PIC to a much larger pitch compatible with the PCB, which we create in our in-house fabrication facility.

This design is similar to designing a photonic chip, using a scripted layout tool with a

Python interface called Nazca Design [4]. With this interface, we can craft the inner and outer pad structure and enforce all of the routings to enable the fan-out of the signals. We see the final design of one of these interposer designs in Fig. 6.1 a), with a close-up of the PIC and bond area in b).

With a/the design in hand, we submit this for mask-writing. Once the mask is written, we proceed to the process, as shown in Fig. 6.1 c-i). This process involves multiple steps, beginning with substrate preparation by cleaning the silicon wafer and growing a thin oxide layer on its surface using oxygen plasma enhanced chemical vapor deposition (PECVD) c-d). A photoresist layer is spin-coated onto the metal surface and exposed to UV light through a photomask to define the desired pattern e-f).

The next step is depositing a metallic layer of Ti-Au on the oxide-coated and exposed photoresist wafer, typically with a target thickness of 100 nm, using an electron beam evaporation system f). The unpatterned metallic layer is then removed through a lift-off process using a solvent, such as acetone, which dissolves the photoresist layer and lifts off the metal along the pattern defined by the photoresist g-h). The resulting patterned metallic layer is rinsed with DI water and dried.

An additional step to the process we have added is to protect the metal and only open the pads, we can call this the passivation step. We conduct additional oxide deposition using tetraethylorthosilicate (TEOS) in a low-pressure chemical vapor deposition system. The oxide is deposited to a thickness of around 100 nm (targetting 50 nm above the metal) i). Finally, the oxide layer is etched using a wet etching process with hydrofluoric acid or a dry etching process with a plasma system to create openings in the desired locations defined by the photoresist, additional exposure, and etching j-k). Finally, we

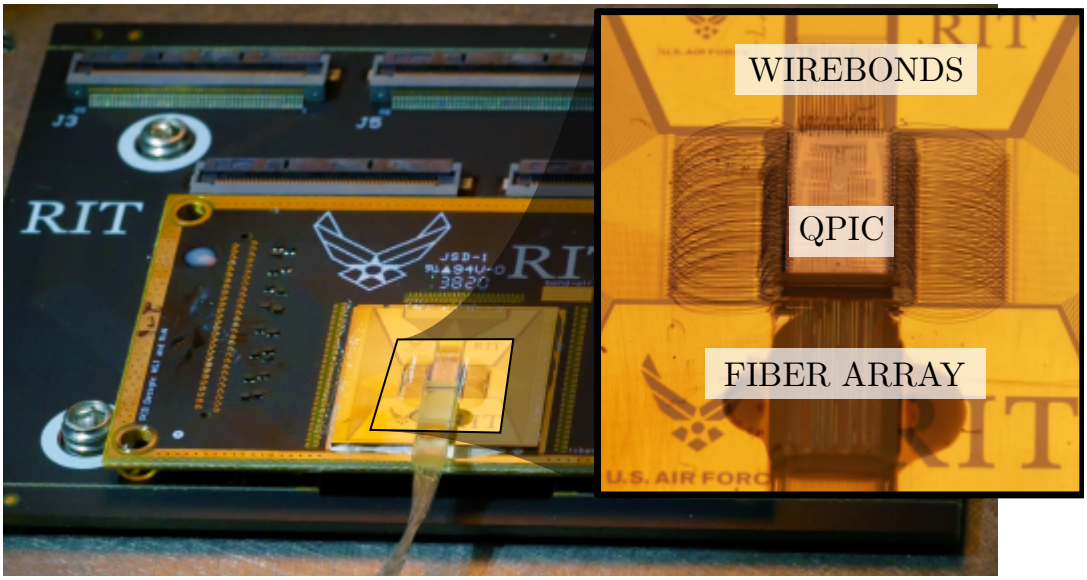

Figure 6.2: An example of the two-board solution for AFRL, which has a PIC on an electrical interposer chip, wirebonded in two stages from PIC-interposer and interposer-PCB. The PCB is then socket plugged into the receiver PCB, which then interfaces with the electronics. We also had this part fiber attached, demonstrating a significant step in our packaging efforts, particularly for quantum applications.

etch the oxide l), performed until we have removed the oxide fully and exposed the metal in preparation for electrical I/O.

6.2 Printed Circuit Boards

Along with the electrical interposer design, we placed some effort into designing the printed circuit boards (PCBs). For the design of the printed circuit board, we used the software platform Altium Designer. Our board is not highly complicated, as there are no active components or power delivery required. Primarily, we are using this board as a secondary interposer or interface to the source-measure units off-chip. We show a diagram of the PCB in Fig. 6.2 This PCB is fabricated and populated with the connector component, a 50-pin FFC connector. This type of connector is often found in displays,

but for us it interfaces with the SMU we use – the Qontrol [5]. The Qontrol system has enough channels so that we can address up to 300 active connections at once with a common ground. The large-scale solution leverages the interposer and the PCB to create the electrical I/O. However, we also have mechanically designed the electrical and optical I/O to be orthogonal to each other, making co-packaging efforts possible. For example, we often will attach a fiber array with multiple optical channels or use a photonic wire bonder to optically address a PIC which has been packaged using this platform. We use wirebonds or the flip-chip process for electrical addressing, described in the following sections.

For our higher I/O designs, we developed a plug-and-play design with the help of Shelton Jacinto at the Air Force Research Laboratory's (AFRL) Information Directorate. The idea here is to keep the part that holds and maintains the PIC to the correct size for our tooling: the wire bonder and fiber attach system. But then allow this board to plug into a receiver board that interfaces with the off-chip electronics. We designed this to accommodate a project for work with AFRL and then adapted a simpler version for use with the work in Chapter 5.

6.3 Electronic Interposer for Flip Chip

Following the same procedure for the wire bonding interposer, we can create an interposer for the flip-chip process. In this design, the oxide cladding protection is a crucial part of the discussion, as it improves the quality of the flip chip bonding by protecting the traces near the bond pads. Since we do not have a second metal layer or vias, we must route the

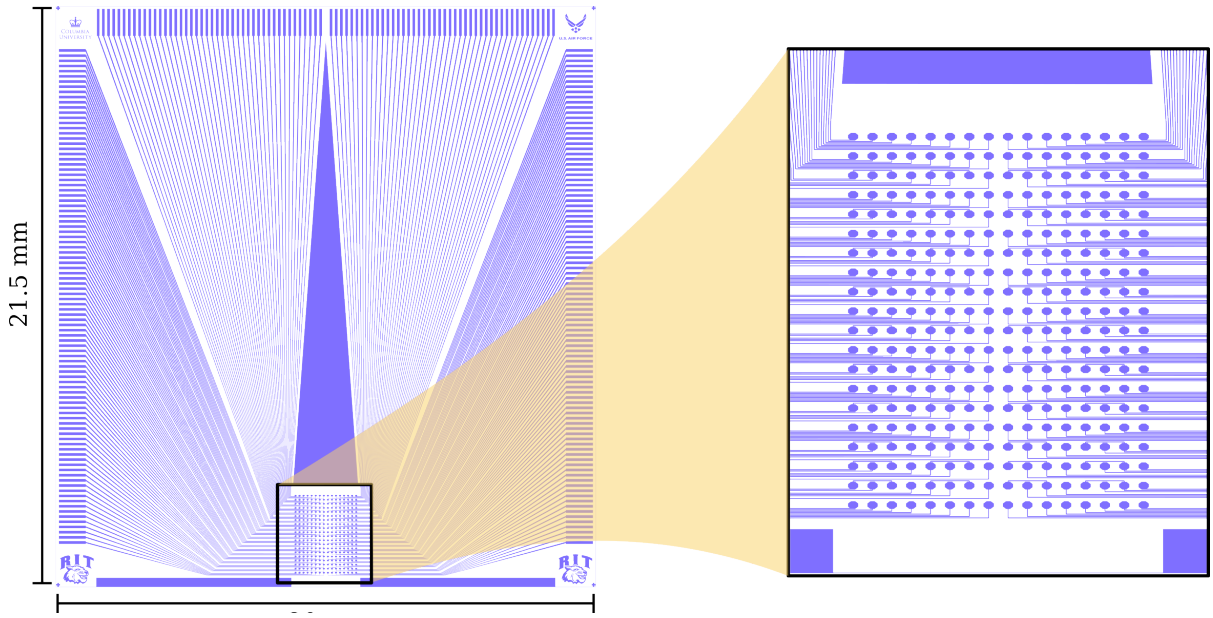

Figure 6.3: Our design for the three-sided flip-chip electronic interposer. The total size is near the max of the inhouse photolithography stepper, at 20 mm × 21.5 mm

traces between the bond pads. This led to the discovery of process issues as we noticed metallic speckling of features around $300-700$ nm, which could link up and join adjacent traces in the tightest regions of the routing. We iterated over two new designs, trying to shrink the traces and pads enough to create at least 5 mum gaps while preserving a trace width larger than 5 mum to support the current of the electrical signal without adding too much resistance to the trace. We choose 5 mum because this is the minimum width of the traces on-chip, with a similar thickness metallic layer. From the inner pads, we route the electronic traces to three sides of the interposer to prepare for wire-bonding this area of the PCB. This is shown in Fig. 6.3, with an inset for the flip chip region. The flip chip bond pads are on the same grid as the PIC pads, 150 μm × 150 μm.

6.4 Wirebonding

Wire bonding is a process common to microelectronics wherein we connect the package/substrate with the microchip. We can also use this process for photonics to connect wirebonds from our PICs to the outside world, be it a PCB like above or some other packaged substrate or electronic chip. Here we describe the two types of wire bonding we use commonly in our lab, ball bonding and wedge bonding. There are some obvious differences that we will highlight; however, the primary reason to use one type of wire bonding versus the other often comes down to material compatibility, temperature requirements, and technical comfort with the tool. Gold ball bonding often has a higher ability to rework failed wirebonds and seems to produce neater wires due to the material's malleability. Aluminum wedge bonding, on the other hand, has a better chance of adhering to challenging materials and requires no heat.

6.4.1 Gold Ball Bonding

The process of forming a ball on a gold wire using a ball bonder involves several steps. First, the gold wire is fed through a capillary, which is a small tube that guides the wire and controls its position. The capillary is typically made of a hard material such as tungsten or ceramic to withstand the high temperatures and pressures involved in the process.

The wire must be prepared to form the ball, this is done by “pinching" off the excess wire sticking through the capillary. With a small amount of wire sticking through the capillary, a small metal paddle is mechanically moved into position (commonly referred

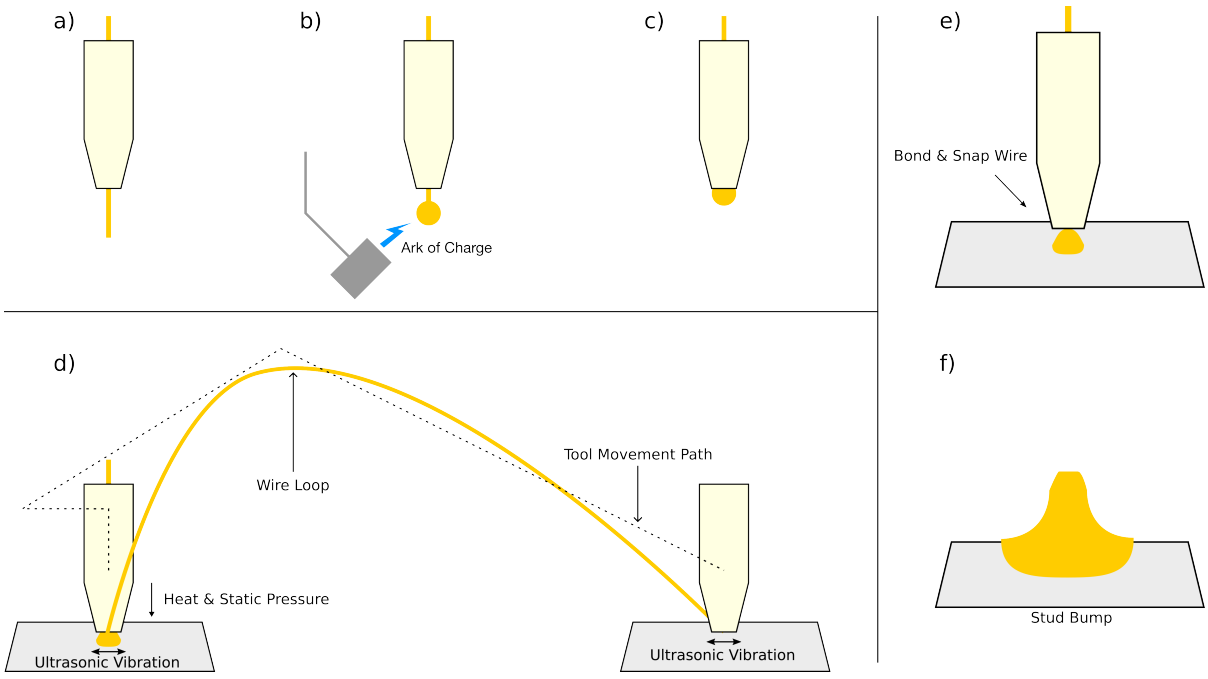

Figure 6.4: A completed package using the thermo-compression flip chip package, showing an inset of the stud-bumping process.

to as the flame). This metal paddle sits a small distance (100s microns) from the exposed wire. A current is generated within the tool, which is large enough to traverse the air gap. This next part actually seems like magic, but it's merely science. The molten metal, which is the state of the small exposed wire after the high current jumps across the air gap, is formed into a ball due to the surface tension of the metal. The amount of charge and time can create a smaller or bigger ball. However, we tend to stick to about a 2x ball diameter vs. wire diameter, but this also depends on the capillary. For the tool at RIT, we have 25-micron wire and generally operate with a 45-55 micron ball size, which is large enough to not get sucked back through the capillary and small enough for our wirebonding pads. After the ball is formed, as alluded to, the tool applies a vacuum to the wire to fix the ball at the end of the capillary. This prepares the wire for the first bond.

The capillary lower to the surface for the first bond, usually aligned via a microscope or camera overhead. Once it touches down on the wire bond pad, it applies an ultrasonic burst of energy through the attached transducer, and along with the slightly increased substrate temperature (80-130 degrees Celsius), the ball adheres to the surface. For gold-aluminum bonding, a common bond type we use as our PIC pads are aluminum; the bond forms an intermetallic layer at the time of the bond, which is the primary vehicle for adhesion. For gold-gold bonding, which we also use to connect the wirebond to our interposer, the heat and ultrasonic energy melt the donor and acceptor gold together to form the bond. From this bond point, the capillary is moved up and away in the loop style decided by the tool operator and makes its way to the next location. At this destination bond, the wire is crushed into the substrate and pinched off, similar to the above wire-preparation stage. This completes the bond from the ball to the tail and the loop.

6.4.1.1 Stud-Bumping

However, we can also perform a bond called a stud bump. This is something we use in the flip-chip process. Here, we follow the steps above to the point of the first bond of the gold ball. At this stage, the tool only moves a short distance (20-50 microns) away from the bond location. Here, the tool clamps the wire and pulls up, severing the wirebond to leave a small stud connected to the substrate. Sometimes, this is used to add insurance to the tail bond of gold wirebonding as "security", but also we can use it for the flip-chip process.

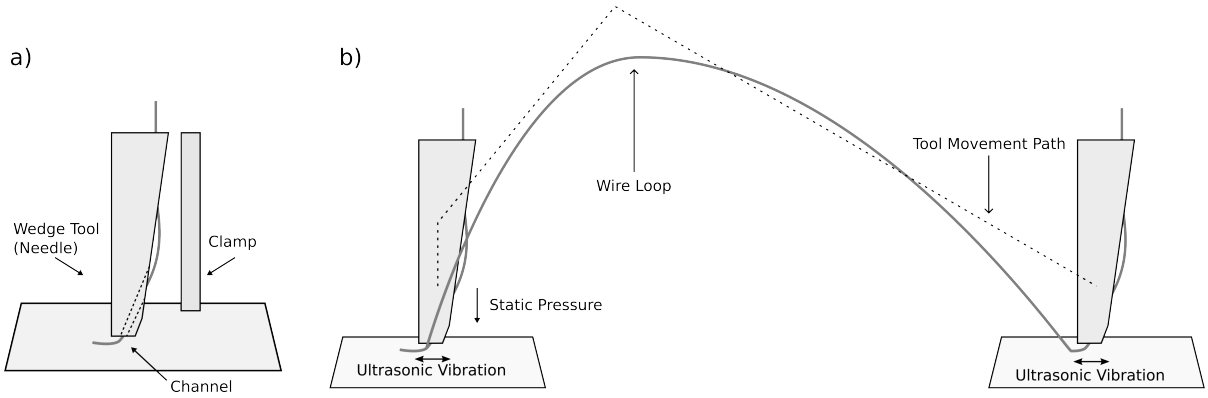

Figure 6.5: A completed package using the thermo-compression flip chip package, showing an inset of the stud-bumping process.

6.4.2 Aluminum Wedge Bonding

Compared to ball bonding, the wedge bonding process operates at ambient temperature and does not necessitate the formation of a ball. The aluminum wire is inserted through a needle and fastened by the wedge, usually constructed from a hard metal such as tungsten. The wire is initially prepared by severing the excess to establish the appropriate "tail" length protruding outwards. During the bonding process, the needle's sharp edge sandwiches the wire between it and the substrate when the tool is lowered to the bond pad. The ultrasonic vibrations and downward force cause the wire to adhere to the pad, and the tool then moves away to form the loop. Subsequently, the tool pushes the wire down and performs the same bond as the first. Instead of continuing to the loop, it travels to the tail length, closes the wire clamp, pulls, and finalizes the entire bond.

6.5 Thermo-Compression Filp Chip

A thermo-compression flip chip process can be devised using two chip regions with matching pad structures. Specifically, a photonic integrated circuit (PIC) with a wire bond pad

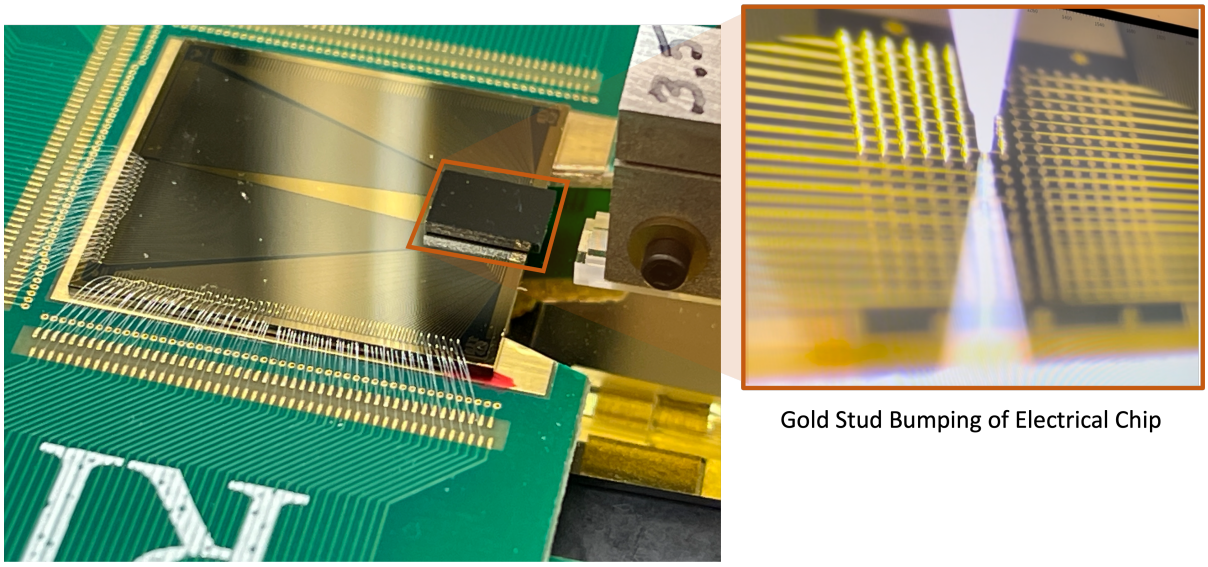

Figure 6.6: A completed package using the thermo-compression flip chip package, showing an inset of the stud-bumping process. This is the process used to assemble the parts in the final section of Chapter 5.

grid and an interposer with identical pads that connect to the external world can be employed. Initially, gold ball bonds are placed on all the pads in the grid of both parts via the stud-bumping method as previously described. A coining step is then performed on the pads using a die-bonding tool, such as the FineTech Lambda 2, to flatten the studs to ensure the levelness of both sets of pads. Next, the PIC is picked up using the same die-bonding tool, flipped over, and aligned with the interposer by utilizing a dichroic mirror that splits the camera's image for simultaneous viewing of both sets of pads. The PIC is then lowered onto the interposer with a force of approximately 80 N to achieve a target force of $0.2 - 0.5$ N per stud, depending on the number of studs being bonded, without exceeding a maximum force of 100 N. The process is held in place, and the temperature is gradually increased from room temperature to 300 degrees Celsius and then back down over three minutes. Upon completion, the PIC is permanently affixed to the interposer. Finally, the interposer is die-bonded onto a printed circuit board (PCB) using epoxy, and

the outer wire bonds are performed following the standard wire bonding procedure. This process was developed primarily to address the large I/O circuit in **??**.

Appendix A

Silicon Photonic Nonlinear Operator

A.1 Design Conception & Initial Results

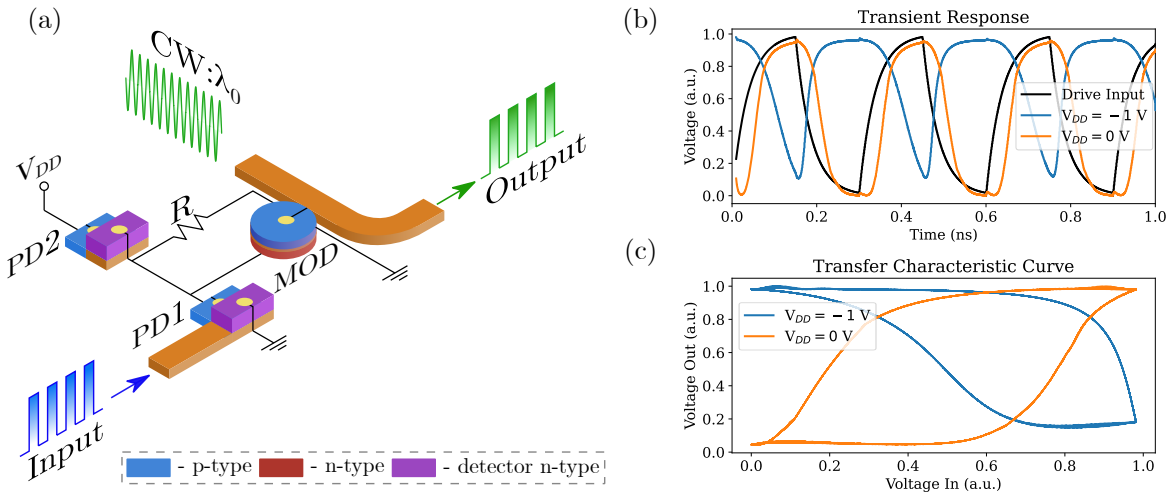

Figure A.1: (a) O/E/O neuron design where PD1 acts as a driver for the voltage that falls upon the MRR PN junction terminals. Note that here the upper PD2 acts as a dummy diode and no optical signal is connected to it. In some implementations, it can be used to provide a bipolar signal. (b) Transient simulation shows the neuron operation at 6.66 Gbit/s, with different biases to V_{DD} of $-1, 0$ V. (c) Transfer Characteristic Curve for the two bias configurations, showing a non-linear regime ideal for operating non-linear activation functions.

Non-linearity is key to realizing integrated photonic deep neural networks. However,

optical nonlinearities of sufficient strength are non-trivial to realize. In contrast, the co-integration of electronic and photonic components has paved the way for faster analog photonic circuits [84,115]. However, characteristics such as linearity and large bandwidth make the realization of non-linear activation functions on a chip a non-trivial task. D. Miller showed [74] that by scaling optoelectronic devices to the micron scale it is possible to realize large effects with low energy usage. Using these principles, here we present an optical-electrical-optical (O/E/O) neuron that realizes the non-linear transfer characteristic that is necessary for neural network algorithms. Electronic/photonic co-simulation is achieved by Verilog-A models of the fundamental devices such as waveguides, directional couplers, free-carrier plasma dispersion effect phase shifters and photodetectors (PDs) [108]. With these basic building blocks, we can design full-scale electronic-photonic circuits, thereby crafting the O/E/O neuron presented. The circuit design that is used to enable an O/E/O neuron is presented in Figure A.1(a). Note that the micro-ring (or micro-disk) resonator (MRR) is designed by connecting a phase shifter to a directional coupler. This circuit is an extension of previous work done on O/E/O neurons in Refs. [84, 115]. PD1 acts as a driver where incident light injects current in the common node. This current is then converted into voltage by the resistor (R) that is connected in parallel with the MRR. As a result, the phase shifter inside the MRR is tuned (by virtue of depletion modulation), shifting the resonance of the ring – allowing for more or less of the continuous wave (CW) input at λ_0 to pass the ring. PD2 acts as a dummy diode, where no optical signal is connected to it. This means that the only purpose of this diode is to supply a bias point for the MRR. Therefore, the supply voltage can be a range of

values that modify the behavior of the neuron to either inverting or non-inverting. Ultimately, this means that light at the output becomes a function of light incident on PD1. It is important to emphasize the fact that optical-to-electrical conversion is done without the need of a trans-impedance amplifier (TIA) [84]. As a result, the ability to scale this architecture is dramatically improved.

The electrical models that represent the PN junction were modified to accurately represent the empirical phase shifter and 'dark' PD I-V curves [108]. The remaining aspects of the models that describe these two modules were not modified, and more details can be found in the cited reference. The modification consisted on adding factors such that the slope changes in the forward and reverse bias regions of operations were accurately represented [31]. Eqs. (A.1) and (A.2) show the models used for the PD and the MRR, respectively. The PD model does not include the low-pass filter response that these typically show as a function of input optical frequency. This is something that may be improved upon in the future.

$$I_D^{PD} = I_s \cdot \left(C_{FB}^{PD}\right) \cdot \left(C_{RB}^{PD}\right) \cdot \left(exp\left(\frac{V_D}{n \cdot V_T}\right) - 1\right) \tag{A.1}$$

$$I_D^{MRR} = I_s \cdot \left(C_{FB}^{MRR}\right) \cdot \left(exp\left(\frac{V_D}{n \cdot V_T}\right) - C_{RB}^{MRR}\right) \tag{A.2}$$

where C_{FB}^{MRR}, C_{RB}^{MRR}, C_{FB}^{PD}, and C_{RB}^{PD} are the added coefficient that handle the derivative changes as the operation regimes transition between forward and reverse bias for the

MRR and the PD, respectively. These are defined by Eqs. A.3 through A.5.

$$C_{FB}^{MRR} = C_{FB}^{PD} = \left(\frac{1}{exp\left(\frac{V_D - V_{TR}}{n_2 \cdot V_T}\right) + 1} \right) \tag{A.3}$$

$$C_{RB}^{MRR} = 2 \cdot exp\left(a_1 \cdot V_D\right) - exp(a_2 \cdot V_D) \tag{A.4}$$

$$C_{RB}^{PD} = \frac{1}{exp\left(\frac{V_{TR} - V_D^{n_e}}{n_{off} \cdot V_T} + 1\right)} \tag{A.5}$$

where I_D is the diode current, I_s is the reverse bias current, V_D is the voltage across the PN junction, n is the ideality factor, V_T is the thermal voltage, V_{TR} is the transition voltage between the reverse and forward bias regions, n_2 is the slope during this transition (this can be related to the thermally injected carriers when V_D remains relatively small; n transitions from a smaller value to a larger value as V_D increases). a_1, a_2, n_e, and n_{off} are parameters fit with MATLAB's curve fitting toolbox [69].

The resistance, R, was set at 1 kΩ, while the zero-bias parasitic capacitance for all three diodes is set to 11 fF. Figure A.1(b) shows the neuron capability for switching on/off with a laser pulse frequency of 6.66 Gbit/s. We can calculate the relationship between the voltage of the input modulated signal and the output detected signal to determine the non-linear transfer characteristic of the O/E/O neuron, shown in A.1 (c). We observe that the neuron is non-linear for the inverting and non-inverting cases, where the latter was achieved with a supply voltage value of $V_{DD} = -1$ V. Further research is needed to define the dynamic limits of the neuron performance as we found that the pulse's rise/fall time affects neuron's transfer characteristics. In conclusion, this work enables the scalable co-simulation of deep photonic neural networks with realistic

nonlinear activation functions.

A.2 Gain Over Unity

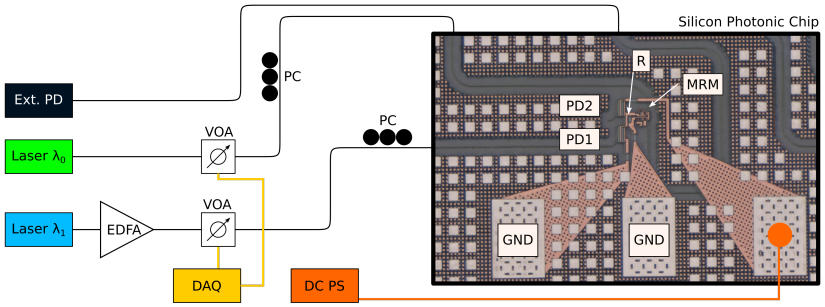

Figure A.2: Experimental setup used to perform DC characterization of the O/E/O conversion circuit. The chip was fabricated in a standard AIM Photonics multi project wafer (MPW) run.

While multiple different non-linear transfer functions are available in the photonic ecosystem, most of them are not suitable for nonlinear activation in neural networks. Characteristics such as $g = \dfrac{\partial P_{out}}{\partial P_{in}} \geq 1$ to maintain fanout, and defined limits such as $\lim_{x \to -\infty} f(x) \leq 0$ for sparsity and $\lim_{x \to \infty} f(x) = a$ for normalization are needed for neuromorphic computing. This work presents a purely silicon photonic circuit that can provide gain over unity ($g \geq 1$) in an open-loop configuration. The circuit shows transistor-like behavior with non-inverting and inverting characteristics akin to PMOS and NMOS devices. Fig. A.2 shows the proposed nonlinear photonic design [96], which utilizes optical-electrical-optical conversion. The optical power incident on the photodetector (PD1) creates a current that biases the resistor (R) connected in parallel with the micro ring modulator (MRM). Since PD2 is reverse biased (by applying a DC bias (DC PS)), the majority of the photo-current passes through R, as it provides the path of least resistance. The current creates a voltage across R, which places the MRM under a new bias.

This change causes the resonance condition of the resonator to change. As the resonance shifts, the CW laser (λ_0) moves in and out of resonance based on the photo-generated current produced by the incident power on PD1. It is essential to recognize that the optical-to-electrical conversion is achieved without the need for a trans-impedance amplifier (TIA). This circuit, therefore, allows for low-overhead, low-power, and scalability (due to gain) for applications such as deep neural networks.

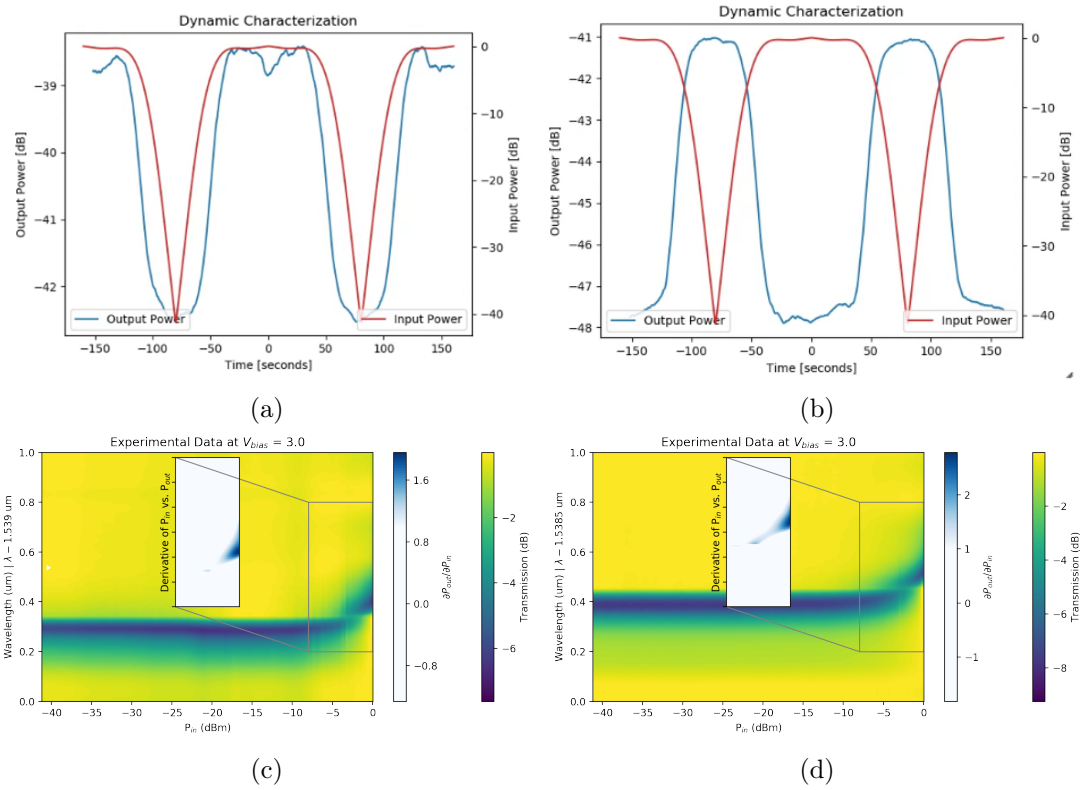

Figure A.3: Quasi-static characterization of the O/E/O conversion circuit with $R = 19.2K\Omega$. Laser wavelength (λ_0) remains constant at $1538.9nm$ and $1539nm$ while its power is swept using the DAQ/VOA setup from $0dB$ to $-40dB$ showing inverting (b) and non-inverting behavior(a) as a function of wavelength. λ and P_{in} sweep for neuron DC characterization and the derivative of the lorentzian fit showing $g \geq 1$ at varying levels of input power for $R = 4.8K\Omega$ (c), and $R = 19.2K\Omega$ (d).

The chip had multiple circuits with varying resistor values (by varying the silicon resistors dimensions and dopings). The resistance values for the circuits presented here

are $R = 4.8K\Omega$, and $R = 19.2K\Omega$, the zero-bias junction capacitance for all three diodes is $\sim$ 11 fF, the PD responsivity is 1 A/W, and the extinction ratio of the MRM is $\sim$ 10 dB. A gain of $g \geq 1$ is achieved with these conditions as shown by Fig. A.3c, and Fig. A.3d. Moreover, the design exhibits both non-inverting and inverting characteristics as a function of wavelength as shown in Fig. A.3a and Fig. A.3b, respectively. The MRM transient behavior causes a discrepancy between obtaining the transfer characteristics using the time-dependent quasi-DC approach presented in these plots compared to the wavelength sweeps shown in Fig. A.3c and Fig. A.3d. This result indicates the need for modulation through a more robust approach than the VOA-based approach presented here. We plan to obtain the neuron's scattering parameters as a function of the frequency of a tone modulating the optical carrier for future work. Further research is needed to establish proper inverter-like transfer characteristics where complete control of the output swing is achieved.

References

[1] Ansys-Lumerical. **CHARGE**: 3D Charge Transport Simulator. `http://lumerical.com`.

[2] Ansys-Lumerical. **FDTD**: 3d Electromagnetic Simulator. `http://lumerical.com`.

[3] Ansys-Lumerical. **MODE**: Waveguide Simulator. `http://lumerical.com`.

[4] Nazca-Design. Photonic Layout Software. `https://nazca-design.org/`.

[5] Qontrol. `https://qontrol.co.uk/`.

[6] Gerhard Abstreiter. Engineering the future of electronics. *Physics World*, 5(3):36, 1992.

[7] Luca Alloatti, Robert Palmer, Sebastian Diebold, Kai Philipp Pahl, Baoquan Chen, Raluca Dinu, Maryse Fournier, Jean-Marc Fedeli, Thomas Zwick, Wolfgang Freude, et al. 100 ghz silicon–organic hybrid modulator. *Light: Science & Applications*, 3(5):e173–e173, 2014.

[8] Ansys-Lumerical. Lumerical inc. `https://www.lumerical.com/`, 2023. [Online; accessed May-2023].

[9] Tom Baehr-Jones, Ran Ding, Yang Liu, Ali Ayazi, Thierry Pinguet, Nicholas C. Harris, Matt Streshinsky, Poshen Lee, Yi Zhang, Andy Eu-Jin Lim, Tsung-Yang Liow, Selin Hwee-Gee Teo, Guo-Qiang Lo, and Michael Hochberg. Ultralow drive voltage silicon traveling-wave modulator. *Opt. Express*, 20(11):12014–12020, May 2012.

[10] B. V. Benjamin, P. Gao, E. McQuinn, S. Choudhary, A. R. Chandrasekaran, J. Bussat, R. Alvarez-Icaza, J. V. Arthur, P. A. Merolla, and K. Boahen. Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations. *Proceedings of the IEEE*, 102(5):699–716, 2014.

[11] Wim Bogaerts, Peter De Heyn, Thomas Van Vaerenbergh, Katrien De Vos, Shankar Kumar Selvaraja, Tom Claes, Pieter Dumon, Peter Bienstman, Dries Van Thourhout, and Roel Baets. Silicon microring resonators. *Laser & Photonics Reviews*, 6(1):47–73, 2012.

[12] Wim Bogaerts, Daniel Pérez, José Capmany, David AB Miller, Joyce Poon, Dirk Englund, Francesco Morichetti, and Andrea Melloni. Programmable photonic circuits. *Nature*, 586(7828):207–216, 2020.

[13] Wim Bogaerts, Dirk Taillaert, Bert Luyssaert, Pieter Dumon, Joris Van Campenhout, Peter Bienstman, Dries Van Thourhout, Roel Baets, V Wiaux, and S Beckx. Basic structures for photonic integrated circuits in silicon-on-insulator. *Optics Express*, 12(8):1583–1591, 2004.

[14] Evangelia Chatzianagnostou, Athanasios Manolis, George Dabos, Dimitra Ketzaki, Amalia Miliou, Nikos Pleros, Laurent Markey, Jean-Claude Weeber, Alain Dereux, Bartos Chmielak, et al. Scaling the sensitivity of integrated plasmo-photonic interferometric sensors. *ACS Photonics*, 6(7):1664–1673, 2019.

[15] Pavel Cheben, Robert Halir, Jens H Schmid, Harry A Atwater, and David R Smith. Subwavelength integrated photonics. *Nature*, 560(7720):565–572, 2018.

[16] Kaixuan Chen, Kyoungsik Yu, and Sailing He. High performance polarization beam splitter based on cascaded directional couplers assisted by effectively anisotropic structures. *IEEE Photonics Journal*, 11(6):1–9, 2019.

[17] Matteo Cherchi, Sami Ylinen, Mikko Harjanne, Markku Kapulainen, and Timo Aalto. Dramatic size reduction of waveguide bends on a micron-scale silicon photonic platform. *Optics express*, 21(15):17814–17823, 2013.

[18] Li Chong, Xue Chun-Lai, Li Ya-Ming, Li Chuan-Bo, Cheng Bu-Wen, and Wang Qi-Ming. High performance silicon waveguide germanium photodetector. *Chinese Physics B*, 24(3):038502, 2015.

[19] Lukas Chrostowski and Michael Hochberg. *Silicon photonics design: from devices to systems*. Cambridge University Press, 2015.

[20] William R. Clements, Peter C. Humphreys, Benjamin J. Metcalf, W. Steven Kolthammer, and Ian A. Walmsley. Optimal design for universal multiport interferometers. *Optica*, 3(12):1460–1465, Dec 2016.

[21] David Coenen, Herman Oprins, Yoojin Ban, Filippo Ferraro, Marianna Pantouvaki, Joris Van Campenhout, and Ingrid De Wolf. Thermal modelling of silicon photonic ring modulator with substrate undercut. *Journal of Lightwave Technology*, 40(13):4357–4363, 2022.

[22] CPU power dissipation. Cpu power dissipation — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/wiki/CPU_power_dissipation#cite_note-6`, 2020. [Online; accessed 9-July-2020].

[23] John E. Cunningham, Ivan Shubin, Xuezhe Zheng, Thierry Pinguet, Attila Mekis, Ying Luo, Hiren Thacker, Guoliang Li, Jin Yao, Kannan Raj, and Ashok V. Krishnamoorthy. Highly-efficient thermally-tuned resonant optical filters. *Opt. Express*, 18(18):19055–19063, Aug 2010.

[24] Rotman David. We're note prepared for the end of moore's law. *MIT Technology Review*, 2020.

[25] Andrea De Iacovo, Andrea Ballabio, Jacopo Frigerio, Lorenzo Colace, and Giovanni Isella. Design and simulation of ge-on-si photodetectors with electrically tunable spectral response. *Journal of Lightwave Technology*, 37(14):3517–3525, 2019.

[26] Thomas Ferreira De Lima, Bhavin J Shastri, Alexander N Tait, Mitchell A Nahmias, and Paul R Prucnal. Progress in neuromorphic photonics. *Nanophotonics*, 6(3):577–599, 2017.

[27] R. H. Dennard, F. H. Gaensslen, H. Yu, V. L. Rideout, E. Bassous, and A. R. LeBlanc. Design of ion-implanted mosfet's with very small physical dimensions. *IEEE Journal of Solid-State Circuits*, 9(5):256–268, 1974.

[28] Rui Ding, Yinlei Huang, Yun Han, Fan Zhang, and Jian-Jun He. High-speed silicon modulators for integrated photonic communication systems. *Nanophotonics*, 8(5):809–838, 2019.

[29] Po Dong, Shirong Liao, Dazeng Feng, Hong Liang, Dawei Zheng, Roshanak Shafiiha, Cheng-Chih Kung, Wei Qian, Guoliang Li, Xuezhe Zheng, et al. Low v pp, ultralow-energy, compact, high-speed silicon electro-optic modulator. *Optics express*, 17(25):22484–22490, 2009.

[30] Ali W Elshaari, Iman Esmaeil Zadeh, Andreas Fognini, Michael E Reimer, Dan Dalacu, Philip J Poole, Val Zwiller, and Klaus D Jöns. On-chip single photon filtering and multiplexing in hybrid quantum photonic circuits. *Nature communications*, 8(1):379, 2017.

[31] D. Marcondes F. et al. A pin diode model for finite-element time-domain simulations. *Journal of Microwaves, Optoelectronics and Electromagnetic Applications (JMOe)*, pages 38S–48S, 2009.

[32] Nicholas M Fahrenkopf, Colin McDonough, Gerald L Leake, Zhan Su, Erman Timurdogan, and Douglas D Coolbaugh. The aim photonics mpw: A highly accessible cutting edge technology for rapid prototyping of photonic integrated circuits. *IEEE Journal of Selected Topics in Quantum Electronics*, 25(5):1–6, 2019.

[33] Monireh Moayedi Pour Fard, Ian AD Williamson, Matthew Edwards, Ke Liu, Sunil Pai, Ben Bartlett, Momchil Minkov, Tyler W Hughes, Shanhui Fan, and Thien-An Nguyen. Experimental realization of arbitrary activation functions for optical neural networks. *Optics Express*, 28(8):12138–12148, 2020.

[34] Lantian Feng, Ming Zhang, Jianwei Wang, Xiaoqi Zhou, Xiaogang Qiang, Guangcan Guo, and Xifeng Ren. Silicon photonic devices for scalable quantum information applications. *Photonics Research*, 10(10):A135–A153, 2022.

[35] S. B. Furber, F. Galluppi, S. Temple, and L. A. Plana. The spinnaker project. *Proceedings of the IEEE*, 102(5):652–665, 2014.

[36] Alexander L. Gaeta, Michal Lipson, and Tobias J. Kippenberg. Photonic-chip-based frequency combs. *Nature Photonics*, 13(3):158–169, Mar 2019.

[37] Genalyte. `https://www.genalyte.com/offering/`, 2020. [Online; accessed 16-July-2020].

[38] I-L Gheorma and RM Osgood. Fundamental limitations of optical resonator based high-speed eo modulators. *IEEE Photonics Technology Letters*, 14(6):795–797, 2002.

[39] George Giamougiannis, Apostolos Tsakyridis, George Mourgias-Alexandris, Miltiadis Moralis-Pegios, Angelina Totovic, George Dabos, Nikos Passalis, Manos Kirtas, Nikos Bamiedakis, Anastasios Tefas, et al. Silicon-integrated coherent neurons with 32gmac/sec/axon compute line-rates using eam-based input and weighting cells. In *2021 European Conference on Optical Communication (ECOC)*, pages 1–4. IEEE, 2021.

[40] Ken Giewont, Karen Nummy, Frederick A. Anderson, Javier Ayala, Tymon Barwicz, Yusheng Bian, Kevin K. Dezfulian, Douglas M. Gill, Thomas Houghton, Shuren Hu, Bo Peng, Michal Rakowski, Stewart Rauch, Jessie C. Rosenberg, Asli Sahin, Ian Stobert, and Andy Stricker. 300-mm monolithic silicon photonics foundry technology. *IEEE Journal of Selected Topics in Quantum Electronics*, 25(5):1–11, 2019.

[41] GRAPHCORE. Graphcore ipu server. `https://www.graphcore.ai/hubfs/Lead%20gen%20assets/DSS8440%20IPU%20Server%20White%20Paper_2020.pdf`, 2015. [Online; accessed 13-July-2020].

[42] R Halir, A Maese-Novo, A Ortega-Moñux, I Molina-Fernández, JG Wangüemert-Pérez, P Cheben, D-X Xu, JH Schmid, and S Janz. Colorless directional coupler with dispersion engineered sub-wavelength structure. *Optics express*, 20(12):13470–13477, 2012.

[43] Robert Halir, Przemek J Bock, Pavel Cheben, Alejandro Ortega-Moñux, Carlos Alonso-Ramos, Jens H Schmid, Jean Lapointe, Dan-Xia Xu, J Gonzalo Wangüemert-Pérez, Íñigo Molina-Fernández, et al. Waveguide sub-wavelength structures: a review of principles and applications. *Laser & Photonics Reviews*, 9(1):25–49, 2015.

[44] Robert Halir, Pavel Cheben, José Manuel Luque-González, Jose Darío Sarmiento-Merenguel, Jens H Schmid, Gonzalo Wangüemert-Pérez, Dan-Xia Xu, Shurui Wang, Alejandro Ortega-Moñux, and Íñigo Molina-Fernández. Ultra-broadband nanophotonic beamsplitter using an anisotropic sub-wavelength metamaterial. *Laser & Photonics Reviews*, 10(6):1039–1046, 2016.

[45] Kyunghun Han, Sangsik Kim, Justin Wirth, Min Teng, Yi Xuan, Ben Niu, and Minghao Qi. Strip-slot direct mode coupler. *Optics express*, 24(6):6532–6541, 2016.

[46] E. Hecht. *Optics*. Pearson education. Addison-Wesley, 2002.

[47] John Heebner, Rohit Grover, Tarek Ibrahim, and Tarek A Ibrahim. *Optical microresonators: theory, fabrication, and applications*, volume 138. Springer Science & Business Media, 2008.

[48] Hao Hu, Xiaolong Xiao, Yu Yu, and Xiaoguang Liu. Resonant silicon modulators for high-speed and energy-efficient optical interconnects: A review. *Journal of Lightwave Technology*, 38(17):4629–4646, 2020.

[49] Wei-Ping Huang. Coupled-mode theory for optical waveguides: an overview. *J. Opt. Soc. Am. A*, 11(3):963–983, Mar 1994.

[50] IBM. Ibm silicon photonics. `https://www.zurich.ibm.com/st/photonics/devices.html`, 2020. [Online; accessed 16-July-2020].

[51] Saman Jahani and Zubin Jacob. Transparent subdiffraction optics: nanoscale light confinement without metal. *Optica*, 1(2):96–100, Aug 2014.

[52] Saman Jahani and Zubin Jacob. Photonic skin-depth engineering. *JOSA B*, 32(7):1346–1353, 2015.

[53] Saman Jahani and Zubin Jacob. All-dielectric metamaterials. *Nature nanotechnology*, 11(1):23, 2016.

[54] Saman Jahani, Sangsik Kim, Jonathan Atkinson, Justin C Wirth, Farid Kalhor, Abdullah Al Noman, Ward D Newman, Prashant Shekhar, Kyunghun Han, Vien Van, et al. Controlling evanescent waves using silicon photonic all-dielectric metamaterials for dense integration. *Nature communications*, 9(1):1–9, 2018.

[55] Aashu Jha, Chaoran Huang, and Paul R Prucnal. Reconfigurable all-optical nonlinear activation functions for neuromorphic photonics. *Optics Letters*, 45(17):4819–4822, 2020.

[56] Norman P Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, et al. In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the 44th Annual International Symposium on Computer Architecture*, pages 1–12, 2017.

[57] Yimin Kang and Yurii A Vlasov. Silicon modulators for high-speed and low-power optical interconnects. *Selected Topics in Quantum Electronics, IEEE Journal of*, 20(4):1–7, 2014.

[58] J. Komma, C. Schwarz, G. Hofmann, D. Heinert, and R. Nawrodt. Thermo-optic coefficient of silicon at 1550 nm and cryogenic temperatures. *Applied Physics Letters*, 101(4):041905, 2012.

[59] J. Koomey, S. Berard, M. Sanchez, and H. Wong. Implications of historical trends in the electrical efficiency of computing. *IEEE Annals of the History of Computing*, 33(3):46–54, 2011.

[60] M.G. Kuzyk. *Nonlinear Optics: A Student's Perspective - With Python Problems and Examples*. NLO Source. CreateSpace Independent Publishing Platform, 2017.

[61] Lightelligence. `https://www.lightelligence.ai`, 2020. [Online; accessed 16-July-2020].

[62] Lightmatter. `https://lightmatter.co`, 2020. [Online; accessed 16-July-2020].

[63] T. Liow, K. Ang, Q. Fang, J. Song, Y. Xiong, M. Yu, G. Lo, and D. Kwong. Silicon modulators and germanium photodetectors on soi: Monolithic integration, compatibility, and performance optimization. *IEEE Journal of Selected Topics in Quantum Electronics*, 16(1):307–315, 2010.

[64] Callum G Littlejohns, David J Rowe, Han Du, Ke Li, Weiwei Zhang, Wei Cao, Thalia Dominguez Bucio, Xingzhao Yan, Mehdi Banakar, Dehn Tran, et al. Cornerstone's silicon photonics rapid prototyping platforms: Current status and future outlook. *Applied Sciences*, 10(22):8201, 2020.

[65] José Manuel Luque-González, Alaine Herrero-Bermello, Alejandro Ortega-Moñux, Íñigo Molina-Fernández, Aitor V Velasco, Pavel Cheben, Jens H Schmid, Shurui Wang, and Robert Halir. Tilted subwavelength gratings: controlling anisotropy in metamaterial nanophotonic waveguides. *Optics letters*, 43(19):4691–4694, 2018.

[66] Alexander I Lvovsky, Barry C Sanders, and Wolfgang Tittel. Optical quantum memory. *Nature photonics*, 3(12):706–714, 2009.

[67] Sasikanth Manipatruni, Ari Novack, Kunal Shastri, and Ashok V Krishnamoorthy. Silicon photonics modulators: a review of recent advances. *Reports on Progress in Physics*, 84(4):046501, 2021.

[68] Riccardo Marchetti, Cosimo Lacava, Lee Carroll, Kamil Gradkowski, and Paolo Minzioni. Coupling strategies for silicon photonics integrated chips *Invited. Photon. Res.*, 7(2):201–239, Feb 2019.

[69] MATLAB. *Curve Fitting Toolbox, version 9.6.0 (R2020b)*. The MathWorks Inc., 2020.

[70] WR McKinnon, D-X Xu, C Storey, E Post, A Densmore, A Delâge, P Waldron, JH Schmid, and S Janz. Extracting coupling and loss coefficients from a ring resonator. *Optics express*, 17(21):18971–18982, 2009.

[71] Paul A. Merolla, John V. Arthur, Rodrigo Alvarez-Icaza, Andrew S. Cassidy, Jun Sawada, Filipp Akopyan, Bryan L. Jackson, Nabil Imam, Chen Guo, Yutaka Nakamura, Bernard Brezzo, Ivan Vo, Steven K. Esser, Rathinakumar Appuswamy, Brian Taba, Arnon Amir, Myron D. Flickner, William P. Risk, Rajit Manohar, and Dharmendra S. Modha. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science*, 345(6197):668–673, 2014.

[72] Md Borhan Mia, Syed Z Ahmed, Ishtiaque Ahmed, Yun Jo Lee, Minghao Qi, and Sangsik Kim. Exceptional coupling in extreme skin-depth waveguides for extremely low waveguide crosstalk. *arXiv preprint arXiv:2002.10579*, 7(8):881–887, Aug 2020.

[73] D. A. B. Miller. Device requirements for optical interconnects to silicon chips. *Proceedings of the IEEE*, 97(7):1166–1185, 2009.

[74] David AB Miller. Are optical transistors the logical next step? *Nature Photonics*, 4(1):3–5, 2010.

[75] George Mourgias-Alexandris, A Totovic, N Passalis, George Dabos, Anastasios Tefas, and N Pleros. Neuromorphic computing through photonic integrated circuits. In *Smart Photonic and Optoelectronic Integrated Circuits XXII*, volume 11284, page 1128403. International Society for Optics and Photonics, 2020.

[76] George Mourgias-Alexandris, Angelina Totović, Apostolos Tsakyridis, Nikolaos Passalis, Konstantinos Vyrsokinos, Anastasios Tefas, and Nikos Pleros. Neuromorphic photonics with coherent linear neurons using dual-iq modulation cells. *Journal of Lightwave Technology*, 38(4):811–819, 2019.

[77] George Mourgias-Alexandris, A Tsakyridis, N Passalis, Anastasios Tefas, K Vyrsokinos, and Nikolaos Pleros. An all-optical neuron with sigmoid activation function. *Optics express*, 27(7):9620–9630, 2019.

[78] Xin Mu, Sailong Wu, Lirong Cheng, and H.Y. Fu. Edge couplers in silicon photonic integrated circuits: A review. *Applied Sciences*, 10(4), 2020.

[79] Vaishnavi Murthy, Anthony Rizzo, Gerald Leake, Nandish Mehta, Asher Novick, Stuart Daudlin, Maarten Haatnik, Matthew van Niekerk, Michael Fanto, Daniel Coleman, et al. Mitigation of parasitic junction formation in compact resonant modulators with doped silicon heaters. In *Laser Resonators, Microresonators, and Beam Control XXIV*, volume 11987, pages 114–125. SPIE, 2022.

[80] Mitchell A Nahmias, Bhavin J Shastri, Alexander N Tait, and Paul R Prucnal. A leaky integrate-and-fire laser neuron for ultrafast cognitive computing. *IEEE journal of selected topics in quantum electronics*, 19(5):1–12, 2013.

[81] Makoto Nakai, Tsuyoshi Nomura, SungWon Chung, and Hossein Hashemi. Geometric loss reduction in tight bent waveguides for silicon photonics. In *CLEO: Science and Innovations*, pages JW2A–70. Optica Publishing Group, 2018.

[82] M. Nedeljkovic, R. Soref, and G. Z. Mashanovich. Free-carrier electrorefraction and electroabsorption modulation predictions for silicon over the 1–14- μm infrared wavelength range. *IEEE Photonics Journal*, 3(6):1171–1180, 2011.

[83] Milos Nedeljkovic, Richard Soref, and Goran Mashanovich. Free-carrier electrorefraction and electroabsorption modulation predictions for silicon over the 1–14- infrared wavelength range. *IEEE Photonics Journal*, 3:1171–1180, 12 2011.

[84] K. Nozaki, S. Matsuo, A. Shinya, and M. Notomi. Amplifier-free bias-free receiver based on low-capacitance nanophotodetector. *IEEE Journal of Selected Topics in Quantum Electronics*, 24(2):1–11, 2018.

[85] NVIDIA. Nvidia tesla v100 gpu architecture. `https://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf`, 2017. [Online; accessed 13-July-2020].

[86] Hyundai Park, Matthew N. Sysak, Hui-Wen Chen, Alexander W. Fang, Di Liang, Ling Liao, Brian R. Koch, Jock Bovington, Yongbo Tang, Kristi Wong, Matt Jacob-Mitos, Richard Jones, and John E. Bowers. Device and integration technology for silicon photonic transmitters. *IEEE Journal of Selected Topics in Quantum Electronics*, 17(3):671–688, 2011.

[87] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, et al. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[88] Alexander Y Piggott, Jesse Lu, Konstantinos G Lagoudakis, Jan Petykiewicz, Thomas M Babinec, and Jelena Vučković. Inverse design and demonstration of a compact and broadband on-chip wavelength demultiplexer. *Nature Photonics*, 9(6):374–377, 2015.

[89] Milos A. Popovic. Theory and design of high-index-contrast microphotonic circuits. *Ph.D. Thesis*, 2007.

[90] Paul R Prucnal and Bhavin J Shastri. *Neuromorphic photonics*. CRC Press, 2017.

[91] Paul R. Prucnal, Bhavin J. Shastri, Thomas Ferreira de Lima, Mitchell A. Nahmias, and Alexander N. Tait. Recent progress in semiconductor excitable lasers for photonic spike processing. *Adv. Opt. Photon.*, 8(2):228–299, Jun 2016.

[92] PsiQuantum. Building the world's first useful quantum computer. `https://psiquantum.com/news/building-the-worlds-first-useful-quantum-computer/`, 2020. [Online; accessed 16-July-2020].

[93] Michael Reck, Anton Zeilinger, Herbert J. Bernstein, and Philip Bertani. Experimental realization of any discrete unitary operator. *Phys. Rev. Lett.*, 73:58–61, Jul 1994.

[94] Graham T Reed, G Mashanovich, F Yand Gardes, and DJ Thomson. Silicon optical modulators. *Nature photonics*, 4(8):518–526, 2010.

[95] Andrew Rickman. The commercialization of silicon photonics. *Nature Photonics*, 8(8):579–582, 2014.

[96] Hector A Rubio Rivera, Matthew van Niekerk, and Stefan F Preble. Silicon photonic optical-electrical-optical modulator neuron verilog-a model. In *CLEO: QELS_ Fundamental Science*, pages JTu3A–125. Optica Publishing Group, 2021.

[97] Anthony Rizzo, Asher Novick, Vignesh Gopal, Bok Young Kim, Xingchen Ji, Stuart Daudlin, Yoshitomo Okawachi, Qixiang Cheng, Michal Lipson, Alexander L Gaeta, et al. Integrated kerr frequency comb-driven silicon photonic transmitter. *arXiv preprint arXiv:2109.10297*, 2021.

[98] Karl Rupp. Microproccesor-trend-data. `https://github.com/karlrupp/microprocessor-trend-data`, 2020.

[99] S Rytov. Electromagnetic properties of a finely stratified medium. *Soviet Physics JEPT*, 2:466–475, 1956.

[100] Johannes Schemmel, Laura Kriener, Paul Müller, and Karlheinz Meier. An accelerated analog neuromorphic hardware system emulating nmda-and calcium-based non-linear dendrites. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2217–2226. IEEE, 2017.

[101] Bhavin J Shastri, Alexander N Tait, T Ferreira de Lima, Wolfram HP Pernice, Harish Bhaskaran, C David Wright, and Paul R Prucnal. Photonics for artificial intelligence and neuromorphic computing. *Nature Photonics*, 15(2):102–114, 2021.

[102] Bhavin J Shastri, Alexander N Tait, Thomas Ferreira de Lima, Mitchell A Nahmias, Hsuan-Tung Peng, and Paul R Prucnal. Principles of neuromorphic photonics. *arXiv preprint arXiv:1801.00016*, 2017.

[103] Bing Shen, Peng Wang, Randy Polson, and Rajesh Menon. An integrated-nanophotonics polarization beamsplitter with 2.4× 2.4 μm 2 footprint. *Nature Photonics*, 9(6):378–382, 2015.

[104] Yichen Shen, Nicholas C Harris, Scott Skirlo, Mihika Prabhu, Tom Baehr-Jones, Michael Hochberg, Xin Sun, Shijie Zhao, Hugo Larochelle, Dirk Englund, et al. Deep learning with coherent nanophotonic circuits. *Nature Photonics*, 11(7):441, 2017.

[105] SiPhox. `https://www.siphox.bio/#COVID`, 2020. [Online; accessed 16-July-2020].

[106] B.W. Smith, K. Suzuki, and J.R. Sheats. *Microlithography: Science and Technology*. Taylor & Francis, 1998.

[107] Jeong Hwan Song, Tangla D Kongnyuy, Peter De Heyn, Sebastien Lardenois, Roelof Jansen, and Xavier Rottenberg. Low-loss waveguide bends by advanced shape for photonic integrated circuits. *Journal of Lightwave Technology*, 38(12):3273–3279, 2020.

[108] Cheryl Sorace-Agaskar et al. Electro-optical co-simulation for integrated cmos photonic circuits with veriloga. *Opt. Express*, 2015.

[109] Richard A Soref. Silicon-based optoelectronics. *Proceedings of the IEEE*, 81(12):1687–1706, 1993.

[110] Isabelle Staude and Jörg Schilling. Metamaterial-inspired silicon nanophotonics. *Nature Photonics*, 11(5):274, 2017.

[111] Sebastian Strunck, Onur Hamza Karabey, Christian Weickhmann, Alexander Gaebler, and Rolf Jakoby. Continuously tunable phase shifters for phased arrays based on liquid crystal technology. In *2013 IEEE International Symposium on Phased Array Systems and Technology*, pages 82–88. IEEE, 2013.

[112] X. Sui, Q. Wu, J. Liu, Q. Chen, and G. Gu. A review of optical neural networks. *IEEE Access*, 8:70773–70783, 2020.

[113] Haoyang Sun, Qifeng Qiao, Qingze Guan, and Guangya Zhou. Silicon photonic phase shifters and their applications: A review. *Micromachines*, 13(9):1509, 2022.

[114] Peng Sun and Ronald M. Reano. Submilliwatt thermo-optic switches using free-standing silicon-on-insulator strip waveguides. *Opt. Express*, 18(8):8406–8411, Apr 2010.

[115] Alexander N Tait, Thomas Ferreira De Lima, Mitchell A Nahmias, Heidi B Miller, Hsuan-Tung Peng, Bhavin J Shastri, and Paul R Prucnal. Silicon photonic modulator neuron. *Physical Review Applied*, 11(6):064043, 2019.

[116] Alexander N Tait, Thomas Ferreira De Lima, Ellen Zhou, Allie X Wu, Mitchell A Nahmias, Bhavin J Shastri, and Paul R Prucnal. Neuromorphic photonic networks using silicon photonic weight banks. *Scientific reports*, 7(1):1–10, 2017.

[117] Alexander N Tait, Mitchell A Nahmias, Bhavin J Shastri, and Paul R Prucnal. Broadcast and weight: an integrated network for scalable photonic spike processing. *Journal of Lightwave Technology*, 32(21):4029–4041, 2014.

[118] Alexander N. Tait, Allie X. Wu, Thomas Ferreira de Lima, Mitchell A. Nahmias, Bhavin J. Shastri, and Paul R. Prucnal. Two-pole microring weight banks. *Opt. Lett.*, 43(10):2276–2279, May 2018.

[119] Nicholas Thomas-Peter, Nathan K Langford, Animesh Datta, Lijian Zhang, Brian J Smith, Justin B Spring, Benjamin J Metcalf, Hendrik B Coldenstrodt-Ronge, Michael Hu, Joshua Nunn, et al. Integrated photonic sensing. *New Journal of Physics*, 13(5):055024, 2011.

[120] E. Timurdogan. Wafer-scale integrated active silicon photonics for manipulation and conversion of light. 2016.

[121] Erman Timurdogan, Cheryl M Sorace-Agaskar, Ehsan Shah Hosseini, and Michael R Watts. An interior-ridge silicon microring modulator. *Journal of lightwave technology*, 31(24):3907–3914, 2013.

[122] Erman Timurdogan, Cheryl M Sorace-Agaskar, Jie Sun, Ehsan Shah Hosseini, Aleksandr Biberman, and Michael R Watts. An ultralow power athermal silicon modulator. *Nature communications*, 5(1):1–11, 2014.

[123] Angelina Totovic, George Giamougiannis, Apostolos Tsakyridis, David Lazovsky, and Nikos Pleros. Programmable photonic neural networks combining wdm with coherent linear optics. *Scientific Reports*, 12(1):1–13, 2022.

[124] Angelina R Totović, George Dabos, Nikolaos Passalis, Anastasios Tefas, and Nikos Pleros. Femtojoule per mac neuromorphic photonics: An energy and technology roadmap. *IEEE Journal of Selected Topics in Quantum Electronics*, 26(5):1–15, 2020.

[125] Alan Turing. Lecture on the automatic computing engine (1947). *B. Jack Copeland*, page 362, 2004.

[126] Alfred B U'Ren, Konrad Banaszek, and Ian A Walmsley. Photon engineering for quantum information processing. *arXiv preprint quant-ph/0305192*, 2003.

[127] S. van der Walt, S. C. Colbert, and G. Varoquaux. The numpy array: A structure for efficient numerical computation. *Computing in Science Engineering*, 13(2):22–30, 2011.

[128] Matthew van Niekerk, Saman Jahani, Justin Bickford, Pak Cho, Stephen Anderson, Gerald Leake, Daniel Coleman, Michael L Fanto, Christopher C Tison, Gregory A Howland, et al. Two-dimensional extreme skin depth engineering for cmos photonics. *arXiv preprint arXiv:2005.14265*, 38(4):1307–1316, 2020.

[129] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake Vand erPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1. 0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[130] John Von Neumann. Probabilistic logics and the synthesis of reliable organisms from unreliable components. *Automata studies*, 34:43–98, 1956.

[131] John Von Neumann. First draft of a report on the edvac. *IEEE Annals of the History of Computing*, 15(4):27–75, 1993.

[132] M Mitchell Waldrop. The chips are down for moore's law. *Nature News*, 530(7589):144, 2016.

[133] Michael R Watts, William A Zortman, Douglas C Trotter, Ralph W Young, and Anthony L Lentine. Vertical junction silicon microdisk modulators and switches. *Optics express*, 19(22):21989–22003, 2011.

[134] Wouter J Westerveld and H Paul Urbach. *Silicon Photonics Electromagnetic Theory*. IOP Publishing, 2017.

[135] I. A. D. Williamson, T. W. Hughes, M. Minkov, B. Bartlett, S. Pai, and S. Fan. Reprogrammable electro-optic nonlinear activation functions for optical neural networks. *IEEE Journal of Selected Topics in Quantum Electronics*, 26(1):1–12, 2020.

[136] Ian AD Williamson, Tyler W Hughes, Momchil Minkov, Ben Bartlett, Sunil Pai, and Shanhui Fan. Reprogrammable electro-optic nonlinear activation functions for optical neural networks. *IEEE Journal of Selected Topics in Quantum Electronics*, 26(1):1–12, 2019.

[137] Yufei Xing, Umar Khan, Antônio Ribeiro Alves Júnior, and Wim Bogaerts. Behavior model for directional coupler. In *Proceedings Symposium IEEE Photonics Society Benelux*, pages 128–131, 2017.

[138] Hongnan Xu, Daoxin Dai, and Yaocheng Shi. Anisotropic metamaterial-assisted all-silicon polarizer with 415-nm bandwidth. *Photonics Research*, 7(12):1432–1439, 2019.

[139] Qianfan Xu, Brad Schmidt, Jagat Shakya, and Michal Lipson. Cascaded silicon micro-ring modulators for wdm optical interconnection. *Optics express*, 14(20):9431–9436, 2006.

[140] Qianfan Xu, Bradley Schmidt, Sameer Pradhan, and Michal Lipson. Micrometre-scale silicon electro-optic modulator. *nature*, 435(7040):325–327, 2005.

[141] H Zhang, M Gu, XD Jiang, J Thompson, H Cai, S Paesani, R Santagati, A Laing, Y Zhang, MH Yung, et al. An optical neural chip for implementing complex-valued neural network. *Nature Communications*, 12(1):1–11, 2021.

[142] Jiadi Zhu, Teng Zhang, Yuchao Yang, and Ru Huang. A comprehensive review on emerging artificial neuromorphic devices. *Applied Physics Reviews*, 7(1):011312, 2020.