DISSERTATION

DESIGN EXPLORATION AND OPTIMIZATION OF SILICON PHOTONIC INTEGRATED CIRCUITS UNDER FABRICATION-PROCESS VARIATIONS

Submitted by

Asif Anwar Baig Mirza

Department of Electrical and Computer Engineering

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2024

Doctoral Committee:

Advisor: Mahdi Nikdast
Co-Advisor: Sudeep Pasricha

Jesse Wilson
Samuel Brewer

ABSTRACT

DESIGN EXPLORATION AND OPTIMIZATION OF SILICON PHOTONIC INTEGRATED CIRCUITS UNDER FABRICATION-PROCESS VARIATIONS

Silicon photonic integrated circuits (PICs) have become a key solution to handle the growing demands of large data transmission in emerging applications by consuming less power and low heat dissipation while offering ultra-high data bandwidth than electronic circuits. With Moore's Law slowing down and the end of Dennard scaling, PICs offer a logical step to improve data movement and processing performance in future computing systems. On PICs, light is processed and routed by means of optical waveguides. Silicon has a unique feature of high refractive index contrast in the silicon-on-insulator (SOI) platform which allows for tight confinement of light in nanometer waveguide cores and bends with a radius of only a few microns. PICs comprise of a diverse set of elements such as waveguide splitters, combiners, crossings, and couplers which help with distribution, routing, and computation of optical signals. Optical signals are converted to electrical signals with the help of photodiodes which in silicon photonics are implemented using Germanium. To enable PICs for wavelength-division multiplexing (WDM), there is a need for efficient wavelength filters consisting of optical delay lines or resonators. Optical delay lines are usually built using Mach-Zehnder Interferometers (MZIs) which consists of a splitter, two waveguides with a given group delay, and a combiner. Other devices such as microring resonators (MRRs) can be used as wavelength filters when the input wavelength matches a whole multiple times in the circumference of the ring. Other components such as grating coupler help couple the light into and out of a PIC. PICs can be fabricated on the infrastructure developed for complimentary metal–oxide–semiconductor (CMOS) electronics. This technology now enables deep submicron features with unprecedented accuracy in large volumes along with close integration of photonics

and electronic circuits. The use of silicon as a base material makes reuse of these manufacturing tools possible, but photonics imposes different demands on the processes.

Although silicon photonics offers data transmission and computation at light speed with high bandwidth and low power consumption, the fundamental building blocks in PICs (e.g., optical waveguides) are extremely sensitive to nanometer-scale fabrication-process variations (FPVs) caused due to slight randomness in optical lithography processes. Active compensation by means of electronic circuits (a.k.a. tuning) is necessary to compensate for FPVs. Tunable microheaters can be used for active compensation which affect the material properties of silicon to improve PIC's performance under FPVs. However, the total power consumed due to tuning in a working PIC can be drastically high. For example, variations as small as 1 nm in an MRR can deviate the optical frequency response of the device by 2 nm that leads to approximately 25% increase in the tuning power consumption to compensate for variations of a single MRR. Additionally, a system can have thousands of such MRRs that can easily add up the total power consumption of the system. In order to address FPVs we need to observe the reliability not just at a system level but down to the device level by enabling reliable, FPV-aware devices to enable FPV-resilient PICs and photonic systems. Designing more reliable and FPV-tolerant photonic devices should not only help us with reducing the total power consumption but also build more reliable circuits with fault-free operational behavior for data transmission and computation in future computing systems.

This PhD thesis covers the impact of process variations on photonic devices primarily MRRs. We take a bottom-up approach in improving the reliability of an MRR towards FPVs. We propose an improved and optimized MRR designs which can be used in any PIC to reduce the overall shift in resonant wavelength of the device due to FPVs, further reducing the total power consumption required to tune the device. We confirmed our findings by further fabricating such MRRs and comparing the improved and optimized designs against conventional MRRs. Furthermore, we study the impact these improved MRRs have in photonic artificial intelligence (AI) accelerators and how they can further improve the network accuracy and overall power consumption. Finally, we also compile our work into a device exploration tool that allows photonic designer to set design

parameters in an MRR and study its behavior under different FPV profiles. With this tool we aim to give the designer the ability to determine desired MRR designs based on desired design and performance requirements and budget constraints set on a photonic system.

ACKNOWLEDGEMENTS

I am deeply grateful to my advisors, Dr. Mahdi Nikdast and Dr. Sudeep Pasricha, for their unwavering support, guidance, and expertise throughout my PhD journey. Their mentorship has been instrumental in shaping my research and academic growth. I also extend my sincere thanks to my committee members, Dr. Jesse Wilson and Dr. Samuel Brewer for their thoughtful feedback and insightful questions, which have greatly enhanced this thesis. I am grateful to Colorado State University and the National Science Foundation under grant CCF-1813370, CCF-2006788, and CNS-2046226 generous financial support, which made this research possible.

I am deeply grateful to the all my industry mentors who guided and supported me throughout my various internships during my PhD journey. Especially Dr. Leimeng Zhuang, Dr. Sriharsha Kota, and Dr. Matthew Sysak for my internships with imec and Ayar Labs. Your willingness to share your expertise, provide insightful feedback and provide career advice has been invaluable in shaping my professional development. I am particularly thankful for the opportunities you provided to apply my academic knowledge to real-world challenges, which has enriched my understanding of my field and prepared me for the next chapter of my career. I am deeply grateful to my colleagues Vipin, Saideep, Febin, Kamil, Ebad, Amin S., Sanmitra, Rashad, Zahra, and Amin M. Your support and encouragement have been invaluable throughout my PhD journey. Your dedication to your own work has served as a constant source of inspiration. I am incredibly fortunate to have shared this experience with such a talented and supportive group of individuals.

Finally, I extend my heartfelt gratitude to my family, Rakhib, Hafeez, and Asfia, for their unconditional love, encouragement, support, and unwavering belief in me. And to the very long list of friends, Aarti, Pavitra, Krishna, Dinesh, Neha, Kavitha, Manasa, Mey, Vivek, Delaney, Rena, Saloni, Ben E., Alana, Marilynn, Ashton, Sandeep, Karthik, Ben P., Bryce, Sofia, Ren, Ethan, Tala, Matthew, Lucy, Elizabeth, Kendal, and Jack thank you for all the laughter, companionship, and countless moments of encouragement. You have truly seen me go through everything in closer proximity. Your friendship has been invaluable.

LIST OF RESEARCH PUBLICATIONS

- **A. Mirza**, R. E. Gloekler, J. Thompson, S. Pasricha, and M. Nikdsast, "Experimental Analysis of Adiabatic Silicon Photonic Microring Resonators under Process Variations," IEEE Photonics Technology Letters (PTL), 2024.

- **A. Mirza**, A. Shafiee, S. Banerjee, K. Chakrabarty, S. Pasricha, and M. Nikdast, "Characterization and optimization of coherent MZI-based nanophotonic neural networks under fabrication non-uniformity," IEEE Transactions on Nanotechnology (TNANO), vol. 21, pp. 763–771, 2022.

- **A. Mirza**, F. Sunny, P. Walsh, K. Hassan, S. Pasricha, and M. Nikdast, "Silicon photonic microring resonators: A comprehensive design-space exploration and optimization under fabrication-process variations," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), vol. 41, no. 10, pp. 3359–3372, 2022.

- F. Sunny, **A. Mirza**, M. Nikdast, and S. Pasricha, "ROBIN: A robust optical binary neural network Accelerator," ACM Transactions on Embedded Computing Systems (TECS), vol. 20, no. 5s, Article no. 57, pp. 1–24, October 2021.

- F. Sunny, **A. Mirza**, I. Thakkar, M. Nikdast, and S. Pasricha, "ARXON: A framework for approximate communication over photonic networks-on-chip," IEEE Transactions on Very Large Scale Integration Systems (TVLSI), vol. 29, no. 6, pp. 1206-1219, June 2021.

- Sysak M, Roucka R, Raval M, Luna F, El-Henawy S, Frey J, Li C, Zhang C, Pavan SK, **A. Mirza**, Yang LF. High Wavelength Count Laser Sources for WDM CMOS Optical Interconnects. 28th International Semiconductor Laser Conference (ISLC) 2022 Oct 16 (pp. 1-2). IEEE.

- A. Shafiee, **A. Mirza**, F. Sunny, S. Banerjee, K. Chakrabarty, S. Pasricha, and M. Nikdast, "Inexact silicon photonics: From devices to applications," OSA Photonics in Switching and Computing (PSC) Conference, September 2021, paper M3C.2. (Invited)

- F. Sunny, **A. Mirza**, M. Nikdast, and S. Pasricha, "ROBIN: A robust optical binary neural network accelerator," IEEE/ACM International Conference on Compilers, Architectures, and Synthesis for Embedded Systems (CASES), October 2021.

- F. Sunny, **A. Mirza**, M. Nikdast, and S. Pasricha, "CrossLight: A cross-layer optimized silicon photonic neural network accelerator," IEEE/ACM Design Automation Conference (DAC), San Francisco, CA, December 2021, pp. 1069-1074.

- **A. Mirza**, S. Pasricha, and M. Nikdast, "Variation-aware inter-device matching in silicon photonic microring resonator demultiplexers," IEEE Photonics Conference (IPC), 2020, pp. 1-2.

- F. Sunny, **A. Mirza**, I. Thakkar, S. Pasricha, and M. Nikdast, "LORAX: Loss-aware approximations for energy-efficient silicon photonic networks-on-chip", ACM Great Lakes Symposium on VLSI (GLSVLSI), September 2020, pp. 235-240.

- **A. Mirza**, F. Sunny, S. Pasricha, and M. Nikdast, "Silicon photonic microring resonators: Design optimization under fabrication non-uniformity", IEEE/ACM Design, Automation and Test in Europe (DATE) Conference and Exhibition, Grenoble, France, March 2020, pp. 484-489.

- **A. Mirza**, S. Manafi Avari, E. Taheri, S. Pasricha, and M. Nikdast, "Opportunities for cross-layer design in high-performance computing systems with integrated silicon photonic networks", IEEE/ACM Design, Automation and Test in Europe (DATE) Conference and Exhibition, Grenoble, France, March 2020, pp. 1622-1627.

TABLE OF CONTENTS

ABSTRACT ii
ACKNOWLEDGEMENTS v
LIST OF RESEARCH PUBLICATIONS vi
LIST OF TABLES xi
LIST OF FIGURES xii

Chapter 1 Introduction 1
1.1 Silicon Photonic Devices 5
1.1.1 Optical Waveguides 5
1.1.2 Photonic Microring Resonators (MRRs) 6
1.1.3 Mach-Zehnder Interferometers (MZIs) 9
1.1.4 Other Optical Devices 11
1.2 Photonic Accelerators 11
1.3 Fabrication-Process Variations 13
1.4 Dissertation Overview 15

Chapter 2 Fabrication-Process Variation Analysis in Silicon Photonics[1] 17
2.1 Overview 17
2.2 Related Work 19
2.3 Virtual Wafer-Level FPV Maps 21
2.4 Conclusion 24

Chapter 3 Microring Resonators under Fabrication-Process Variations[1] 26
3.1 Introduction 26
3.2 Related Work 28
3.3 Modeling Silicon Photonic MRR Under Fabrication-Process Variation . . . 30
3.4 MRR Design-Space Exploration under Fabrication-Process Variations . . . 35
3.4.1 MRR performance analysis when $w_i = w_r$ 38
3.4.2 MRR performance analysis when $w_i \neq w_r$ 41
3.4.3 Fabrication Results 43
3.5 MRR Design Optimization under Fabrication-Process Variations 45
3.6 Experimental Analysis of Adiabatic Silicon Photonic Microring Resonators under Process Variations[2] 51
3.6.1 Adiabatic MRR Design and Analysis 52
3.6.2 Experimental Results and Discussion 54
3.7 Conclusion 59

Chapter 4 Applications of FPV-aware Optimized Photonic Devices 60
4.1 Introduction 60
4.2 A Wavelength-Selective MRR-Based Demultiplexer[1] 62

4.3 CrossLight: A Cross-Layer Optimized Silicon Photonic Neural Network . . . 68
4.3.1 Background and Related Work . . . 70
4.3.2 Noncoherent Photonic Computation Overview . . . 71
4.3.3 Crosslight Architecture . . . 74
4.3.3.1 MRR device engineering and fabrication . . . 75
4.3.3.2 Tuning circuit design . . . 77
4.3.3.3 Architecture design . . . 79
4.3.3.3.1 Decomposing vector operations in CONV/FC layers . . . 79
4.3.3.3.2 Vector dot product (VDP) unit design . . . 81
4.3.3.3.3 Optical wavelength reuse in VDP units . . . 82
4.3.4 Evaluation and Simulation Results . . . 83
4.3.4.1 Simulation setup . . . 83
4.3.4.2 Results: *CrossLight* resolution analysis . . . 84
4.3.4.3 Results: *CrossLight* sensitivity analysis . . . 86
4.3.4.4 Results: Comparison with state-of-the-art accelerators . . . 86
4.3.5 Conclusion . . . 89
4.4 ROBIN: A Robust Optical Binary Neural Network Accelerator . . . 89
4.4.1 Related Work . . . 91
4.4.2 Overview of Noncoherent Optical Computation . . . 93
4.4.3 Binarized Neural Network . . . 95
4.4.4 *ROBIN* Architecture . . . 96
4.4.4.1 Tuning Circuit Design . . . 97
4.4.4.2 Device-Level Optimization . . . 99
4.4.4.2.1 Fabrication-Process Variation Resilience . . . 99
4.4.4.2.2 Multi-Bit Precision MRRs . . . 100
4.4.4.2.3 Single-bit MRRs . . . 102
4.4.4.2.4 Broadband MRRs . . . 102
4.4.4.3 Architecture Design . . . 104
4.4.4.3.1 Vector dot Product (VDP) Unit Design . . . 104
4.4.4.3.2 Optical Wavelength Reuse in VDP Units . . . 106
4.4.4.3.3 *ROBIN* Pipeline and Scheduling . . . 107
4.4.5 Results . . . 109
4.4.5.1 Simulation Setup . . . 109
4.4.5.2 Fabrication-Process Variation Analysis . . . 111
4.4.5.3 *ROBIN* Architecture Optimization Analysis . . . 114
4.4.5.4 Comparison with State-of-the-art Optical and Electronic DNN/BNN Accelerators . . . 114
4.4.5.5 Comparison to CPU Based Inference . . . 118
4.4.6 Conclusion . . . 119
4.5 Characterization and Optimization of Coherent MZI-based Nanophotonic Neural Networks under Fabrication Non-Uniformity[3] . . . 120
4.5.1 Background and Related Work . . . 121

4.5.1.1 Mach–Zehnder Interferometer (MZI) . . . 121
4.5.1.2 Coherent SPNNs based on MZIs . . . 123
4.5.1.3 Fabrication-Process Variations (FPVs) . . . 124
4.5.1.4 Related Work on FPV Analysis in SPNNs . . . 125
4.5.2 Modeling FPVs in Coherent SPNNs . . . 126
4.5.2.1 Device Level: MZI Performance under FPVs . . . 126
4.5.2.2 Network Level: OIU Performance under FPVs . . . 128
4.5.3 SPNN Design Optimization Under FPVs . . . 130
4.5.4 Simulation Results and Discussions . . . 134
4.5.5 Conclusion . . . 139

Chapter 5 ProVAT: An Automated Design and Analysis Framework for Process-Variation-Resilient Design of Silicon Photonic Microring Resonators . . . 141
5.1 Introduction . . . 141
5.2 Proposed Design Framework . . . 142
5.2.1 FPV Maps . . . 144
5.2.2 Waveguide Analysis . . . 144
5.2.3 MRR Analysis . . . 146
5.3 Results and Discussions . . . 148
5.4 Conclusion . . . 150

Chapter 6 Conclusion and Future Work . . . 152

Bibliography . . . 158

Appendix A Cross-Over Coupling in Unconventional MRRs . . . 176

Appendix B Impact of Radius Variations on Resonant Wavelength Shift . . . 179

Appendix C Resonant-Wavelength Shift in MRRs . . . 181

LIST OF TABLES

2.1 Different parameters used to generate FPV maps 23

3.1 Different parameters used in our simulations 38
3.2 Measurement results in MRR1-3 45
3.3 MRR test structure design parameters (see Fig. 3.10). 54
3.4 Characterized device performance (Avg.: Average, SD: Standard Deviation, λ_R: Resonant Wavelength, ER: Extinction Ratio). 57

4.1 Different parameters used to generate FPV maps 62
4.2 Models and datasets considered for evaluation 83
4.3 Parameters considered for analysis of photonic accelerators 83
4.4 Average EPB and kiloFPS/Watt values across accelerators 89
4.5 Models and Datasets used for Evaluations 109
4.6 Parameters Considered for Analysis of Photonic Accelerators 111
4.7 Inference time on *ROBIN-PO* and Intel i7 Desktop for the Four Models 119
4.8 Parameters used to generate FPV maps. 124
4.9 Architectures of the SPNNs considered. FC(x,y): Fully connected layer with x inputs and y outputs, SP: Softplus activation, LSM: LogSoftMax activation, PhS: phase shifters. 136
4.10 Network-2 accuracy with worst-case-tolerant MZIs designed using strip waveguides under width variations. 140

5.1 Characterized device performance (Avg.: Average, SD: Standard Deviation) 151

LIST OF FIGURES

1.1 The graph from Nokia illustrates the exponential growth of global network traffic, with projections indicating a dramatic increase in data demand in the coming years. This surge in data demand poses a significant challenge for traditional network infrastructure, which is struggling to keep pace with the ever-growing data demands. 2
1.2 Illustration of increase in computing power (represented by the total power per package), the energy required for off-chip communication (I/O) at an unsustainable pace. As data centers continue to expand to handle the explosive growth in data, the power consumption for data movement is becoming a major bottleneck. This not only leads to increased operational costs but can also cause signal degradation, data loss, and ultimately limit the performance of advanced computing systems. 3
1.3 This Figure illustrates the growing disparity between the exponential growth in model parameter size (red line) and the comparatively slower increase in AI hardware memory capacity (blue line) hindering progress in high-performance computing, autonomous vehicles, and other data-intensive applications. 4
1.4 The figure (a) illustrates an all-pass microring resonator (MRR) coupled to two waveguides, showcasing its transmission spectrum at a resonant wavelength. In this configuration, light enters the MRR from the input port and circulates within the ring. At specific resonant wavelengths, determined by the ring's parametrs, constructive interference occurs, allowing light to efficiently couple to the drop port. (b) The transmission spectrum reveals a distinct dip at the resonant wavelength, indicating maximum transmission to the drop port and minimal transmission to the through port. 7
1.5 A 2×2 MZI structure with two integrated phase shifters (θ and ϕ) and two beam splitters based on directional couplers (B_{DCs}). Where a fraction of the optical signal (defined as κ) at an input port(I) is transmitted to an output port(O), and the remaining signal is coupled to the other output port. 9
1.6 Overview of the dissertation capturing the methodology for the design and optimization of silicon photonic integrated circuits under fabrication-process variations. 15

2.1 (a) A 200 mm × 200 mm random distribution map $z(x, y)$ generated with σ = 4.2; (b) a Gaussian filter map $g(x, y)$ generated with l = 4.5 mm; (c) the correlated variation wafer map $m(x, y)$. 21
2.2 (a) A 200 mm × 200 mm correlated variation wafer map $m(x, y)$, which is simulated using a coarse simulation mesh; (b) and (c) are the variation maps for a 9 mm × 9 mm die located at the top right corner of the wafer before and after interpolation, respectively. 21
2.3 Virtual FPV wafer maps ((a), (b), and (c)) and interpolated die maps ((d), (e), and (f)) that are correlated and mimic radial-variation effects. The maps are generated using the parameters in Table 2.1 with a mean of zero. 24

3.1 An overview of an MRR add-drop filter showing waveguide width (w), SOI thickness (t), and slab thickness (h) in (a) a passive and (b) an active MRR with (c) a P-N junction. Cross section of (d) a strip and (e) a ridge waveguide. Here, Si and SiO_2 denote silicon and silicon dioxide, respectively. 31
3.2 (a) Resonant-wavelength shift ($\Delta\lambda_R$) in an active MRR calculated using (3.4) with $w = 400$ nm, $t = 220$ nm, $h = 90$ nm, and $R = 10$ μm under variations in the waveguide width, SOI thickness, slab thickness, and radius (x-axis). (b) Optical spectrum of the MRR simulated using transfer-matrix method [1] with $\lambda_{R0} = 1550$ nm where the resonant wavelength shifts because of width variations ($\nu_w = \pm$ 5 nm). 33
3.3 A cross section of the coupling region in an active MRR add-drop filter with different physical- and device-level design parameters. Here, w_i is the input/drop waveguide width and w_r denotes the ring waveguide width. Note that $h = 0$ for a passive MRR. . 35
3.4 Resonant-wavelength shift slopes and device performance in passive ((a)–(d)) and active ((e)–(h)) MRRs when the input/drop and ring waveguide widths increase from 300 to 1500 nm (x-axis). Here, (a) and (e) also show the total resonant-wavelength shift ($T_{\Delta\lambda_R}$). Results are for the fundamental TE mode with the parameters in Table 3.1 when $w_i = w_r$. 36
3.5 Bending and propagation loss in strip and ridge waveguide as the waveguide width increases. 37
3.6 Resonant-wavelength shift slopes and device performance in passive ((a)–(d)) and active ((e)–(h)) unconventional MRRs. Here, (a) and (e) also show the total resonant-wavelength shift ($T_{\Delta\lambda_R}$). Results are for the fundamental TE mode with the parameters in Table 3.1 when $w_i = 400$ nm (considered as an example) and $w_i \neq w_r$. The x-axis shows the ring waveguide width (w_r) changes from 300 to 1500 nm. 39
3.7 (a) An example of the cell layout (bottom, MRR1 and MRR2 only) of the three designed passive TE-polarized MRRs with their design specifications (top). Note that 500-nm-wide strip waveguide is used for routing for all the MRRs. (b) Measured through- and drop-port responses obtained by testing 30 identical copies of MRRs in Fig. 3.7(a). All the MRRs were designed to resonate at 1550 nm, specified as desired resonance in the figures. 44
3.8 Performance of a passive MRR designed using the parameters in Table 3.1 and with $g = 100$ nm where both the input and ring waveguide widths change from 350 to 1200 nm. The desired design points are selected and shown with magenta squares and the optimal MRR design region in (f) satisfies all the requirements in (a)–(e). Moreover, the yellow triangle in (f) shows the design point at which the total resonant-wavelength shift is minimum ($T_{\Delta\lambda_R} = 3.8$ nm). 49
3.9 Performance of an active MRR designed using the parameters in Table 3.1 and with $g = 200$ nm where both the input and ring waveguide widths change from 350 to 1200 nm. The desired design points are selected and shown with magenta squares and the optimal MRR design region in (f) satisfies all the requirements in (a)–(e). Moreover, the yellow triangle in (f) shows the design point at which the total resonant-wavelength shift is minimum ($T_{\Delta\lambda_R} = 2.8$ nm). 50
3.10 (a) Adiabatic MRR with its design parameters ($w' > w$). (b) Rate of changes in the effective index (n_{eff}) w.r.t. the variations in the waveguide width (w) and thickness (t) as the waveguide width increases from 300 nm to 1500 nm (x axis). 52

3.11 (a) Optical loss and (b) transmission efficiency in a tapered ring with different radii and input–output waveguide widths. Results are based on simulating the curved, tapered quarter ring (inset in (a)) with the radius of R and nonidentical input–output waveguide widths (w and w') using Lumerical FDTD. The legend shows the difference in waveguide width (i.e., 0 nm $\leq |w' - w| \leq$ 700 nm). 53
3.12 Scanning electron microscopic (SEM) images of fabricated (a) conventional and (b) adiabatic MRRs taken from part of the chip. (c) The unit cell of the fabricated MRRs. . 55
3.13 Through-port and drop-port response in (a) conventional and (b) adiabatic MRRs. Figures show the resonant wavelength versus corresponding group index (n_g) to find responses that belong to the same resonant mode. Black-dotted lines show the ideal (nominal) resonant wavelengths. 56
3.14 Comparison of various performance metrics represented with different colors between each pair of MRRs placed at different positions on the chip (x axis). Yellow triangles show the average within each box, to which a linear fit (shown as black line) is depicted in (a). For each box, the red plus signs represent the outliers that fall outside the typical range and the red lines show the median within each box. 58

4.1 A two-channel passive wavelength-selective MRR-based demulitplexer with two MRRs placed at a distance d and with a channel spacing c_s, designed based on Table 2.1. Here, we assume $\lambda_{R1} = 1550$ nm and $\lambda_{R2} = 1553$ nm (radius in MRR2 is slightly different). Therefore, $c_s = 3$ nm. 62
4.2 Virtual FPV wafer maps ((a), (c), and (e)) and interpolated die maps ((b), (d), and (f)) that are correlated and mimic radial-variation effects. The maps are generated using the parameters in Table 4.1 with a mean of zero. 63
4.3 An overview of the channel-spacing accuracy optimization for the MRR demultiplexer in Fig. 4.1. 65
4.4 Statistical analysis of channel-spacing variations (Δc_s) in the demultiplexer in Fig. 4.1 while considering normal and optimized MRRs and different distances (d) between the two MRRs. In the normal MRR design, $w_{i1,r1} = w_{i2,r2} = 400$ nm. The optimized MRR design is based on the procedure in Fig. 4.3 with $w_{i1,i2} = 450$ nm and $w_{r1,r2} \in$[570, 820] nm. The legends show the mean (μ) and standard deviation (σ) of the normal-distribution fit of each histogram. The nominal channel spacing in Fig. 4.1 is 3 nm. 67
4.5 Noncoherent Broadcast-and-weight (B&W) based photonic neuron. 72
4.6 An all-pass MRR with output spectral characteristics at the through port with extinction ratio (ER) and free spectral range (FSR) specified in the figure. 74
4.7 An all-pass MRR with output spectral characteristics at the through port with extinction ratio (ER) and free spectral range (FSR) specified in the figure. 76
4.8 Phase crosstalk ratio and tuning power consumption in a block of 10 fabricated MRRs with variable distance between adjacent pair of MRRs. 78
4.9 Inference accuracy of the four DNN models considered, across quantization (resolution) range from 1 bit to 16 bits (for both weights and activations). 85
4.10 Scatterplot of average FPS vs. average EPB vs. area of various *CrossLight* configurations. The configuration with highest FPS/EPB (and FPS) is highlighted. 86

4.11 Power consumption comparison among variants of *CrossLight* vs. photonic accelerators (DEAP-CNN, Holylight), and electronic accelerator platforms (P100, Xeon Platinum 9282, Threadripper 3970x, DaDianNao, EdgeTPU, Null Hop) 87
4.12 Comparison of EPB values of the photonic DNN accelerators) 88
4.13 (a) A recurrent noncoherent B&W MAC based design [2]; (b) An MRR bank consisting of MRRs with individual resonant wavelength (λ_i) coupled to the MRRs at cross-over coupling (κ) and the output spectrum, showing free spectral range (FSR). . . 94
4.14 The accuracy sensitivity study conducted by varying activation parameter precision (number of bits). Weights are kept as binary values in all cases. The study was performed across four different models and their datasets (described later in Section 4.4.5.1 97
4.15 Tuning power compensation in a block of 10 MRRs placed with and without considering thermal eigen- mode decomposition (TED) for different MRR radius. The orange line represents phase crosstalk ratio variation with distance between MRRs. 98
4.16 (a) Resonant-wavelength shift slopes with respect to changes in waveguide width, thickness, and radius, and corresponding cross-over coupling(κ), when the input waveguide (w_i) is set to 400 nm the marked point represents our selected MRR design; (b) The different MRR designs considered in this work. 100
4.17 An overview of the *ROBIN* architecture, showing the electronic control unit, the photonic vector dot product (VDP) unit array, and the photonic summation unit, along with a detailed view of the VDP unit internal structure. 104
4.18 Pipelined scheduling of operations during BNN execution on the *ROBIN* accelerator. . 107
4.19 The training accuracy vs epoch for the BNN models considered for (a) Sign MNIST, (b) CIFAR10,(c) STL10, and (d) SVHN datasets. (a) shows top-1 accuracy, while (b)-(d) show top-5 accuracy. 110
4.20 Inference accuracy versus level of tuning applied. At 80% tuning, the inference accuracy saturates, rendering further tuning unnecessary, and providing an opportunity to save tuning power. 113
4.21 Scatterplot of average FPS vs. average EPB vs. area of various *ROBIN* configurations. The configuration with highest FPS/Watt (energy optimized or EO) and the one with best FPS (performance optimized or PO) are specified. 115
4.22 Power consumption comparison among variants of *ROBIN* versus other optical accelerators (DEAP- CNN, Holylight, LightBulb), and electronic accelerator platforms (P100, SIGMA, EdgeTPU, DaDianNao, Null Hop, FINN, and FBNA). 115
4.23 EPB comparison between electrical BNN accelerators, optical accelerators, and the *ROBIN* variants. 116
4.24 Average FPS/Watt among different accelerator platforms, visualized. 117
4.25 FPS comparison between the *ROBIN* variants and the electronic BNN accelerators. . . 118
4.26 (a) Overview of singular value decomposition (SVD) of a weight matrix related to a fully connected layer (L_m) with N_1 as the number of input ports and N_2 as the number of output ports. (b) An optical-interference unit (OIU). (c) A 2×2 MZI structure with two integrated phase shifters (θ and ϕ) and two beam splitters based on directional couplers (DCs). 122

4.27 An MZI device structure with waveguide tapers mapped to FPV maps (top), based on [3], with a mesh size of 10 μm. The MZI can use strip waveguides or ridge waveguides, both with the SOI thickens of 220 nm and varying waveguide width (w) on each arm. The design of slab thickness (150 nm) in the ridge waveguide is discussed in Section IV. Note that variation-free directional couplers (DCs) are considered. 126
4.28 Rate of changes in waveguide effective index (see the strip and ridge waveguides in Fig. 4.27) under FPVs $\frac{\partial n_{\text{eff}}}{\partial X}$, where X shows the design parameter under FPVs, in (a) a strip and (b) a shallow-etched ridge waveguide, when the waveguide width (w) increases from 350 to 1200 nm. Results are for $t = 220$ nm and $h = 150$ nm (for the ridge waveguide in (b)). 130
4.29 Minimum taper length required to keep the optical transmission between two waveguides of different widths consistent (i.e., at 1 in the figure) and to avoid mode distortion. The inset zooms in the results for the taper length of 0–2 μm. 132
4.30 Different region sizes (R1, R3, and R6) and related MZIs in a single 8×8 OIU unit. R12 (not shown) can be obtained similarly. Each MZI block size is 30×340 μm^2. . . . 133
4.31 (a) RVD for different region sizes and under FPVs with different correlation lengths. (b) High R^2 values denote a strong linear correlation between RVD and accuracy. FPVs are based on the parameters in Table 4.8. We consider the linear correlation as an example to calculate R^2. 135
4.32 Accuracy of two SPNNs in Table 4.9 under correlated FPVs in width, SOI thickness, and slab thickness before and after the optimization. Here, ST and RG denote strip and shallow-etched ridge waveguide, respectively. The parameters listed inside (.) show the variations considered, with W, T, and H denoting waveguide width, SOI thickness, and slab thickness variations, respectively. Note that results for ST (W+T) and RG (W+T+H) are the same for R1. The second y-axis shows the average difference between No Optim. accuracy (i.e., using the conventional MZI) and the accuracy obtained using RG (W+T+H) for R1, R3, R6, and R12. 137

5.1 ProVAT workflow for designing FPV-resilient microring MRRs. The workflow encompasses: (a) generating realistic FPV maps across a wafer and individual dies, (b) analyzing the impact of FPVs on strip and ridge waveguides; (c) exploring adiabatic MRRs, considering coupling effects and variations on ring parameters; (d) calculating cross-over coupling coefficients for various waveguide dimensions; and (e) exploring and optimizing MRR designs to meet user-specified performance requirements. Top graphs in (e) show variations in free spectral range (FSR) and extinction ratio with ring radius, while bottom graphs demonstrate optimized values for specific design requirements. 143
5.2 ProVAT workflow showcasing: (a) the effect of width and thickness variations on n_{eff} and n_{g} for strip/ridge waveguides, (b) the impact of FPVs on MRR performance metrics (resonant wavelength shift ($\Delta\lambda_r$), FSR, transmission spectrum), and (c) selection of optimal adiabatic MRR design parameters (see figure inset) aimed to reduce impact of $\Delta\lambda_r$ under FPVs. 145
5.3 Through-port and drop-port response in (a) conventional and (b) adiabatic MRRs. Black-dotted lines show the ideal (nominal) resonant wavelengths. 150

A.1 (a) Difference between the effective indices of the symmetric (n_e) and antisymmetric (n_o) supermodes in a conventional and unconventional DC. (b) Rate of changes in the DC cross-over coupling (κ) w.r.t. the changes in Δn. Cross-over coupling in (c) a conventional and (d) an unconventional DC (x-axis shows w_2). In these simulations, $L = 5$ μm and $g = 100$ nm. Also, $w_1 = 400$ nm in the unconventional DC. 177

B.1 (a) MRR showing inner and outer radius R_i and R_o respectively. The average radius (R) without any radius variations, (b) variations in MRR radius (R') due to changes in MRR radius, and (c) variations in MRR radius (R') 180

Chapter 1

Introduction

The relentless pursuit of faster, more efficient and higher bandwidth communication systems has propelled the field of photonics to the forefront of technological innovation [4, 5]. Silicon photonics, leveraging the mature complementary metal-oxide-semiconductor (CMOS) fabrication-processes ubiquitous in the electronics industry, has emerged as a promising platform to address these demands [6]. This integration of photonics onto silicon substrates offers the potential to revolutionize various applications, including optical interconnects [7], data centers [5], telecommunications [8], and sensing [1]. Inherent compatibility with CMOS technology enables the co-integration of photonic and electronic components on a single chip, leading to miniaturized, low-cost, and high-performance devices [8]. This convergence of photonics and electronics holds the promise of overcoming the limitations imposed by traditional electrical interconnects, such as bandwidth bottlenecks and power consumption constraints [7].

Although electronic integrated circuits (ICs) have undeniably revolutionized modern technology, their inherent limitations are becoming increasingly apparent as demands for data transmission and processing continue to escalate. The scaling of transistors, a cornerstone of Moore's Law, is facing fundamental physical constraints [9]. Furthermore, electrical interconnects suffer signal degradation due to resistive losses, capacitive coupling, and electromagnetic interference, which is exacerbated with increasing data rates and transmission distances [10]. These challenges have spurred the exploration of alternative paradigms, with photonics emerging as a prime candidate to complement and extend the capabilities of electronics.

Fig. 1.1 captures trends in global network traffic that paints a clear picture of the exponential surge in data demand, with traffic projections skyrocketing in the coming years. This impending data deluge poses a significant challenge for traditional electronic network infrastructure. Silicon photonics, with its ability to transmit massive amounts of data at lightning speeds using light, offers a promising solution to meet this escalating demand. It has the potential to revolutionize network

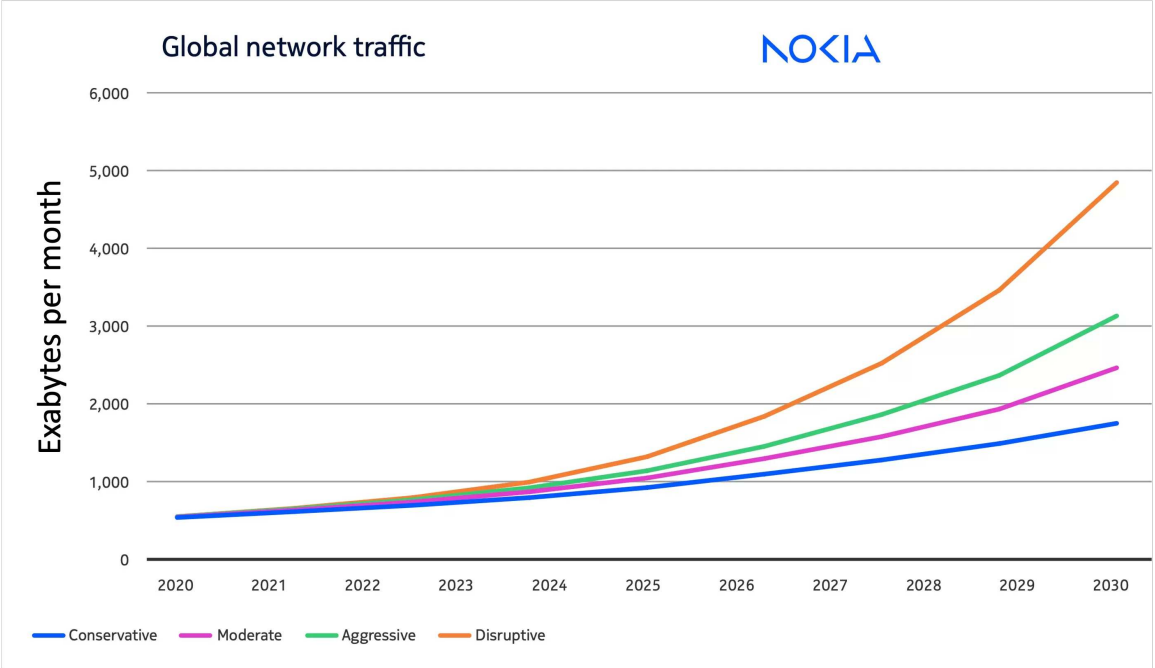


Figure 1.1: The graph from Nokia illustrates the exponential growth of global network traffic, with projections indicating a dramatic increase in data demand in the coming years. This surge in data demand poses a significant challenge for traditional network infrastructure, which is struggling to keep pace with the ever-growing data demands.

capacity and efficiency, enabling us to build the high-bandwidth, low-latency networks necessary to support the future of the internet and emerging technologies.

Fig. 1.2 illustrates the exponential growth of global network traffic, with projections indicating a dramatic increase in data demand in the coming years. This surge in data demand poses a significant challenge for traditional network infrastructure, which is struggling to keep pace with the ever-growing data demands. Silicon photonics, with its ability to transmit massive amounts of data at lightning speeds using light, offers a promising solution to meet this escalating demand. With the potential to revolutionize network capacity and efficiency, enabling us to build the high-bandwidth, low-latency networks necessary to support the future of the internet and emerging technologies.

Fig. 1.3 starkly illustrates the widening gap between the exponential growth in model parameter size and the comparatively slower increase in artificial intelligence (AI) hardware memory capacity. This disparity, known as the "memory wall," is a critical bottleneck hindering progress in

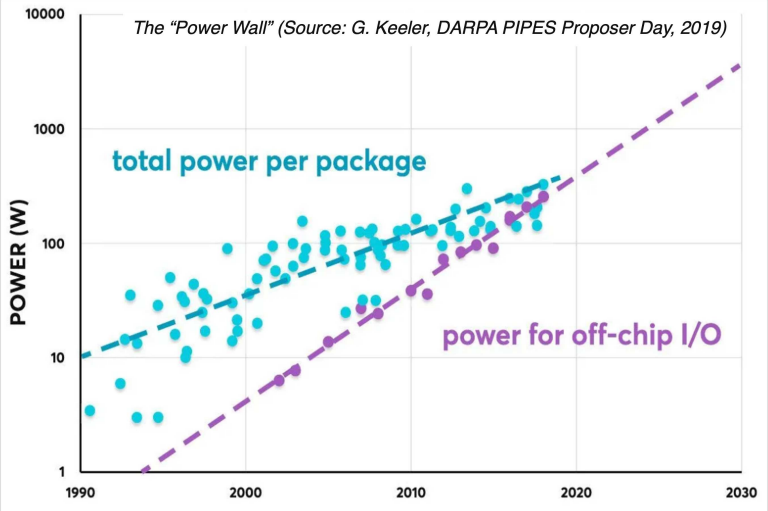


Figure 1.2: Illustration of increase in computing power (represented by the total power per package), the energy required for off-chip communication (I/O) at an unsustainable pace. As data centers continue to expand to handle the explosive growth in data, the power consumption for data movement is becoming a major bottleneck. This not only leads to increased operational costs but can also cause signal degradation, data loss, and ultimately limit the performance of advanced computing systems.

high-performance computing, autonomous vehicles, and other data-intensive applications. As the volume of data generated and processed continues to explode, the limitations of traditional electrical interconnects become increasingly apparent. Silicon photonics, with its ability to provide high-bandwidth, low-latency data transfer, offers a promising solution to overcome the memory wall. By enabling faster and more efficient communication between processors and memory, silicon photonics can unlock the full potential of these data-hungry applications and pave the way for the next generation of computing systems.

Silicon photonics, with its unique properties and compatibility with CMOS fabrication, presents several key advantages over electrical interconnects. Firstly, optical signals transmitted through silicon waveguides exhibit significantly lower losses compared to electrical signals in copper wires, especially over long distances. This translates to higher bandwidths and longer reach capabilities for silicon photonic interconnects [11]. Secondly, the absence of electrical currents in optical transmission eliminates resistive heating and reduces power consumption, making silicon photonics a more energy-efficient solution [8]. Additionally, optical signals are immune to electromagnetic interference, ensuring signal integrity in noisy environments [7]. The ability to densely integrate

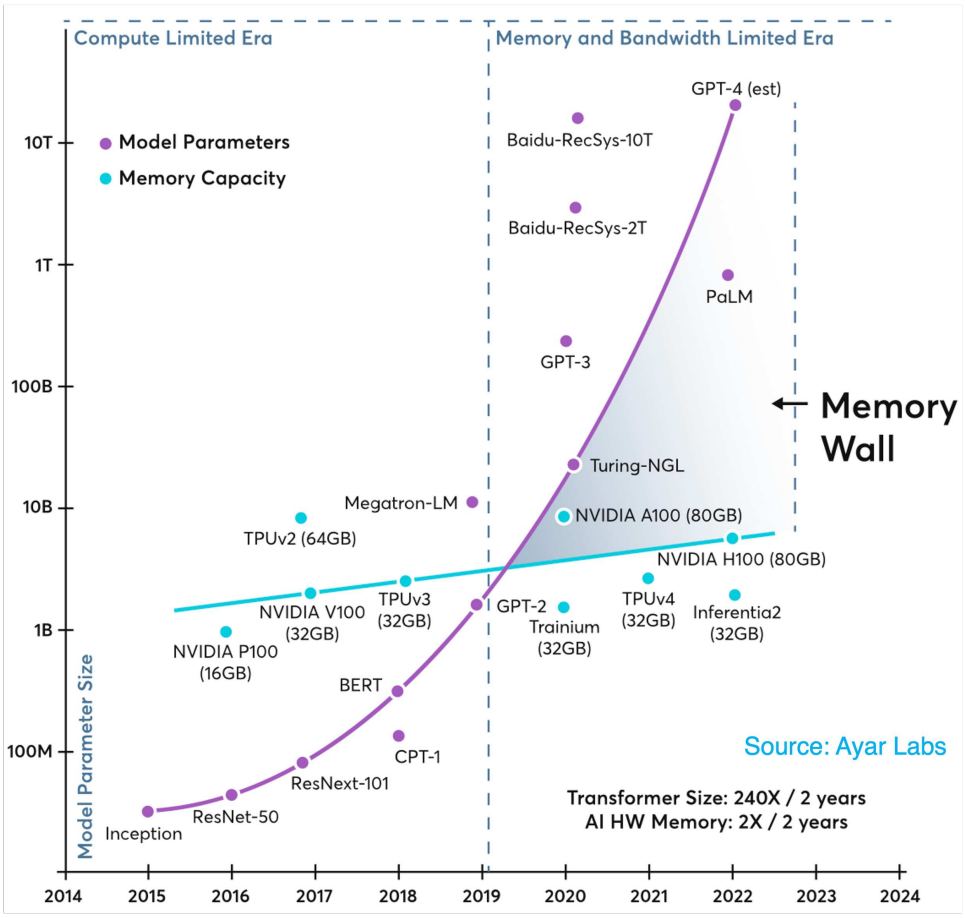


Figure 1.3: This Figure illustrates the growing disparity between the exponential growth in model parameter size (red line) and the comparatively slower increase in AI hardware memory capacity (blue line) hindering progress in high-performance computing, autonomous vehicles, and other data-intensive applications.

photonic components alongside electronics on a single chip further enhances functionality and miniaturization, paving the way for compact, high-performance systems [1].

To fully harness the potential of silicon photonics and overcome the challenges inherent in its implementation, a comprehensive understanding of various aspects is essential. First, we go a thorough exploration of fundamental photonic devices and their roles in photonic integrated circuits (PICs) . This includes an in-depth analysis of their design principles, operating characteristics, and performance metrics. Next, the impact of fabrication-process variations on the behavior of photonic devices and PICs will be investigated. This analysis will encompass the identification of key variation sources, their statistical modeling, and the development of techniques to mitigate their adverse effects. Finally, the concept of photonic accelerators, leveraging the advantages of silicon photonics to enhance computational tasks, will be introduced. This section will focus on their design considerations, potential applications, and the challenges associated with their integration.

1.1 Silicon Photonic Devices

Photonic devices are the fundamental building blocks of photonic integrated circuits (PICs), analogous to transistors in electronic ICs. These devices manipulate light to perform various functions, such as generating, amplifying, modulating, routing, and detecting optical signals. Understanding their design principles, operating characteristics, and interactions is crucial for developing high-performance and reliable PICs.

1.1.1 Optical Waveguides

The primary mechanism enabling photonics on a chip is total internal reflection, a phenomenon where light traveling within a high-refractive index material (such as silicon) is reflected back into the material when it encounters an interface with a lower refractive index medium (such as silicon dioxide). This principle is exploited in the design of waveguides, which are the fundamental structures for guiding and confining light on a silicon chip [8]. Waveguides can be compared to as optical "wires" that channel light signals from one device to another within a PIC.

The light propagating within a waveguide is not a uniform beam but rather consists of distinct spatial patterns known as optical modes. Each mode represents a specific solution to the electromagnetic wave equation that governs light propagation within the waveguide. The cross-sectional dimensions and refractive index contrast of a waveguide determine the number of supported modes and their respective mode profiles, which describe the distribution of light intensity across the waveguide. Single-mode waveguides, which support only one optical mode, are often preferred for their well-defined propagation characteristics and minimal dispersion [1]. The design of waveguides involves careful consideration of factors such as bending losses (These losses occur when a waveguide is curved, causing some light to radiate away from the waveguide), propagation losses (These losses result from absorption and scattering of light within the waveguide material), and coupling efficiency (This refers to the percentage of light transferred from one waveguide to another or from a waveguide to a device) to other devices.

Waveguides, with their ability to confine and direct light, serve as the backbone for constructing various photonic devices on a chip. These devices exploit the properties of light manipulation within waveguides to achieve specific functionalities, such as filtering, modulation, switching, and wavelength conversion. One such versatile and widely used device is the microring resonator, which we will explore in detail in the following section.

1.1.2 Photonic Microring Resonators (MRRs)

Microring resonators (MRRs) are compact, high-performance devices that have found widespread applications in silicon photonics due to their exceptional properties and versatility. Their operating principle is based on the resonance phenomenon, where light circulating within a ring-shaped waveguide interferes constructively or destructively with itself, depending on the wavelength of the light and the circumference of the ring. MRRs are essentially ring-shaped waveguides coupled to one or more bus waveguides, as illustrated in Fig. 1.4. Light propagating in the bus waveguide can couple into the ring and circulate multiple times before coupling back out to the bus waveg-

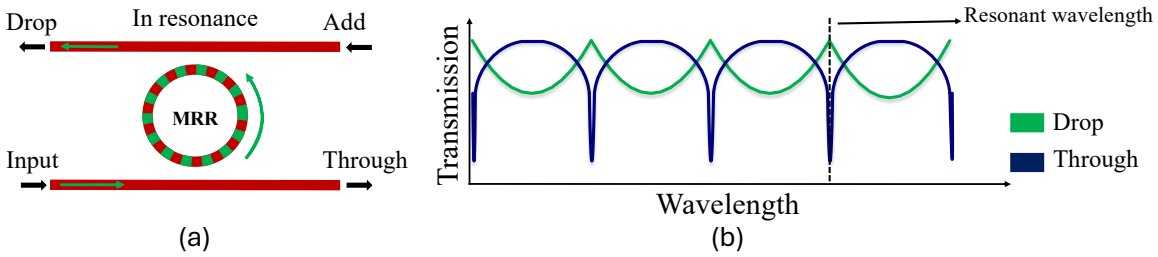


Figure 1.4: The figure (a) illustrates an all-pass microring resonator (MRR) coupled to two waveguides, showcasing its transmission spectrum at a resonant wavelength. In this configuration, light enters the MRR from the input port and circulates within the ring. At specific resonant wavelengths, determined by the ring's parametrs, constructive interference occurs, allowing light to efficiently couple to the drop port. (b) The transmission spectrum reveals a distinct dip at the resonant wavelength, indicating maximum transmission to the drop port and minimal transmission to the through port.

uide. The resonant nature of MRRs arises from the constructive and destructive interference of the circulating light waves.

The effective index (n_{eff}) represents the equivalent refractive index experienced by the light propagating within the waveguide, taking into account the waveguide geometry and material properties. The resonance condition for an MRR occurs when the optical path length of the ring, determined by the product of the effective index and the circumference of the ring ($n_{\text{eff}}L$), is an integer multiple of the wavelength of the light [12]:

$$\lambda_R = \frac{n_{\text{eff}} \cdot L}{m}, \tag{1.1}$$

Where, λ_R is wavelength at which resonance occurs, L is the round trip length and, m is an integer representing the resonant mode order. The design of MRRs involves careful consideration of several parameters that influence their performance. The ring radius (R) plays a crucial role in determining the free spectral range (FSR), which is the spacing between adjacent resonant wavelengths. Smaller radii result in larger FSRs, allowing for denser wavelength channels in applications like wavelength-division multiplexing (WDM) [1]. The coupling coefficient (κ) governs the amount of light transferred between the bus and ring waveguides. A higher coupling coefficient leads to broader resonance linewidths and lower quality factors (Q-factor), while a lower coupling coefficient results in narrower linewidths and higher Q-factor. The round-trip loss (α), encom-

passing losses due to scattering, absorption, and bending within the ring, significantly impacts the Q-factor. Minimizing these losses is essential for achieving high Q-factors and sharp resonances, which are desirable for various applications.

The performance of MRRs is typically evaluated using metrics such as the FSR, resonance linewidth ($\Delta\lambda$), and quality factor (Q) [12]. The FSR, calculated as

$$\text{FSR} = \frac{\lambda_R^2}{n_\text{g} L}, \tag{1.2}$$

where n_g is the group index of the ring waveguide, quantifies the spectral separation between adjacent resonances and is a measure of how fast light travels in the waveguide which is related to the effective index by

$$n_\text{g} = n_\text{eff} - \lambda_R \frac{dn_\text{eff}}{d\lambda_R} \tag{1.3}$$

These unique characteristics of MRRs make them versatile building blocks for various applications in silicon photonics. Their ability to selectively filter specific wavelengths makes them indispensable in WDM systems, where multiple optical signals are transmitted over a single fiber using different wavelengths. By dynamically adjusting the refractive index of the ring, MRRs can function as optical switches and modulators, enabling the routing and manipulation of light signals. In addition, the sensitivity of the resonant properties of MRRs to changes in their surroundings makes them excellent candidates for optical sensors capable of detecting various physical and chemical parameters with high precision. The inherent time delay experienced by circulating light within MRRs also opens up possibilities for their use as optical delay lines in applications such as optical buffering and signal processing. The broad range of applications and their compatibility with CMOS fabrication-processes firmly establish MRRs as essential components in the advancement of silicon photonics.

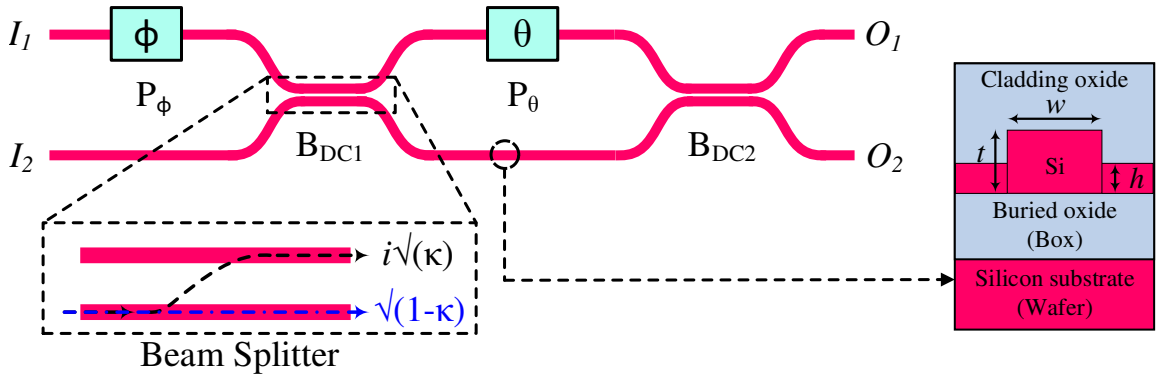


Figure 1.5: A 2×2 MZI structure with two integrated phase shifters (θ and ϕ) and two beam splitters based on directional couplers (B_{DCs}). Where a fraction of the optical signal (defined as κ) at an input port(I) is transmitted to an output port(O), and the remaining signal is coupled to the other output port.

1.1.3 Mach-Zehnder Interferometers (MZIs)

Mach-Zehnder interferometers (MZIs) are another essential building block in silicon photonics, renowned for their ability to manipulate the amplitude and phase of optical signals. As depicted in Fig. 1.5, an MZI consists of two 3dB couplers (which ideally split the input light into two equal paths) connected by two waveguide arms. The light traveling through the two arms experiences different phase shifts due to their distinct path lengths or refractive indices. When the light waves recombine at the second coupler, they interfere, resulting in either constructive or destructive interference depending on the relative phase difference between the two arms.

The output intensity of an MZI can be expressed as:

$$I_{\text{out}} = I_{\text{in}} \left[1 + \cos(\Delta\phi)\right] /2, \tag{1.4}$$

where I_{out} is the output intensity, I_{in} is the input intensity and, $\Delta\phi$ is the phase difference between the two arms.

The phase difference ($\Delta\phi$) between the MZI arms, which governs its behavior, is intricately linked to several design parameters. The arm length difference is a fundamental factor, as a longer arm introduces a greater phase shift due to the increased propagation distance. Additionally, variations in the refractive index of the waveguide material or its surrounding medium can significantly

alter the phase velocity of light within the arms, leading to changes in the overall phase difference. This refractive index can be influenced by various factors, including the thermo-optic effect, where temperature variations induce changes in the refractive index. This effect provides a convenient mechanism for thermally tuning the phase difference. Furthermore, certain materials exhibit the electro-optic effect, wherein an applied electric field modifies their refractive index. This enables rapid electrical control of the phase difference, making MZIs suitable for high-speed modulation and switching applications.

By judiciously manipulating these design parameters, MZIs can be configured to fulfill diverse functionalities. Introducing a π phase shift in one arm, for instance, transforms the MZI into an optical switch, directing the light output to either of the two output ports. Dynamic modulation of the phase difference enables the encoding of information onto the optical signal, making MZIs key components in optical modulators for high-speed communication systems. MZIs can also function as optical filters by adjusting the phase difference to achieve constructive or destructive interference for specific wavelengths. Moreover, their sensitivity to environmental changes, reflected in the phase difference, allows them to serve as optical sensors for various physical and chemical parameters. The versatility and control offered by MZIs, combined with their compatibility with silicon photonics, solidify their importance in advancing photonic integrated circuits.

The flexibility, controllability, and wide range of functionalities offered by MZIs, coupled with their compatibility with silicon photonics, have solidified their role as indispensable building blocks in the design and realization of advanced PICs. However, the rich landscape of silicon photonics encompasses a diverse array of other essential devices that warrant exploration. In addition to the fundamental devices discussed thus far, a plethora of other photonic components play critical roles in the operation and functionality of silicon photonic integrated circuits. These devices, contribute to the overall versatility and performance of PICs.

1.1.4 Other Optical Devices

Beyond MRRs and MZIs, a variety of other photonic devices contribute to the versatility and functionality of silicon photonic integrated circuits (PICs). These devices play essential roles in routing, splitting, combining, and coupling light signals within the PIC.

Y-Splitters: These devices split an incoming optical signal into two equal parts, typically with a 50/50 splitting ratio [13]. They are commonly used to distribute signals to multiple components or to create balanced interferometers.

Directional Couplers: These devices couple light between adjacent waveguides [12]. The coupling ratio can be controlled by the separation and interaction length between the waveguides, enabling the design of various splitting ratios and power transfer functionalities. Directional couplers are widely used in optical switches, modulators, and filters.

Grating Couplers: These devices couple light between optical fibers and on-chip waveguides [14]. They utilize a periodic grating structure to diffract light at specific angles, enabling efficient coupling from the fiber mode to the waveguide mode. Grating couplers are essential for interfacing PICs with external optical systems.

Photodetectors: These devices convert optical signals into electrical signals [15]. They are typically based on semiconductor materials that generate electron-hole pairs upon absorbing photons. Photodetectors are used for receiving and demodulating optical signals in communication systems and for sensing applications.

These miscellaneous devices, along with MRRs and MZIs, constitute a comprehensive toolkit for designing and implementing complex photonic integrated circuits. Their precise fabrication and integration on a single chip enable the realization of high-performance, compact, and low-cost photonic systems for a wide range of applications.

1.2 Photonic Accelerators

The rapid growth of data-intensive applications, such as artificial intelligence (AI), machine learning (ML), and high-performance computing (HPC), has intensified the demand for faster and

more energy-efficient computational platforms. Traditional electronic processors, while versatile, are increasingly encountering limitations in terms of power consumption and interconnect bandwidth [16]. Photonic accelerators, leveraging the inherent advantages of photonics, offer a promising avenue to address these challenges and potentially revolutionize computational capabilities.

Photonic accelerators are specialized hardware architectures designed to accelerate specific computational tasks using light as the information carrier. By exploiting the high speed, low latency, and parallel processing capabilities of photonics, these accelerators can outperform their electronic counterparts in certain domains. In particular, operations such as matrix-vector multiplication, convolution, and Fourier transforms, which are fundamental building blocks in many computational algorithms, can be efficiently implemented using photonic circuits [17].

The potential of photonic accelerators extends beyond these basic operations. They have shown promise in accelerating complex tasks like deep neural network (DNN) inference, where the massive parallelism of optics can be harnessed to perform numerous multiply-accumulate (MAC) operations simultaneously [18]. Additionally, photonic accelerators have been explored for solving optimization problems, such as the Ising model, which has applications in various fields including drug discovery and financial modeling [19]. The operational bandwidth of such photonic accelerators can approach the photodetection rate (hundreds of GHz), which is significantly higher than electronic systems today that operate at a few GHz [18].

The convergence of silicon photonics and machine learning has given rise to a new class of DNN accelerators that employ photonic integrated circuits (PICs) for low-latency and energy-efficient data transport and computation. PICs, fabricated using CMOS-compatible processes, integrate both photonic and electronic components on a single chip, enabling seamless data transfer and processing. The high bandwidth and low energy consumption of photonic interconnects alleviate the communication bottlenecks that plague electronic accelerators, while the parallel nature of optical computation allows for efficient execution of matrix operations that are central to DNN algorithms.

1.3 Fabrication-Process Variations

Fabrication-Process variations are an unavoidable consequence of manufacturing silicon photonic devices. These variations, essentially create deviations from the intended design, are inherent to the fabrication-processes and can significantly impact the performance, reliability, and yield of photonic circuits. Among these variations, geometric differences play a crucial role. These differences are evident in the altered physical dimensions of the photonic structures that form the device. For instance, the width and thickness of waveguides, critical components for guiding light, can vary from their intended values due to imperfections in the lithography, etching, and deposition processes involved in their creation [1]. These imperfections can arise from factors like mask misalignment during lithography, non-uniform etching rates, or variations in film deposition thickness. Even minute deviations on the scale of nanometers can have a substantial impact on the propagation of light within the waveguide, leading to alterations in the device's optical properties, such as its transmission spectrum and propagation loss [20].

Beyond geometric variations, material variations also play a significant role. These variations relate to fluctuations in the properties of the materials used to construct the photonic device. Primarily, this involves silicon, the core material for silicon photonics, but can also extend to other materials used in the device, such as dielectrics or metals in case of active photonic structures. These variations can include fluctuations in the refractive index of silicon, a critical parameter determining how light interacts with the material, or variations in the doping concentration, which affects the electrical properties of the silicon [21]. Such inconsistencies can stem from factors like non-uniformities in the deposition process, leading to variations in film thickness or composition, or differences in annealing temperatures, which can influence the material's crystal structure and thereby its optical properties. Both geometric and material variations can have a cascading effect on the performance of silicon photonic devices, leading to discrepancies between the designed and actual device behavior. Understanding and quantifying these variations is thus crucial for developing effective design to mitigate their unfavorable effects and ensure the reliable operation of silicon photonic circuits.

Within the confines of a single photonic chip, variations can arise even across its surface, referred to as within-die variations. These variations stem from spatial non-uniformities in the processing conditions during fabrication. For instance, temperature gradients within the deposition chamber can lead to inconsistencies in the film thickness and material properties across the chip [22]. Even slight temperature differences can cause the deposition rate to vary, resulting in regions of the chip with thicker or thinner films. These variations can directly affect the optical properties of the photonic structures within those regions, potentially leading to performance deviations. Similarly, variations in gas flow rates during etching can cause non-uniform etching depths, affecting the waveguide dimensions concurrently their optical behavior.

Zooming out to the wafer level, die-to-die variations come into play. These variations reflect inconsistencies between different chips or dies on the same wafer. While chips are fabricated simultaneously on a wafer, subtle differences in processing conditions can accumulate across the wafer surface, leading to variations in device performance from one die to another [23]. This could be due to factors like wafer edge effects, where the edges of the wafer experience slightly different conditions than the center, or variations in the uniformity of chemical baths used for processing. These die-to-die variations can result in a spread in device performance across the wafer, impacting the overall yield of functional chips.

Further broadening the scope, wafer-to-wafer variations encompass inconsistencies between different wafers within the same fabrication batch. These variations can arise from batch-to-batch inconsistencies in processing conditions and material properties. For example, slight differences in the initial quality of the silicon wafers, variations in the composition of the chemical precursors used for deposition, or even fluctuations in ambient temperature or humidity during processing can lead to variations in device performance between wafers [23]. These variations are of particular concern for large-scale manufacturing, where consistency across multiple wafers is essential to ensure high yields and reliable device performance. In essence, fabrication-process variations manifest at multiple levels, from within a single chip to across multiple wafers. Understanding and

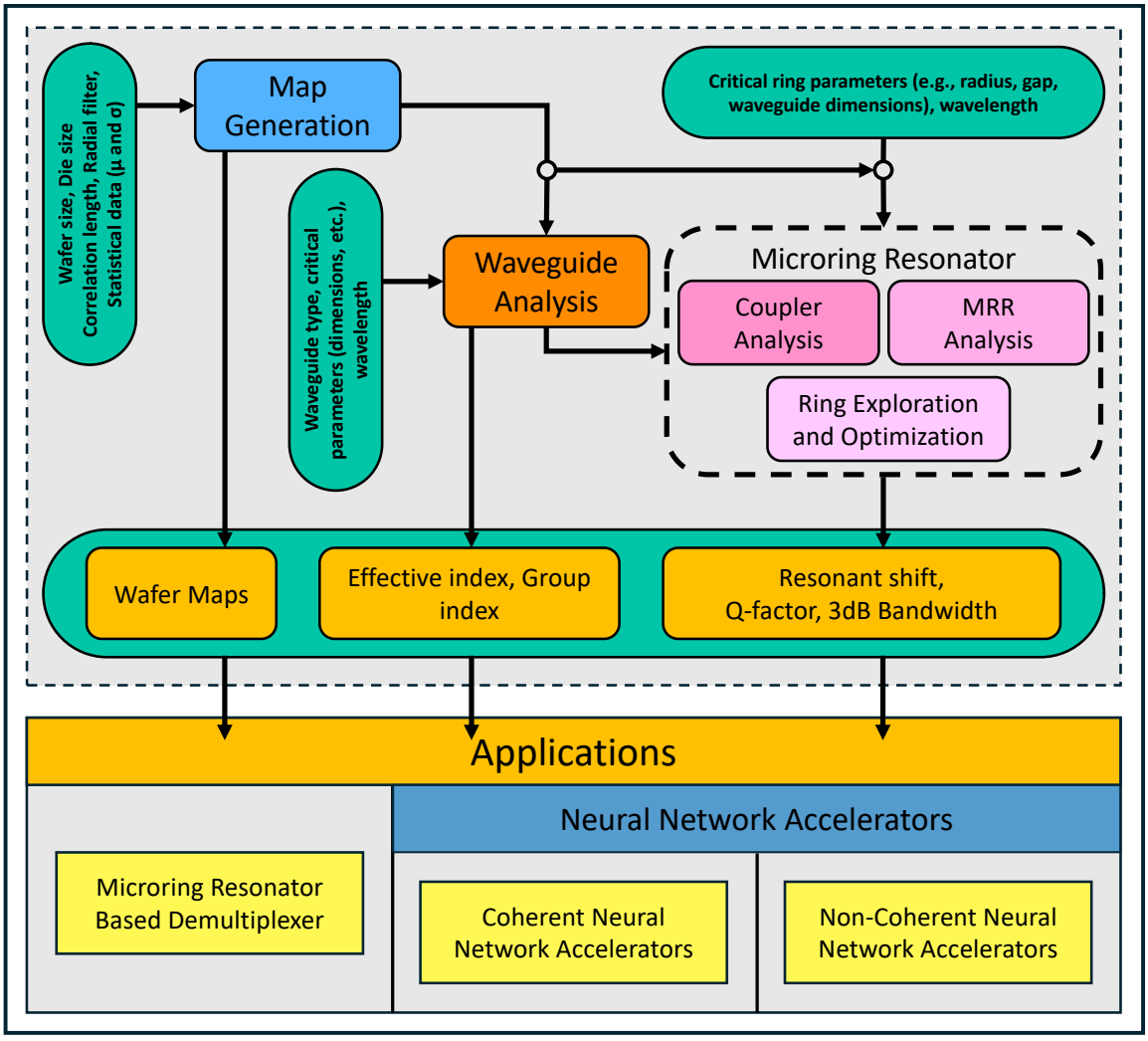


Figure 1.6: Overview of the dissertation capturing the methodology for the design and optimization of silicon photonic integrated circuits under fabrication-process variations.

mitigating these variations are crucial for ensuring the reliable and high-performance operation of silicon photonic devices and circuits.

1.4 Dissertation Overview

In summary, a comprehensive understanding of the impact of fabrication-process variations (FPVs) on photonic devices and their subsequent effects on the overall performance of photonic integrated circuits (PICs) is crucial for the advancement of silicon photonics technology. This

dissertation embarks on a detailed exploration of these intricate relationships, aiming to unravel the underlying mechanisms and develop strategies to mitigate the detrimental effects of FPVs.

As illustrated in Fig. 1.6, the dissertation follows a structured approach to tackle the challenges posed by FPVs. In Chapter 2, we delve into the intricacies of FPVs in the context of photonics, meticulously modeling realistic variations that mirror real-world scenarios. Armed with this understanding, we proceed in Chapter 3 to conduct an extensive design space exploration (DSE) of photonic devices, particularly focusing on microring resonators (MRRs). By analyzing the response of these devices to FPVs, we identify optimal design parameters that enhance their resilience and maintain performance under process fluctuations. These findings are further validated through fabrication and characterization of optimized MRR designs, as presented in Chapter 3.

The knowledge gained from understanding FPVs and optimizing device designs paves the way for innovative applications in Chapter 4. We explore the utilization of MRRs in diverse scenarios, such as wavelength demultiplexers and photonic neural network accelerators. Both coherent and non-coherent architectures are investigated, showcasing the versatility and inherent parallelism of silicon photonics, making it a compelling candidate for addressing the ever-growing demands of high-performance computing and information processing applications.

To empower researchers and designers in the field, Chapter 5 introduces ProVAT (Process Variation Analysis Tool), a comprehensive software framework developed to facilitate the design of FPV-tolerant MRRs. ProVAT enables users to navigate through a vast parameter space, considering design characteristics, variation parameters, and performance metrics, ultimately leading to the creation of tailored MRRs optimized for specific applications and fabrication constraints.

In conclusion, this dissertation contributes to the advancement of silicon photonics by not only shedding light on the impact of FPVs but also by providing practical tools and design methodologies for mitigating their adverse effects. By doing so, it lays a solid foundation for the realization of robust, high-performance photonic integrated circuits that can meet the stringent requirements of emerging technologies.

Chapter 2

Fabrication-Process Variation Analysis in Silicon Photonics[1]

2.1 Overview

Silicon photonics, leveraging the mature complementary metal-oxide-semiconductor (CMOS) fabrication infrastructure, has emerged as a promising platform for realizing integrated photonic circuits with diverse applications ranging from optical communications and data centers to sensing and quantum computing [24]. The ability to integrate various photonic components, such as lasers, modulators, detectors, and waveguides, onto a single chip holds the potential for miniaturization, cost reduction, and enhanced functionality. However, the successful realization of high-performance, reliable, and cost-effective silicon photonic circuits hinges on addressing a critical challenge: fabrication-process variations (FPVs).

All photonic devices are susceptible to FPVs. These variations originate in optical-lithography process imperfections, contributing to deviations in waveguide thickness and linewidth, waveguide edge roughness and sidewall slope, dopant concentration, and other critical parameters [25]. In photonic integrated circuits (PICs) employing microring resonators (MRRs), FPVs impose resonant-wavelength deviations, leading to severe PIC performance degradation, or in the worst-case, a total circuit failure [26]. For example, prior work shows an approximate 2 nm shift in the resonant wavelength of an MRR with only a single nanometer change in its waveguide thickness [27]. Such a small resonant-wavelength shift is of critical concern in dense wavelength division multiplexing (DWDM) systems with a large number of MRRs, each tuned to a specific optical channel (i.e., wavelength), with a channel spacing as small as 0.8 nm [28].

[1] A. Mirza, F. Sunny, P. Walsh, K. Hassan, S. Pasricha, and M. Nikdast, "Silicon photonic microring resonators: A comprehensive design-space exploration and optimization under fabrication-process variations," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), vol. 41, no. 10, pp. 3359–3372, 2022.

FPVs also pose a significant challenge for silicon photonic neural networks (SPNNs), an emerging application area for silicon photonics. These variations lead to undesired changes in the critical dimensions of SPNNs' building blocks, such as the waveguide width and thickness in Mach-Zehnder interferometers (MZIs) used in coherent SPNNs. Such inaccuracies can disrupt matrix-vector multiplications, a fundamental operation in SPNNs [29]. Research has shown that even in mature fabrication-processes, random uncertainties due to FPVs and thermal crosstalk can result in a catastrophic 70% accuracy loss [29]. Additionally, El-Henawy et al. [30] virtually simulated the effect of line edge roughness (LER) on the performance of a Y-branch splitter, a fundamental component of photonic integrated circuits (PICs). The authors found that LER caused an imbalance in the transmission between the two output ports of the Y-branch, with the imbalance increasing with both the amplitude and correlation length of the LER. The maximum imbalance observed was 15% for an LER amplitude of 15 nm and correlation length of 60 nm. In addition, LER was found to increase the excess loss of the Y-branch compared to the nominal (smooth) case, with losses reaching as high as 16.8% for an LER amplitude of 15 nm and correlation length of 60 nm.

In the context of silicon photonics, where device dimensions are on the order of hundreds of nanometers, even minute variations can have significant consequences. For instance, variations in waveguide dimensions can alter the propagation of light, leading to increased losses or shifts in operating wavelengths, while fluctuations in material properties can affect the efficiency of light generation, modulation, or detection. As such, a comprehensive understanding and effective management of fabrication-process variations are paramount for the continued advancement and widespread adoption of silicon photonics technology.

In the subsequent sections of this chapter, we delve into the intricate relationship between fabrication-process variations and photonic circuit design. We embark on a comprehensive review of the relevant literature and then explore the realm of modeling these variations, with a particular focus on creating realistic FPV maps. These maps serve as invaluable tools for understanding the behavior of devices during fabrication and for predicting the performance of entire systems under

the influence of FPV, ultimately paving the way for more resilient and high-performance silicon photonic technologies.

2.2 Related Work

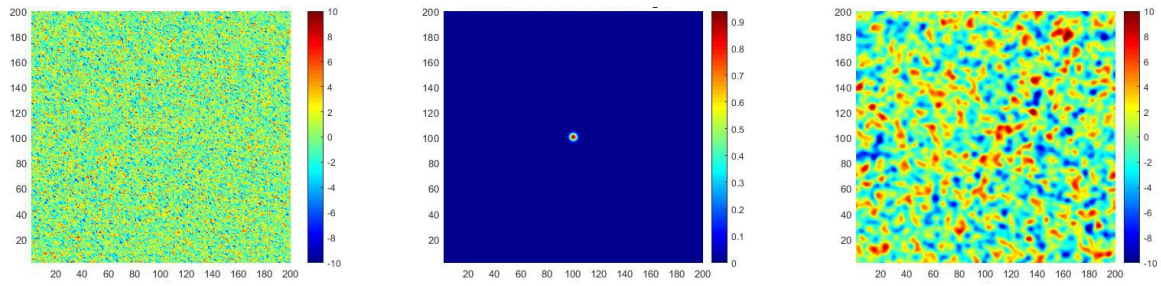
The analysis and mitigation of fabrication-process variations (FPVs) in silicon photonics have been a subject of ongoing research, with several notable contributions advancing our understanding and capabilities in this domain. Xing et al. [31] presents a hierarchical model to analyze spatial process variations in integrated photonics across different levels, including intra-wafer and intra-die variations. The authors conducted automated wafer measurements on a set of Mach-Zehnder interferometers (MZIs) to extract geometric parameters like linewidth and thickness, as well as behavioral parameters like effective and group index. By averaging parameter wafer maps for all devices with unique locations on the die, the systematic wafer variation is obtained. The offset between individual device wafer maps and the averaged wafer map represents the systematic intra-die variation, while the remaining residue indicates random die-to-die variation. Such top to bottom approach does help us better understand the trends in variations however, to design for variability we would need a better bottom to top approach where we need to design devices and circuits while taking process variations into considerations for robustness.

Reda et al. [32] introduce a statistical framework to model the impact of process variations on semiconductor circuits using process-sensitive test structures. The authors employ a multivariate normal (MVN) distribution to represent the parametric measurements from different wafers, assuming that the data set follows this distribution with potential outliers. The study analyzes the systematic and random sources of process variations. The systematic component is represented by the mean values of the measurements at each location on the wafers, revealing spatial trends and dependencies. The random component is captured by the residuals, which are the deviations of individual measurements from the mean. Variograms are employed to analyze the spatial structure of the residuals, providing insights into the correlation and dependency between measurements at different locations. The estimated values are compared to the actual measurements to validate the

effectiveness of the approach and they conclude that their statistical framework accurately models process variations and enables efficient test cost reduction. However, the impact of such variations have not been directly studied on the performance of photonic devices hence, there is a gap to further validate the credibility of these process variations.

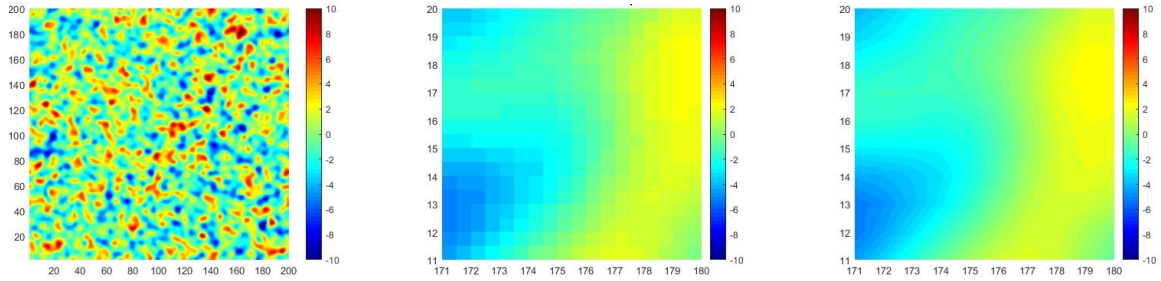
Boning et al. [23] introduces a comprehensive framework for variation-aware design in silicon photonics, addressing systematic, random, and particle defect variations. The authors emphasize the need for compact models that capture the relationship between process variations and device performance, enabling designers to predict and optimize designs for high yield and performance. This information can be used to decompose variations into die-to-die and within-die components, providing valuable insights for process control and optimization. The only issues with this work would be that some of the methods presented in this work, such as the virtual fabrication approach for analyzing random variations, can be computationally intensive, potentially limiting their applicability to large-scale circuits.

Lu et al. [33] proposes a method to characterize and predict the impact of manufacturing variations on the performance of silicon photonic integrated circuits (PICs). Authors developed an enhanced Monte Carlo (MC) simulation methodology that accounts for layout-dependent correlated variations, which are fluctuations in waveguide dimensions and material properties that occur during fabrication and can significantly affect circuit performance. To incorporate these variations into circuit simulations, the authors developed models that map the spatially correlated physical variations onto each circuit component based on its layout coordinates. This allows the simulation to account for the correlated variations between components and accurately predict the performance of the entire circuit under the influence of manufacturing variations. The authors applied their enhanced MC simulation methodology to several ring resonator filter circuits and successfully predicted common-mode and differential-mode variations in circuit performance. This work highlights the importance of considering layout-dependent correlated variations in the design and simulation of PICs to ensure their reliable operation and performance.



(a) (b) (c)

Figure 2.1: (a) A 200 mm $\times$ 200 mm random distribution map $z(x, y)$ generated with σ = 4.2; (b) a Gaussian filter map $g(x, y)$ generated with l = 4.5 mm; (c) the correlated variation wafer map $m(x, y)$.



(a) (b) (c)

Figure 2.2: (a) A 200 mm $\times$ 200 mm correlated variation wafer map $m(x, y)$, which is simulated using a coarse simulation mesh; (b) and (c) are the variation maps for a 9 mm $\times$ 9 mm die located at the top right corner of the wafer before and after interpolation, respectively.

Such works motivate us further to understand and design out photonic devices under fabrication-process variations by considering fabrication models that align closely with realistic expectations. However there is still a small gap to fill in that takes into consideration realistic maps where the variations in the center is slightly uniform and lower in magnitude than the variations experienced at the edges. The following section looks into modeling realistic process variations we build on the existing work by [33] to help us realize our proposed variation maps which can be used to understand the design space of photonic devices and help us further optimize them for robustness. In the following section we observe how these process variation maps are modelled.

2.3 Virtual Wafer-Level FPV Maps

Initial design of the virtual wafer model has been derived from [33]. The within-wafer manufacturing variations are modeled using a correlated surface roughness function [34]. Firstly, an

uncorrelated random distribution map, $z(x, y)$ is generated, based on discrete mesh of points in x-y plane, and the value at each discrete point is a random number following a normal distribution with a mean of 0 and a standard deviation of σ. Then, the random height function $z(x, y)$ is convolved with a Gaussian filter, $g(x, y)$, which is given by:

$$g(x, y) = \frac{1}{\sqrt{\pi}\frac{l}{2}} e^{-(\frac{x^2}{l^2/2} + \frac{y^2}{l^2/2})} \tag{2.1}$$

in which l, is correlation length for variation. The convoluted wafer with correlated variations is given by:

$$m(x, y) = \mathcal{F}^{-1}\left[\mathcal{F}[g(x, y)] \cdot \mathcal{F}[z(x, y)]\right] \tag{2.2}$$

where, $\mathcal{F}$ and $\mathcal{F}^{-1}$ are denotations for the fast Fourier transform and the inverse fast Fourier transform, respectively. Figs. 2.1(a)–2.1(c) show a random distribution map $z(x, y)$, a Gaussian filter map $g(x, y)$, and a correlated variation wafer map $m(x, y)$, respectively.

To include wafer-to-wafer variations in the wafer model, the mean of the correlated variation can be adjusted, which is given by:

$$m(x, y) = \mathcal{F}^{-1}\left[\mathcal{F}[g(x, y)] \cdot \mathcal{F}[z(x, y)]\right] + c \tag{2.3}$$

where c is a random number following a normally distribution with a mean of 0 and standard deviation σ^*. In total, there are three input parameters to the virtual wafer model, which are σ, l, and σ^* [33].

To increase computation efficiency, virtual wafers are simulated using a coarse simulation mesh of 500 μm $\times$ 500 μm while generating virtual wafer maps [33]. Furthermore, the simulated width and thickness variations on the selected die are interpolated to have a higher resolution mesh of 9 μm $\times$ 9 μm or die size appropriate to the user. The interpolated high resolution variation maps are used to map variations in photonic circuits. As an example, Fig. 2.2 illustrates the variations maps before and after interpolation.

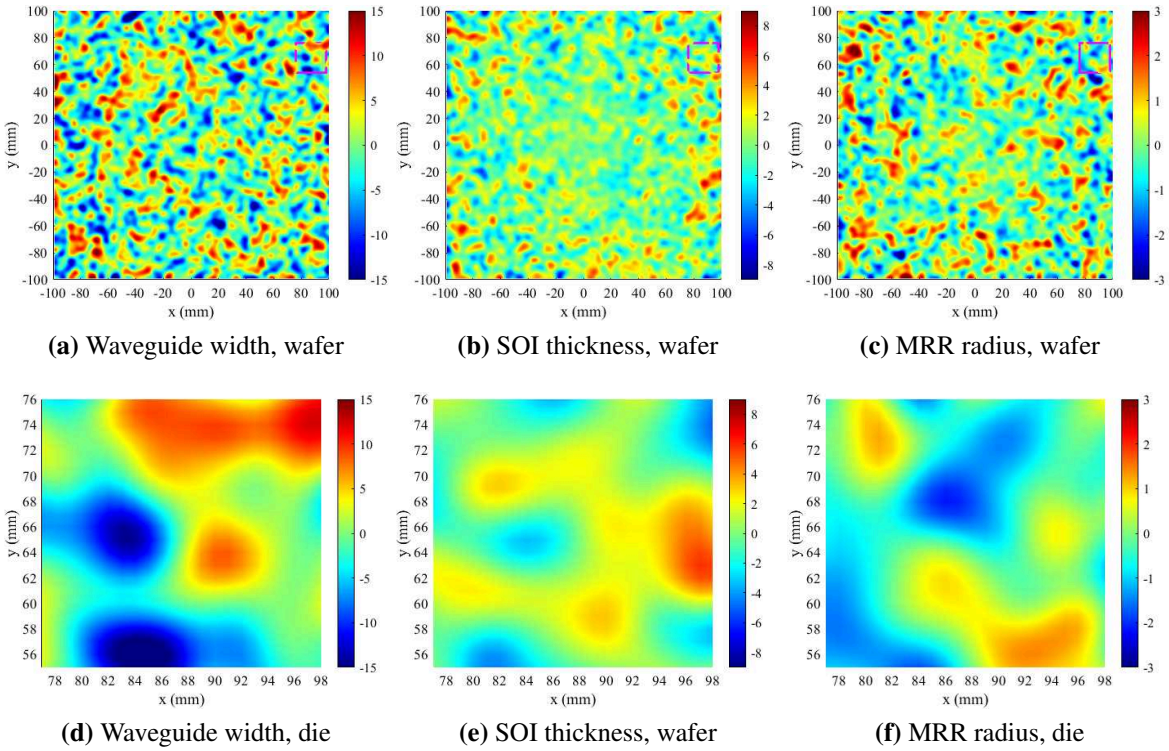
Table 2.1: Different parameters used to generate FPV maps

Design Parameter	Correlation Length	Standard Deviation
Waveguide width	$l_w = 4.5$ mm	Center: $\sigma_w = 4.2$ nm Edges: $\sigma_w = 5.5$ nm
SOI thickness	$l_t = 4.5$ mm	Center: $\sigma_t = 0.7$ nm Edges: $\sigma_t = 2.2$ nm
MRR radius	$l_R = 4.5$ mm	Center: $\sigma_R = 0.5$ nm Edges: $\sigma_R = 1$ nm

To capture radial-variation effects, we first characterized the waveguide width and SOI thickness variations (i.e., σ_w and σ_t) at the center of several 200 mm wafers, and then repeated the same at multiple points while moving towards the edges of the wafers. The standard deviations were then averaged over the points within the same distance to the wafer center. For instance, Table 2.1 shows the standard deviations averaged at the center and edges of several 200 mm wafers that we characterized in collaboration with CEA-Leti (σ_w and σ_t only). Accordingly, we enhance our virtual wafer-map models with a variable standard deviation that increases almost linearly as moving from the center towards the wafer edges.

Leveraging the aforementioned method, we develop waveguide width, SOI thickness, and MRR radius virtual FPV maps with means of zero ($\mu_w = \mu_t = \mu_R = 0$) and correlation lengths (l_w, l_t, and l_R) and standard deviations (σ_w, σ_t, and σ_R) listed in Table 2.1. In this table, $\sigma_{w,t}$ are analyzed through experimentally characterizing several 200 mm wafers at CEA-Leti, $l_{w,t}$ are taken from [33], and σ_R and l_R are considered as an example. Moreover, the table only shows the $\sigma_{t,w,R}$ at the wafer center and edges. Figs. 2.3(a), 2.3(b), and 2.3(c) show the resulting waveguide width, SOI thickness, and radius correlated wafer maps, respectively. Also, from each wafer map, we select a die with a size of 22×22 mm^2, which is then interpolated as shown in Figs. 2.3(d), 2.3(e), and 2.3(f).

In this section, we have detailed a comprehensive methodology for generating virtual wafer maps that incorporate both within-wafer and wafer-to-wafer variations. The model leverages correlated surface roughness functions to accurately capture the spatial correlations inherent in fabrication-processes. Additionally, the model accounts for radial variation effects by incorporat-



(a) Waveguide width, wafer **(b)** SOI thickness, wafer **(c)** MRR radius, wafer

(d) Waveguide width, die **(e)** SOI thickness, wafer **(f)** MRR radius, die

Figure 2.3: Virtual FPV wafer maps ((a), (b), and (c)) and interpolated die maps ((d), (e), and (f)) that are correlated and mimic radial-variation effects. The maps are generated using the parameters in Table 2.1 with a mean of zero.

ing variable standard deviations that increase towards the wafer edges, as observed in experimental characterization. By using a higher resolution model, this model can be applied to photonic circuits with greater precision. This virtual wafer model serves as a powerful tool for understanding the impact of FPVs on device performance, as demonstrated by the generated waveguide width, SOI thickness, and MRR radius variation maps. In the following chapters, we will utilize this model to analyze specific photonic devices and circuits under the influence of FPVs, providing valuable insights for robust design and optimization strategies.

2.4 Conclusion

In conclusion, this chapter has provided a comprehensive overview of fabrication-process variations (FPVs) in silicon photonics, highlighting their significant impact on device performance, yield, and reliability. We have explored the diverse types of these geometric variations. We have

delved into the development of a virtual wafer model, incorporating both within-wafer and wafer-to-wafer variations, to simulate realistic FPVs scenario. This model, combined with the analysis of existing literature, serves as a powerful tool for predicting device behavior under the influence of FPVs. By understanding the intricate relationship between fabrication-processes and device performance, we can pave the way for more robust design methodologies and improved fabrication techniques.

Chapter 3

Microring Resonators under Fabrication-Process Variations[1]

3.1 Introduction

Silicon photonics (SiPh) offers new and unique solutions where today's conventional technologies are approaching their limits in terms of speed, capacity, and accuracy. Silicon photonic integrated circuits (PICs) are emerging to boost the communication performance in high-performance computing systems [35], data centers [36], and are rapidly evolving into other applications, such as neural networks [37]. Among various SiPh devices designed for PICs, microring resonators (MRRs, an example of which is shown in Fig. 3.1) are widely considered as the primary building block in such circuits in many applications because of their compact footprint (*e.g.*, radius of 3 μm [38]) and capability to perform a variety of functions such as ability to perform a variety of functions such as optical filtering [39], modulation [40], and spatial switching [41,42]. The central optical frequency (i.e., resonant wavelength) in MRRs is an essential parameter that is determined by several key factors in the MRR design space, including waveguide width, SOI thickness, and MRR radius [12]. For dense wavelength-division multiplexing (DWDM) applications, where a large number of optical channels (i.e., wavelengths) are narrowly spaced to boost the bandwidth performance in PICs [43], it is critical to maintain accurate channel spacing and match the central resonant wavelengths of different MRRs to achieve a reliable communication [44]. Nevertheless, the resonant wavelength in an MRR is highly sensitive to the variations in the critical dimensions of the MRR due to inevitable fabrication-process variations (FPVs).

[1] A. Mirza, F. Sunny, P. Walsh, K. Hassan, S. Pasricha, and M. Nikdast, "Silicon photonic microring resonators: A comprehensive design-space exploration and optimization under fabrication-process variations," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), vol. 41, no. 10, pp. 3359–3372, 2022.

An MRR is on resonance when the input optical wavelength on its input matches with the resonant wavelength of the MRR (see Figs. 3.1(a) and 3.1(b)). This resonant wavelength is jointly determined by several key parameters in the MRR design space, including the radius and the waveguide physical dimensions (*i.e.*, width and thickness), as shown in Figs. 3.1(a) and 3.1(c). Reliable operation in PICs that integrate thousands of MRRs depends on precise matching of central wavelengths of MRRs in such circuits [45]. However, MRRs are considerably sensitive to fabrication-process variations (FPVs) that deviate the resonant wavelength of MRRs ($\Delta\lambda_{MRR}$ in Fig. 3.1(b)), resulting in severe performance and reliability degradation in PICs [46]. FPVs stem from optical lithography imperfections in which silicon-on-insulator (SOI) thickness and waveguide width variations are the major concerns [20,45]. Consequently, PICs require efficient methods to compensate for the impact of such inevitable FPVs.

FPVs originate in optical-lithography process imperfections, contributing to different variations in the waveguide thickness and linewidth, waveguide edge roughness and sidewall slope, dopant, etc. [25]. In PICs employing MRRs, FPVs impose resonant-wavelength deviations, leading to severe PIC performance degradation, or in the worst-case, a total circuit failure [26]. For example, prior work shows $\approx$2 nm shift in the resonant wavelength of an MRR with only a single nanometer change in its waveguide thickness [27]. Such a small resonant-wavelength shift is of critical concern in DWDM systems with a large number of MRRs, each tuned to a specific optical channel (i.e., wavelength), with a channel spacing as small as 0.8 nm [28]. Several methods have been proposed to compensate for the impact of FPVs in PICs at run-time by leveraging the thermo-optic [47] and electro-optic effects [48] of silicon. Nevertheless, such methods lead to considerable increase in power dissipation in PICs [46]. For example, internal integrated heaters can be used in MRRs to tune the resonant wavelength by 40 mW/FSR [49], where FSR is the free-spectral range. This rather small power consumption quickly adds up in PICs that integrate a large number of MRRs, where each MRR's resonant wavelength may need to be adjusted by several nanometers (e.g., 9 nm in [20]). This calls for efficient design solutions to minimize the effect of FPVs at design-time in the current fabless silicon photonic ecosystem [27], thus improving PIC

tolerance to FPVs and reducing required tuning power consumption after the fabrication. Note that post-fabrication trimming of MRRs is also possible [50], but it will significantly increase the fabrication cost. In addition to active compensation methods, MRRs can be designed to be more robust under variations. Prior work in [3, 51–53] has shown that an adiabatic design of MRRs can help improve device robustness under variations. Such a design helps reduce phase errors and scattering losses, enabling higher transmission efficiency and overall improved performance, including better tolerance to FPVs.

In this chapter we propose the first comprehensive design-space exploration for silicon photonic MRRs under different FPVs. In particular, 1) we analyze and model the impact of different FPVs in waveguide width, SOI thickness, slab thickness, and MRR radius (see Fig. 3.1) on the resonant wavelength of passive and active MRRs, identifying the impact of each variation on the resonant wavelength in such MRRs; 2) we present computationally efficient analytical models that capture the impact of such variations at the physical layer on the MRRs' device-layer performance, including coupling efficiency, Quality factor (Q-factor), and 3dB bandwidth; 3) leveraging our detailed design-space exploration, we develop MRR design optimization to find optimal design parameters in MRRs under different FPVs with a goal of enhancing not only the device tolerance to FPVs but also improving other important factors such as Q-factor and 3dB bandwidth; and, finally, we propose corresponding adiabatic MRRs based on the optimization and fabricate these MRRs to show that they are tolerant towards fabrication-process when compared to conventional MRRs.

3.2 Related Work

FPVs in silicon photonics have been studied at different scales, including exploring within-die, die-to-die, wafer-to-wafer, and lot-to-lot variations [31]. In [54], R. G. Beausoleil *et al.* studied the impact of FPVs in identically designed MRRs and reported within-die and within-wafer variations with a variance of 0.5 and 2 nm, respectively. A. V. Krishnamoorthy *et al.* explored the manufacturing tolerance of over 500 four-channel MRRs fabricated in a commercial 130 nm CMOS foundry with 193 nm lithography [55]. Results from characterizing multiple reticles, wafers, and

fabrication lots show that the absolute resonant wavelengths of individual MRRs cannot be controlled across wafers or even across reticles or fields within a wafer. In [56], X. Chen *et al.* studied FPVs in MRRs and racetrack resonators, all identically designed but fabricated in two different establishments (Leti and IMEC). They reported resonant-wavelength variations with a variance of 1.3 nm^2/cm in both the MRRs and racetrack resonators.

Through various examples of wavelength filters, W. Bogaerts *et al.* in [27] showed that variability modeling can guide the design process and make circuits more robust against different FPVs. Their work analyzed the performance of a Mach–Zehnder lattice filter under waveguide width and SOI thickness variations to improve the fabrication yield by optimizing circuit layout. In [33], Z. Lu *et al.* proposed a method to characterize FPVs and predict circuit performance under the impact of correlated FPVs. Using this method, they measured the spectral responses of 2074 identical racetrack resonators on a 200 mm wafer fabricated using a 248 nm deep UV (DUV) lithography photonics process. Moreover, for such a fabrication-process, they reported standard deviations for waveguide width and SOI thickness of 3.9 and 1.3 nm, respectively. Y. Wang *et al.* in [57] proposed a hierarchical approach that takes into account domain specific knowledge, spatial frequency analysis, and low-rank tensor factorization methods to decompose FPVs into wafer-level, intra-die, and inter-die components.

Similar to this chapter, the work in [52, 53, 58] aims at modeling and improving MRRs under FPVs. Y. London *et al.* in [58] proposed a behavioral model for directional couplers with non-identical waveguide widths, and not for the full design space of MRRs, which is a more complex problem to address. Y. Luo *et al.* in [52] developed an MRR modulator design based on using multi-mode waveguides that can improve accuracy and repeatability of the MRR resonant wavelength. In their work, the MRR is based on an unequal design of ring and bus waveguide widths where the ring waveguide width is increased to reduce phase errors associated with side-wall etch. Z. Su *et al.* in [53] proposed the use of adiabatic rings to design MRRs with high Q-factor and improved tolerance to FPVs, but the designed MRRs are too small (radii of 2–3 μm) and suffer from high optical power losses. Unlike the contributions in this chapter, [52, 53, 58] lacks compre-

hensive modeling and optimization of MRRs under FPVs, thereby focusing on the experimental design and improvement of the individual MRR designs considered in these papers. Consequently, the results in [52, 53, 58] are limited to the MRR structures considered in these works. As we will show, our proposed design-space exploration and optimization in this chapter can be applied to any MRR design problem, including those in [52, 53, 58], providing designers with a comprehensive framework to explore and optimize any MRR under FPVs.

Several prior works have investigated adiabatic MRRs. Luo *et al.* in [52] developed an MRR-based modulator with multi-mode waveguides to improve accuracy and repeatability of the MRRs' resonant wavelength. In their work, the ring waveguide width is increased to reduce phase errors associated with side-wall etching. Su *et al.* in [53] proposed the use of adiabatic rings to design MRRs with a high Q-factor of 27000 and improved tolerance to FPVs. The work in [59] presented an adiabatic MRR to enable terabit-per-second chip-to-chip optical I/O. The work in [60] proposed thermally tuned adiabatic MRRs with 0.49 nm/mW tuning efficiency and low resonance loss of 0.085 dB.

3.3 Modeling Silicon Photonic MRR Under Fabrication-Process Variation

In this section, we present the fundamental analytical models required to study the impact of FPVs in MRRs. These models lay the foundation for our proposed design-space exploration and optimization in Sections 3.4 and 3.5. As shown in Figs. 3.1(a) and 3.1(b), we consider passive MRRs designed using strip waveguides and active MRRs constructed using ridge waveguides, which allow for electrical connections to be made to the waveguides [44] (e.g., through P-N junctions), as shown in Fig. 3.1(c). Moreover, Figs. 3.1(d) and 3.1(e) show a cross section of a strip and a ridge waveguide while specifying the waveguide width (w), SOI thickness (t), and slab thickness (h). Hereafter and unless specifically mentioned, we use "MRRs" to refer to both passive and active MRRs.

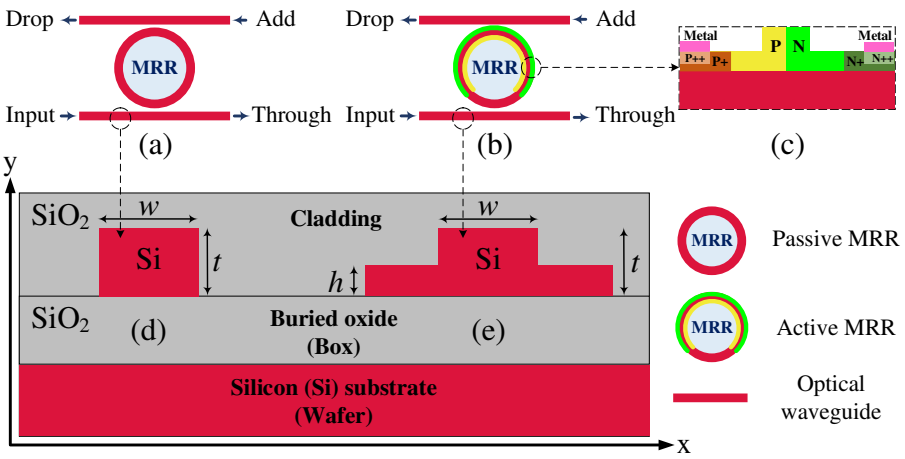


Figure 3.1: An overview of an MRR add-drop filter showing waveguide width (w), SOI thickness (t), and slab thickness (h) in (a) a passive and (b) an active MRR with (c) a P-N junction. Cross section of (d) a strip and (e) a ridge waveguide. Here, Si and SiO_2 denote silicon and silicon dioxide, respectively.

The effective index (n_{eff}) in a waveguide depends on the optical wavelength and the critical dimensions of the waveguide (i.e., w, t, and h—in the case of a ridge waveguide) [12]. The relation can be written as:

$$n_{\text{eff}}(\lambda, w, t, h) = \left(\frac{\lambda}{2\pi}\right) \beta(\lambda, w, t, h), \tag{3.1}$$

where β is the propagation constant and λ is the optical wavelength. Note that numerical solutions of β can be obtained using various numerical and approximation methods (e.g., effective-index method [61]). Leveraging (3.1), we can define the group index (n_{g}) in a waveguide as:

$$n_{\text{g}}(\lambda, w, t, h) = n_{\text{eff}}(\lambda, w, t, h) - \lambda \frac{\partial n_{\text{eff}}(\lambda, w, t, h)}{\partial \lambda}. \tag{3.2}$$

An MRR can be either on or off resonance depending on the resonant wavelength of the MRR and the optical wavelength on the input waveguide. In an MRR add-drop filter (see Fig. 3.1(a) as an example), when the round-trip optical phase in the MRR is an integer multiple of 2π, the MRR is on resonance and it drops the input signal. Otherwise, the MRR is off resonance and the optical signal on the input waveguide passes the MRR towards the through port. The resonant wavelength (λ_R) in an MRR can be calculated based on the effective index (defined in (1)) as:

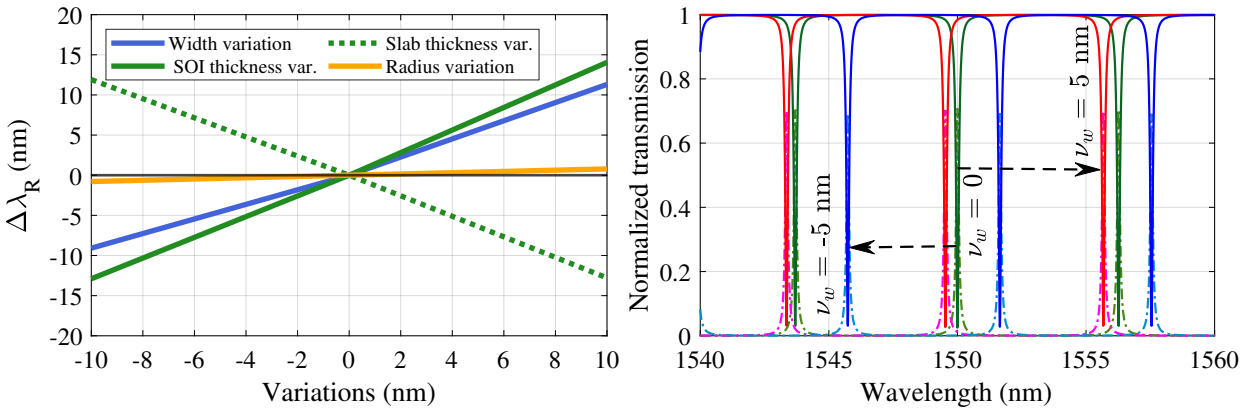
$$\lambda_R(w,t,h,R) = \left(\frac{2\pi R}{m}\right) n_{eff}(\lambda_R, w, t, h), \tag{3.3}$$

where R is the radius of the MRR and m is an integer that denotes the order of the resonant mode. According to (3.3), the resonant wavelength in an MRR depends on the critical dimensions of the MRR, including the waveguide width, SOI thickness, slab thickness (in case of an active MRR), and MRR radius, in which any slight variations will deviate the resonant wavelength. Such a deviation is known as the resonant-wavelength shift. Considering the first-order approximation of the waveguide dispersion, the resonant-wavelength shift ($\Delta\lambda_R$) in an MRR can be modeled as:

$$\begin{aligned}\Delta\lambda_R(\lambda_{R0}, w', t', h', R') =& \frac{\lambda_{R0}\left|n_{\text{eff}}(\lambda_{R0}, w, t, h) - n_{\text{eff}}(\lambda_{R0}, w', t', h')\right|}{n_{\text{g}}(\lambda_{R0}, w, t, h)} + \\ & \left(\frac{\lambda_{R0} R'}{R}\right) \frac{n_{\text{eff}}(\lambda_{R0}, w, t, h)}{n_{\text{g}}(\lambda_{R0}, w, t, h)}.\end{aligned} \tag{3.4}$$

Here, λ_{R0} is the nominal resonant wavelength. Moreover, $w' = w \pm \nu_w$, $t' = t \pm \nu_t$, $h' = h \pm \nu_h$, and $R' = R \pm \nu_R$, where ν_w, ν_t, ν_h, and ν_R are the variations in the waveguide width, SOI thickness, slab thickness, and MRR radius, respectively. In this chapter, we assume independent variations (e.g., ν_R is independent of ν_w). Leveraging (3.4), Fig. 3.2(a) shows the resonant-wavelength shift in an active MRR while considering one FPV at a time and a variation range of $[-10, 10]$ nm (the x-axis), considered as an example. As can be seen, the resonant wavelength changes almost linearly under all the different variations, and the impact of each variation on the MRR resonant wavelength is different. A similar trend was reported in [55]. As an example, Fig. 3.2(b) shows the optical spectrum of the MRR in Fig. 3.2(a)—simulated using the transfer-matrix method [1]—where there is a red and a blue shift with $\nu_w = 5$ nm and $\nu_w = -5$ nm, respectively. There is a good agreement between the resonant-wavelength shift results calculated in Fig. 3.2(a) and simulated in Fig. 3.2(b).

FPVs also impact the through- and drop-port response in MRRs, introducing extra power losses when an optical signal passes or drops into an MRR [46]. Such power loss will degrade the



(a) Resonant-wavelength shift **(b)** MRR optical spectrum

Figure 3.2: (a) Resonant-wavelength shift ($\Delta\lambda_R$) in an active MRR calculated using (3.4) with $w = 400$ nm, $t = 220$ nm, $h = 90$ nm, and $R = 10$ μm under variations in the waveguide width, SOI thickness, slab thickness, and radius (x-axis). (b) Optical spectrum of the MRR simulated using transfer-matrix method [1] with $\lambda_{R0} = 1550$ nm where the resonant wavelength shifts because of width variations ($\nu_w = \pm$ 5 nm).

power efficiency in PICs employing MRRs. Here, we analyze the impact of FPVs on the coupling strength in the coupling region in an MRR (i.e., between the input/drop waveguide and the ring waveguide—see Fig. 3.3) that determines the through- and drop-port response in the MRR. We define κ as the cross-over coupling coefficient between the input/drop waveguide and the ring, and s as the straight-through coefficient associated with the power that remains on the input waveguide [12]. We assume that the input and drop waveguides are coupled symmetrically to the ring, and consider a lossless coupler where $|\kappa|^2 + |s|^2 = 1$. Both κ and s can be calculated precisely based on accurate numerical methods (e.g., FDTD) or using approximate methods such as the supermode theory [46]. A compact systematic model to study the impact of FPVs on the cross-over coupling coefficient (κ) in MRRs can be defined as:

$$\kappa(\lambda_R, w', t', h', R') = f\left(n_{e/o}(\lambda_R, w', t', h'), R', g^{-1}\right). \tag{3.5}$$

Considering (3.5), κ is a function of the effective indices of the even (n_e) and odd (n_o) supermodes in the coupling region of an MRR that change under FPVs [62] (see Appendix A). Moreover, κ is directly proportional to the MRR radius (R) and inversely proportional to the gap (g), hence g^{-1} in (3.5). The gap is defined as the edge-to-edge distance between the input/drop

waveguide and the ring waveguide. Note that $f()$ in (3.5) is a function to calculate the cross-over coupling in MRRs that can be defined based on the method described in [44].

Q-factor is the measure of the sharpness of the resonance relative to its central frequency that impacts the optical channel spacing, crosstalk, bandwidth, and other characteristics in MRRs [12]. In particular, it is desirable to have MRRs with a high Q-factor for DWDM applications. Nevertheless, FPVs can deteriorate the Q-factor in MRRs. Assuming a lossless coupler and using (3.2) and (3.5), the Q-factor in an MRR add-drop filter under FPVs can be defined as:

$$Q(\lambda_R, w', t', h', R') = \left(\frac{\pi L}{\lambda_R}\right) \frac{n_{\mathrm{g}}(\lambda_R, w', t', h')\sqrt{a\left(1-\kappa^2(\lambda_R, w', t', h', R')\right)}}{1-a\left(1-\kappa^2(\lambda_R, w', t', h', R')\right)}. \tag{3.6}$$

Here, $L = 2\pi R$ is the round-trip length in the MRR. Also, a is the single-pass amplitude transmission, which includes the propagation loss in the MRR and loss in the couplers, and can be calculated as $a^2 = exp(-\alpha L)$, where α is the power attenuation coefficient [12].

The optical bandwidth of interest plays an important role in deciding the optimal design parameters in MRRs. The optical bandwidth in an MRR can be defined based on the frequency at which half the power is incident in the channel (i.e., frequency at 3dB). Therefore, employing (3.2) and (3.5), the 3dB bandwidth in an MRR can be attained by observing the full width at half maximum (FWHM) of the resonance spectrum, defined as:

$$FWHM(\lambda_R, w', t', h', R') = \left(\frac{\lambda_R^2}{\pi L}\right) \frac{1-a(1-\kappa^2(\lambda_R, w', t', h', R'))}{n_{\mathrm{g}}(\lambda_R, w', t', h')\sqrt{a(1-\kappa^2(\lambda_R, w', t', h', R'))}}. \tag{3.7}$$

The 3dB bandwidth of an MRR used in a DWDM-based add-drop configuration should be large enough to accommodate the bandwidth of the optical signal to be dropped. A narrow bandwidth will result in undesired truncation of the signal spectrum causing distortions [63]. Nonetheless, a large 3dB bandwidth should be avoided too as it can cause severe crosstalk noise (e.g., inter-channel crosstalk) if the channel density is high [64]. Leveraging (3.7) and considering c to be the speed of light in vacuum, the 3dB bandwidth in an MRR (Δv, in GHz) can be modeled as [44]:

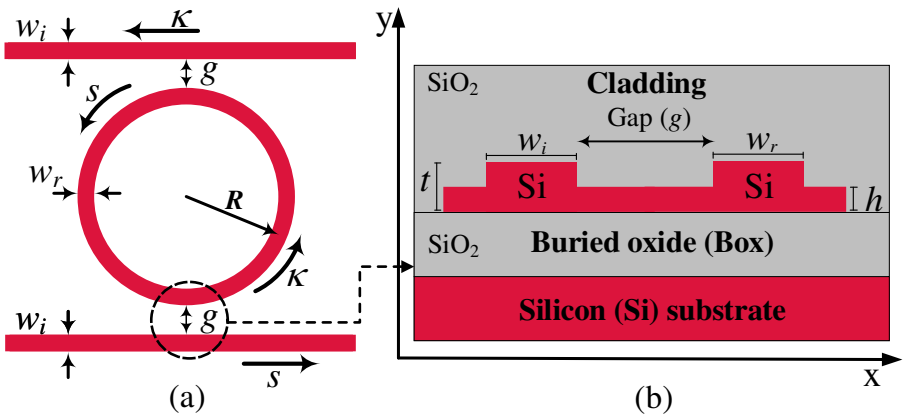


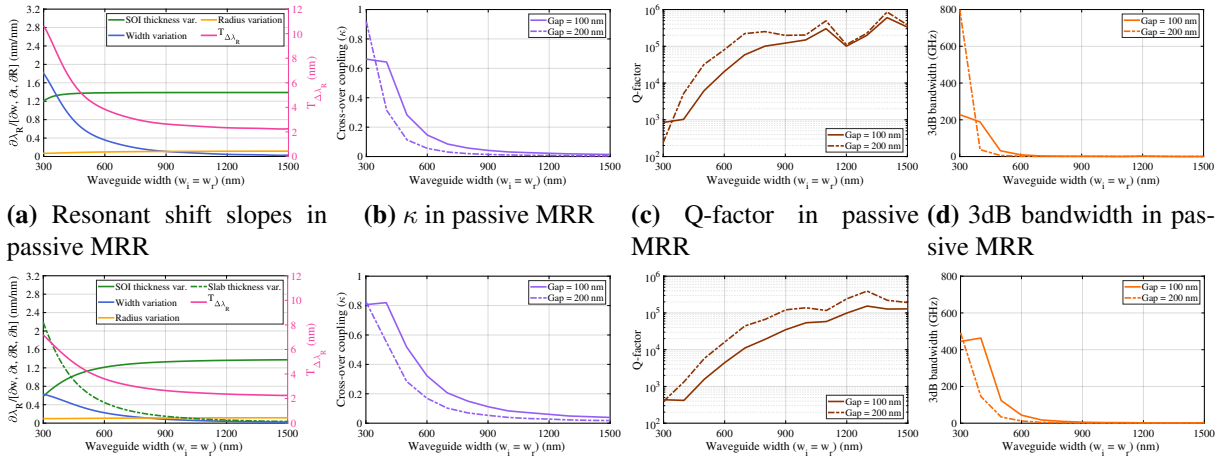
Figure 3.3: A cross section of the coupling region in an active MRR add-drop filter with different physical- and device-level design parameters. Here, w_i is the input/drop waveguide width and w_r denotes the ring waveguide width. Note that $h = 0$ for a passive MRR.

$$\Delta v(\lambda_R, w', t', h', R') = \left(\frac{c}{\lambda_R^2}\right) \text{FWHM}(\lambda_R, w', t', h', R') \tag{3.8}$$

The analytical models proposed in this chapter are computationally inexpensive and derived to capture the impact of various physical-level FPVs on the critical dimensions and device-level performance in MRRs. Such models can enable a design-space exploration in MRRs under different variations, as discussed in the next section.

3.4 MRR Design-Space Exploration under Fabrication-Process Variations

Leveraging the proposed analytical models in Section 3.3, we present a comprehensive design-space exploration for MRRs under different FPVs in this section. In particular, we analyze the impact of different variations on the resonant wavelength, cross-over coupling, Q-factor, and 3dB bandwidth in MRRs. In addition, we explore the impact of changing the design parameters in MRRs (e.g., the waveguide width) on the device performance under different FPVs. As the input/drop and ring waveguides are in proximity, we assume that the variations on the input/drop waveguide and those on the ring waveguide are the same in a single MRR.



(a) Resonant shift slopes in passive MRR **(b)** κ in passive MRR **(c)** Q-factor in passive MRR **(d)** 3dB bandwidth in passive MRR

(e) Resonant shift slopes in active MRR **(f)** κ in active MRR **(g)** Q-factor in active MRR **(h)** 3dB bandwidth in active MRR

Figure 3.4: Resonant-wavelength shift slopes and device performance in passive ((a)–(d)) and active ((e)–(h)) MRRs when the input/drop and ring waveguide widths increase from 300 to 1500 nm (x-axis). Here, (a) and (e) also show the total resonant-wavelength shift ($T_{\Delta\lambda_R}$). Results are for the fundamental TE mode with the parameters in Table 3.1 when $w_i = w_r$.

Considering (3.4), we define the total resonant-wavelength shift in an MRR ($T_{\Delta\lambda_R}$) by distinguishing the contribution of each variation to the resonant-wavelength shift in the MRR. A generalization based on passive and active MRRs can be made using the following model:

$$T_{\Delta\lambda_R}(\lambda_R, w', t', h', R') = \frac{\partial\lambda_R}{\partial w}(\sigma_w) + \frac{\partial\lambda_R}{\partial t}(\sigma_t) + \frac{\partial\lambda_R}{\partial h}(\sigma_h) + \frac{\partial\lambda_R}{\partial R}(\sigma_R). \tag{3.9}$$

Here, $\frac{\partial\lambda_R}{\partial w}$, $\frac{\partial\lambda_R}{\partial t}$, $\frac{\partial\lambda_R}{\partial h}$, and $\frac{\partial\lambda_R}{\partial R}$ denote the rate of changes in the MRR resonant wavelength with respect to the variations in the waveguide width, SOI thickness, slab thickness, and radius, respectively (i.e., *resonant-wavelength shift slopes*). Note that $\frac{\partial\lambda_R}{\partial h} = 0$ in the case of passive MRRs using strip waveguides. Moreover, σ_w, σ_t, σ_h, and σ_R are the standard deviations associated with the variations in the waveguide width, SOI thickness, slab thickness, and radius, respectively. Our prior work in [26, 46, 62], and that from other groups [27, 33] showed that the impact of different variations in MRRs can be combined linearly (see (3.9)). Note that $\nu_{w,t,h,R}$ in (3.4) can be initialized based on these standard deviations. Moreover, such standard deviations can be quantified through various fabrications or obtained from a silicon photonic fabrication vendor.

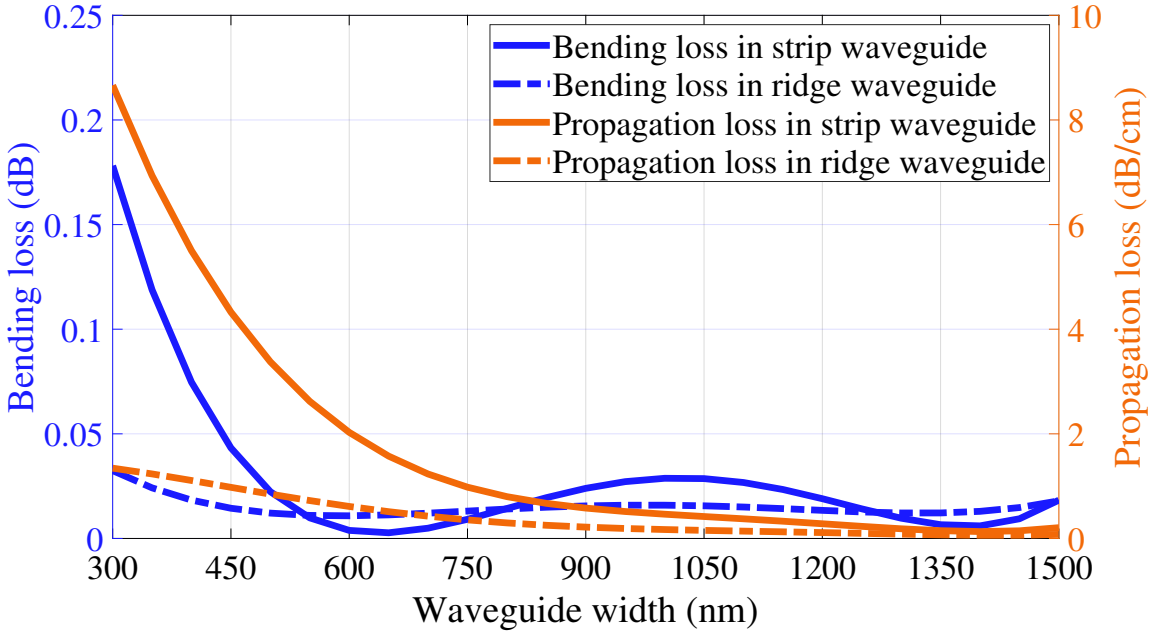


Figure 3.5: Bending and propagation loss in strip and ridge waveguide as the waveguide width increases.

The resonant-wavelength shift slopes ($\frac{\partial \lambda_R}{\partial w}$, $\frac{\partial \lambda_R}{\partial t}$, $\frac{\partial \lambda_R}{\partial h}$, and $\frac{\partial \lambda_R}{\partial R}$) in (3.9) can be approximated using a linear model as we found that the resonant wavelength in MRRs changes almost linearly under different variations (see Fig. 3.2(a)). For instance, using (3.4), $\frac{\partial \lambda_R}{\partial w}$ can be approximated as:

$$\frac{\partial \lambda_R}{\partial w} \approx \left| \frac{\Delta\lambda_R(\lambda_{R0}, w+\epsilon, t, h, R) - \Delta\lambda_R(\lambda_{R0}, w-\epsilon, t, h, R)}{2\epsilon} \right|, \tag{3.10}$$

where ϵ is an arbitrary parameter denoting a slight change in the waveguide width. Similarly, $\frac{\partial \lambda_R}{\partial t}$, $\frac{\partial \lambda_R}{\partial h}$, and $\frac{\partial \lambda_R}{\partial R}$ can be approximated.

As shown in Fig. 3.3, the design space of MRRs can be divided into physical-level parameters (e.g., waveguide width, SOI thickness, and slab thickness) as well as device-level parameters (e.g., radius and gap), all of which can be affected under FPVs. Among these parameters, only the waveguide width, MRR radius, and the gap can be determined (and explored) at design-time as the SOI and slab thickness are limited by the host wafer and available etching depths in the fabrication-process. Note that we do not consider the gap in an MRR in our design-space exploration in this chapter, but we account for the variations in the gap through waveguide width variations in the MRR (i.e., width variations in the coupling region in an MRR that impact the gap). Moreover,

Table 3.1: Different parameters used in our simulations

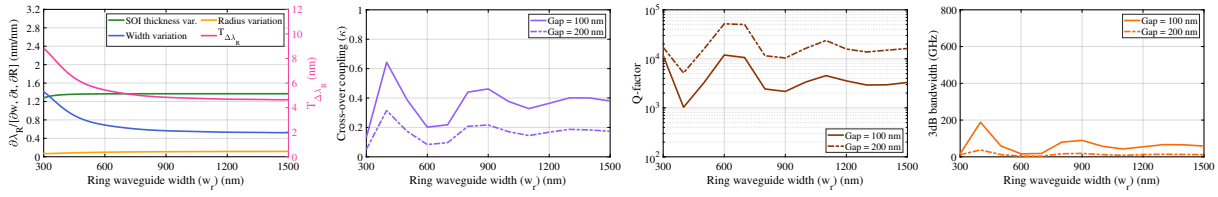
Parameter	Value	Parameter	Value
SOI thickness (t)	220 nm	Slab thickness (h)	90 nm
Radius (R)	10 μm	Wavelength (λ or λ_{R0})	1550 nm
σ_w	4.9 nm	σ_t	1.5 nm
σ_h	1.5 nm	σ_R	0.8 nm

the impact of radius variations in MRRs is minimal (see Fig. 3.2(a)). Therefore, in the rest of this section an effort is made to explore the impact of altering the waveguide width at design-time on the MRR performance under FPVs. In particular, we consider two scenarios where (see Fig. 3.3): *A.* the input/drop waveguide width (w_i) and ring waveguide width (w_r) are the same, and *B.* they are different, as discussed next.

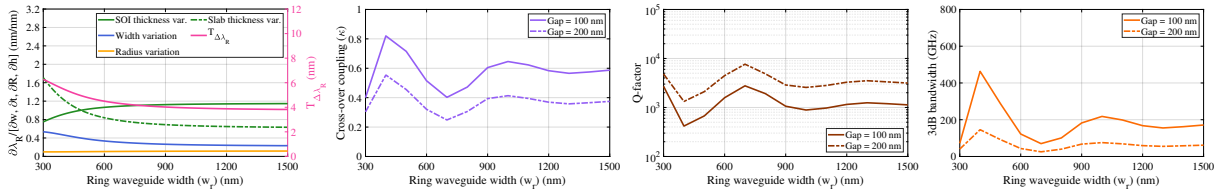
3.4.1 MRR performance analysis when $w_i = w_r$

Using (3.4), (3.9), nominal design parameters in Table 3.1, and Lumerical MODE [65], which is a commercial eigenmode solver and simulator, we quantitatively simulate the rate of changes in the resonant wavelength w.r.t. different variations (i.e., $\frac{\partial \lambda_R}{\partial w}$, $\frac{\partial \lambda_R}{\partial t}$, $\frac{\partial \lambda_R}{\partial h}$, and $\frac{\partial \lambda_R}{\partial R}$) in passive and active MRRs. Results for the fundamental transverse-electric (TE) mode are shown in Figs. 3.4(a) and 3.4(e) for passive and active MRRs, respectively. We vary the input (w_i) and ring (w_r) waveguide width together from 300 to 1500 nm (selected as an example). Note that increasing the waveguide width in MRRs will contribute to higher-order mode excitation (i.e., multi-mode waveguides). While multi-mode waveguides are beneficial in some applications (e.g., mode-division multiplexing), we will discuss feasible solutions to alleviate higher-order mode excitation in Section 3.5. Therefore, our design-space exploration in this section focuses mainly on understanding the impact of increasing the waveguide width on the rate of changes in the MRR resonant wavelength due to the different variations.

One can observe from Figs. 3.4(a) and 3.4(e) that the resonant-wavelength shift slopes corresponding to the different variations are all different. In particular, the impact of waveguide width and slab thickness variations ($\frac{\partial \lambda_R}{\partial w}$ and $\frac{\partial \lambda_R}{\partial h}$) decreases as the waveguide width increases; this de-



(a) Resonant shift slopes in passive MRR **(b)** κ in passive MRR **(c)** Q-factor in passive MRR **(d)** 3dB bandwidth in passive MRR



(e) Resonant shift slopes in active MRR **(f)** κ in active MRR **(g)** Q-factor in active MRR **(h)** 3dB bandwidth in active MRR

Figure 3.6: Resonant-wavelength shift slopes and device performance in passive ((a)–(d)) and active ((e)–(h)) unconventional MRRs. Here, (a) and (e) also show the total resonant-wavelength shift ($T_{\Delta\lambda_R}$). Results are for the fundamental TE mode with the parameters in Table 3.1 when $w_i = 400$ nm (considered as an example) and $w_i \neq w_r$. The x-axis shows the ring waveguide width (w_r) changes from 300 to 1500 nm.

notes that MRR's resonance tolerance to the variations in waveguide width and slab thickness improves as the waveguide width increases. On the other hand, the impact of SOI thickness variations remains high and dominant as the waveguide width increases in Figs. 3.4(a) and 3.4(e) (i.e., increasing the waveguide width does not have much impact here). Another observation is that the impact of radius variations on the resonant-wavelength shift is small in both passive and active MRRs (e.g., ≈0.12 nm shift in the resonance with 1 nm change in the MRR radius [55]). Note that the results in Figs. 3.4(a) and 3.4(e) are independent of the MRR radius and gap. Employing (3.9) and the σ values in Table 3.1—obtained by averaging different σ values in Table 2.1 in Section 2.3—we also show the total resonant-wavelength shift ($T_{\Delta\lambda_R}$) in passive and active MRRs in Figs. 3.4(a) and 3.4(e), respectively (see the second y-axis). We observe that as the waveguide width increases, $T_{\Delta\lambda_R}$ decreases, thereby the MRR tolerance to different FPVs increases.

Leveraging (3.5) and the parameters in Table 3.1, we analyze the cross-over coupling (κ) in passive and active MRRs with $g = 100$ and 200 nm while increasing the waveguide width ($w_i = w_r$) from 300 to 1500 nm. Results, shown in Figs. 3.4(b) and 3.4(f), indicate that while increasing

the waveguide width helps improve an MRR tolerance to FPVs (see Figs. 3.4(a) and 3.4(e)), the cross-over coupling exponentially decreases as the waveguide width increases in the MRR. Accordingly, the drop-port response in MRRs with wider waveguides will be degraded. Note that the κ at $w_i = w_r \approx$300 nm in Figs. 3.4(b) and 3.4(f) is higher for the gap of 200 nm as compared to that for the gap of 100 nm. We believe that this phenomenon is because of the optical signal being confined mostly in the coupling region rather than in the waveguide core for such an MRR design, similar to that of a slot waveguide [66].

As discussed in Section 3.3, the cross-over coupling also determines the Q-factor and 3dB bandwidth in MRRs. Leveraging (3.6) and (3.8), Figs. 3.4(c), 3.4(d), 3.4(g), and 3.4(h) show the results for the Q-factor and 3dB bandwidth performance in passive and active MRRs when the MRR waveguide width ($w_i = w_r$) increases. As can be seen, as the MRR waveguide width increases, the Q-factor increases but the 3dB bandwidth considerably reduces to $\approx$0.1 GHz. Note that the Q-factor and 3dB bandwidth calculations in our work include the impact of increasing the waveguide width on the propagation loss and bending loss in MRRs. In particular, we estimate the waveguide propagation loss using the n_w model approximation [67], which is particularly useful if one only needs to quantify the relative comparisons of the propagation losses among different waveguide geometries under the same fabrication conditions [68]. Moreover, we estimated the changes in bending loss using Lumerical FDTD [65]. We found that both the propagation and bending losses decrease as the waveguide width increases, as shown in Fig. 3.5.

To summarize our findings in this sub-section, we demonstrate that one can improve the MRR tolerance to FPVs by increasing the waveguide width in the MRR at design-time, but at the cost of severe reductions in the cross-over coupling that determines the drop-port response, Q-factor, and 3dB bandwidth in MRRs. To overcome this challenge while maintaining high MRR tolerance to FPVs, in the next sub-section an effort is made to study MRRs under FPVs while considering $w_i \neq w_r$ (i.e., an unconventional MRR).

3.4.2 MRR performance analysis when $w_i \neq w_r$

In this sub-section, we assume that the input and drop waveguide widths (i.e., w_i) are equal and can be different from the ring waveguide width (w_r)—see Fig. 3.3. The cross-over coupling in MRRs is proportional to the overlap and the interaction between the optical modes in the input and the ring waveguide. When increasing $w_i = w_r$ in an MRR, such an overlap reduces as the fundamental optical modes will be more confined inside the waveguide cores, and hence κ decreases. Therefore, here we examine a possible solution based on considering $w_i \neq w_r$ and then increasing w_r only (i.e., an *unconventional* MRR with $w_i < w_r$) to enhance κ in MRRs while improving the MRR tolerance to FPVs. Note that one can increase w_r adiabatically and similar to [53].

To enable design-space exploration in an unconventional MRR, we need to enhance the proposed analytical models in Section 3.3 to account for the effect of unequal input/drop and ring waveguide widths. In particular, the effective index in the MRR is the main parameter that needs to be recalculated when $w_i \neq w_r$ in an MRR. For simplicity and similar to the effective index analysis in waveguide grating structures [1], we consider the average effective index (n_{eff}) of the input/drop and the ring waveguide for our calculations in this sub-section, based on which all the analytical models proposed in Section 3.3 can be updated. Similar to Section 3.4.1, we also consider the impact of changes in bending and propagation losses in waveguides associated with unconventional MRRs (see Fig. 3.5) as the ring waveguide width increases.

Similar to Section 3.4.1 and using the updated analytical models in Section 3.3, Lumerical MODE [65], and the parameters in Table 3.1, we explore MRRs with $w_i = 400$ nm (considered as an example) while increasing the ring waveguide width (w_r) from 300 nm to 1500 nm. Note that starting w_r from 300 nm is considered as an example and to show MRR performance when $w_r < w_i$. Figs. 3.6(a) and 3.6(e) show the resonant-wavelength shift slopes under the different variations in passive and active MRRs, respectively. As can be seen, the impacts of width, SOI thickness, slab thickness, and radius variations on the resonant wavelength of MRRs are similar to those shown in Figs. 3.4(a) and 3.4(e). However, compared to when $w_i = w_r$, the resonant-wavelength shift slopes, and hence the rate of changes in the resonant wavelength of an MRR under

FPVs, are slightly higher when $w_i < w_r$. Consequently, the total resonant-wavelength shift—calculated using the σ parameters from Table 3.1—in both passive and active MRRs is higher when $w_i < w_r$. Nevertheless, $T\Delta\lambda_R$ still decreases as w_r increases in Figs. 3.6(a) and 3.6(e), indicating that the proposed solution can help improve the tolerance of passive and active MRRs to FPVs.

Employing (3.5) and the parameters in Table 3.1, Figs. 3.6(b) and 3.6(f) show the cross-over coupling in, respectively, passive and active MRRs for $g = 100$ and 200 nm when $w_i = 400$ nm and w_r increases from 300 to 1500 nm. Observe that κ increases at first when the ring wveguide width (w_r) increases from 300 nm to 400 nm, where the coupling is the highest at $w_i = w_r = 400$ nm. However, the coupling starts decreasing for $w_r > 400$ nm, until it increases again and reaches its second maximum where $w_r \approx 2w_i$. We simulated κ with other values for w_i and we found the same conclusion as the one in Figs. 3.6(b) and 3.6(f): κ increases when $w_r \approx \rho w_i$, where ρ is an integer. We further discuss such an increase in κ in Appendix A. Moreover, Figs. 3.6(c), 3.6(d), 3.6(g), and 3.6(h) show the Q-factor and 3dB bandwidth performance in unconventional passive and active MRRs while also including the changes in the propagation and bending losses in the MRRs (see our discussion in Section 3.4.1). A similar trend to κ can be observed in the Q-factor and 3dB bandwidth. In particular, unlike when $w_i = w_r$ in Fig. 3.4, as w_r increases in an unconventional MRR, the decrease in Q-factor and 3dB bandwidth is inconstant. This is an interesting finding as it shows that an optimal MRR design is feasible with high tolerance to FPVs and also an acceptable cross-over coupling, Q-factor, and 3dB bandwidth (compare the results in Figs. 3.4 and 3.6).

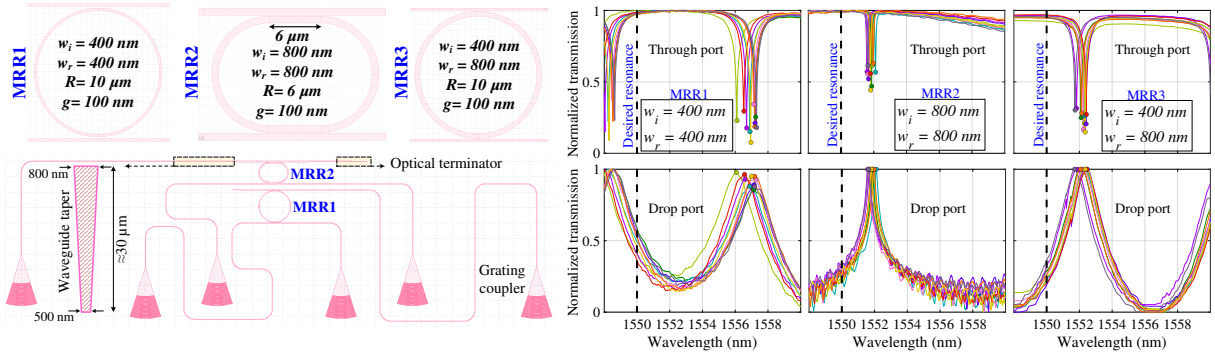
To summarize our findings in this sub-section, we show that passive and active MRRs with unequal and variable w_i and w_r (called unconventional MRRs in this chapter) can not only achieve high tolerance to FPVs but the cross-over coupling in these MRRs—and hence their drop-port response, Q-factor, and 3dB bandwidth—can be improved as well. Note that our conclusions and findings are independent of the assumption $w_i = 400$ nm, considered as an example in this sub-section, and we show results with other w_i values in Section 3.5. To verify our results and findings

in Sections 3.4.1 and 3.4.2, we fabricate three different MRRs with variable w_i and w_r and evaluate the total resonant-wavelength shift and performance of our designs in the next sub-section.

3.4.3 Fabrication Results

We designed three passive TE-polarized MRR add-drop filters with a gap of 100 nm to experimentally validate our design-space exploration and results in this section. Fig. 3.7(a) shows the device layout of the different designed MRRs: MRR1 with $w_i = w_r = 400$ nm, MRR2 with $w_i = w_r = 800$ nm (see Fig. 3.4), and MRR3 with $w_i = 400$ nm and $w_r = 800$ nm (see Fig. 3.6). The radius (R) in MRR1 and MRR3 is set to be 10 μm. For MRR2, a racetrack resonator is considered with a radius and coupler length of 6 μm to ensure sufficient coupling between input/drop and ring waveguide (see κ in Fig. 3.4(b)). Note that the models developed in Section 3.3 can be easily extended to racetrack resonators (also see Appendix A). All the MRRs are designed to resonate at 1550 nm (i.e., desired/nominal resonance). Ten identical copies of each MRR were placed on a 1.5×0.6 mm^2 chip fabricated with a high-resolution electron-beam (EBeam) lithography system. Fig. 3.7(a) indicates an example of a unit cell of the designed MRRs with grating couplers designed for 1550 nm quasi-TE operation. Moreover, 220 nm thick SOI strip waveguides with a width of 500 nm are used for routing. Waveguide tapers of lengths 10–30 μm are employed to ensure a single-mode operation.

Employing an automated silicon photonic testing station, we characterized the through- and drop-port responses of all the 30 MRRs (i.e., 10 identical copies per designed MRR), as shown in Fig. 3.7(b). Note that thermal variations were avoided using a thermal stage to preserve the chip temperature during the test. In Fig. 3.7(b), the desired resonant response is shown (at 1550 nm), and the resonant peaks that are specified by circles all belong to the same resonant mode. Table 3.2 summarizes the measurement results in MRR1–3. As can be seen, these experimental results are consistent with our design-space exploration results in Figs. 3.4 and 3.6. In particular, compared to MRR2 and MRR3, MRR1 has the highest cross-over coupling but lowest Q-factor of $\approx$500 (estimated using a Lorentzian fit) with an average total resonant-wavelength shift of 7.1 nm among



(a) Details of fabricated MRRs (top) and cell layout example (bottom)

(b) Measurement results (30 MRRs)

Figure 3.7: (a) An example of the cell layout (bottom, MRR1 and MRR2 only) of the three designed passive TE-polarized MRRs with their design specifications (top). Note that 500-nm-wide strip waveguide is used for routing for all the MRRs. (b) Measured through- and drop-port responses obtained by testing 30 identical copies of MRRs in Fig. 3.7(a). All the MRRs were designed to resonate at 1550 nm, specified as desired resonance in the figures.

the 10 resonant-wavelength peaks (i.e., the worst tolerance to FPVs among the fabricated MRRs). Moreover, MRR2 shows the best tolerance to FPVs as the average resonant-wavelength shift in MRR2 is only 1.8 nm. Nevertheless, MRR2 compromises cross-over coupling (see Fig. 3.4(b)), hence the noise on the drop-port response in Fig. 3.7(b), with the highest Q-factor of ≈1800. MRR3, designed based on our proposed design-space exploration in Fig. 3.6, shows an average total resonant-wavelength shift of 2.1 nm (i.e., lower than MRR1 and slightly higher than MRR2), indicating a 70% increase in tolerance to FPVs (compared to MRR1) with a higher κ (compared to MRR2) and a Q-factor of ≈600, which is higher than that for MRR1. Moreover, MRR3 has the best extinction ratio of 12 dB. It is worth mentioning that the difference between the calculated Q-factor results in Figs. 3.4 and 3.6 and those reported for our fabricated MRRs is due to the higher propagation (≈3 dB/cm) and coupler losses in our fabricated MRRs (note that our calculations and results in Figs. 3.4 and 3.6 consider a lossless coupler). Note that comparing our fabrication results with those from prior related work (e.g., [52, 53]) is unfair and invalid because the design parameters, objectives, and the fabrications are all different.

Table 3.2: Measurement results in MRR1-3

Parameter	MRR1	MRR2	MRR3
Total shift—Average	7.1 nm	1.8 nm	2.1 nm
Total shift—Standard deviation	0.38 nm	0.17 nm	0.25 nm
Q-factor—Lorentzian fit	500	1800	600

3.5 MRR Design Optimization under Fabrication-Process Variations

Considering Figs. 3.4 and 3.6, employing wider waveguides in MRRs can help improve their tolerance to FPVs. However, considering different performance metrics altogether, several of which are contradicting, our design-space exploration in Section 3.4 does not identify design parameters in MRRs (i.e., w_i and w_r in our paper) required to attain a high tolerance to FPVs while also achieving a desirable Q-factor and 3dB bandwidth and not imposing high overhead (e.g., silicon area) in MRRs. Indeed, choosing design parameters in MRRs based on a single performance metric may cause performance degradation when considering other metrics. For example, if an MRR is solely designed on the basis of having a high Q-factor, a designer might need to compensate in terms of photodiode sensitivity and also limited 3dB bandwidth of such a design. Leveraging the proposed MRR design-space exploration in Section 3.4, we present design optimization for passive and active MRRs under FPVs in this section.

The main goal in our optimization is to minimize the total resonant-wavelength shift in MRRs while considering MRR Q-factor and 3dB bandwidth. Note that κ, propagation loss, and bending loss are all considered in the optimization through their impact on Q-factor and 3dB bandwidth. Furthermore, to provide a designer with a peripheral vision while considering various properties of MRRs, we consider some additional metrics such as MRR footprint and higher-order mode excitation (i.e., multi-mode waveguide effect) in MRRs. As a result, we can define a set of constraints (Ω) for the optimization as:

$$\Omega = \{(Q_m \leq Q \leq Q_M) \cap (B_m \leq B \leq B_M) \cap (A \leq A_M) \cap (\Delta n_{2nd} < \Delta n_{1st})\}. \quad (3.11)$$

Here, Q_m and B_m (Q_M and B_M) denote the minimum (maximum) acceptable Q-factor and 3dB bandwidth that can be determined by the designer. Note that $Q_{m,M}$ and $B_{m,M}$ are often application specific. Moreover, A_M is the maximum silicon-area overhead that is acceptable, and Δn_{1st} and Δn_{2nd}—considered to account for higher-order mode excitation—are the optical-confinement (guidance) strengths of, respectively, the first-order and the second-order TE mode in the MRR. Considering $\Delta n_{2nd} < \Delta n_{1st}$ ensures that the confinement strength of the fundamental mode is always stronger than that of the second-order mode (see our discussion below). We define the confinement (guidance) strength of an optical mode to be the difference between the effective index of that optical mode and the refractive index of cladding/substrate: $\Delta n = n_{\text{eff}} - 1.44$, where 1.44 is the refractive index of silicon dioxide at 1550 nm. The higher the confinement strength of an optical mode, the stronger the guidance of the optical mode will be in an MRR.

Algorithm 1 MRR Design Optimization under FPVs

Given R, g, σ_w, σ_t, σ_h, and σ_R
for $w_a \in [w_{i\text{-min}}, w_{i\text{-max}}]$ **do** /* Input waveguide */
  for $w_b \in [w_{r\text{-min}}, w_{r\text{-max}}]$ **do** /* Ring waveguide */
    $Q_{w_a,w_b} \leftarrow$ calculate MRR Q-factor using (3.6)
    $B_{w_a,w_b} \leftarrow$ calculate MRR 3dB bandwidth using (3.8)
    $\Delta n_{2nd} \leftarrow \max(n_{\text{eff}2nd}(w_a), n_{\text{eff}2nd}(w_b)) - 1.44$
    $A_{w_a,w_b} \leftarrow$ calculate total silicon area in the MRR
    $C_{w_a,w_b} \leftarrow$ calculate $T_{\Delta\lambda_R}$ with (w_a, w_b) using (3.9)
    if $C_{w_a,w_b} < C^*$ ($C_{w_a,w_b} \leq T_M$) **then**
      if Ω_{w_a,w_b} is True **then** /* see (3.11) */
        $C^* \leftarrow C_{w_a,w_b}$
        Update s^* with (w_a, w_b) (add (w_a, w_b) to s)

For a given MRR design problem (i.e., R and g) and FPV parameters (i.e., σ_w, σ_t, σ_h, σ_R), an optimization search O can be used to find the optimal set of input and ring waveguide widths ($s^* =$

$\{w_i^*, w_r^*\}$) with the minimum cost (C^*), where the cost function is the total resonant-wavelength shift in the MRR: $C = T_{\Delta_{MR}}$, while satisfying the constraints in (3.11):

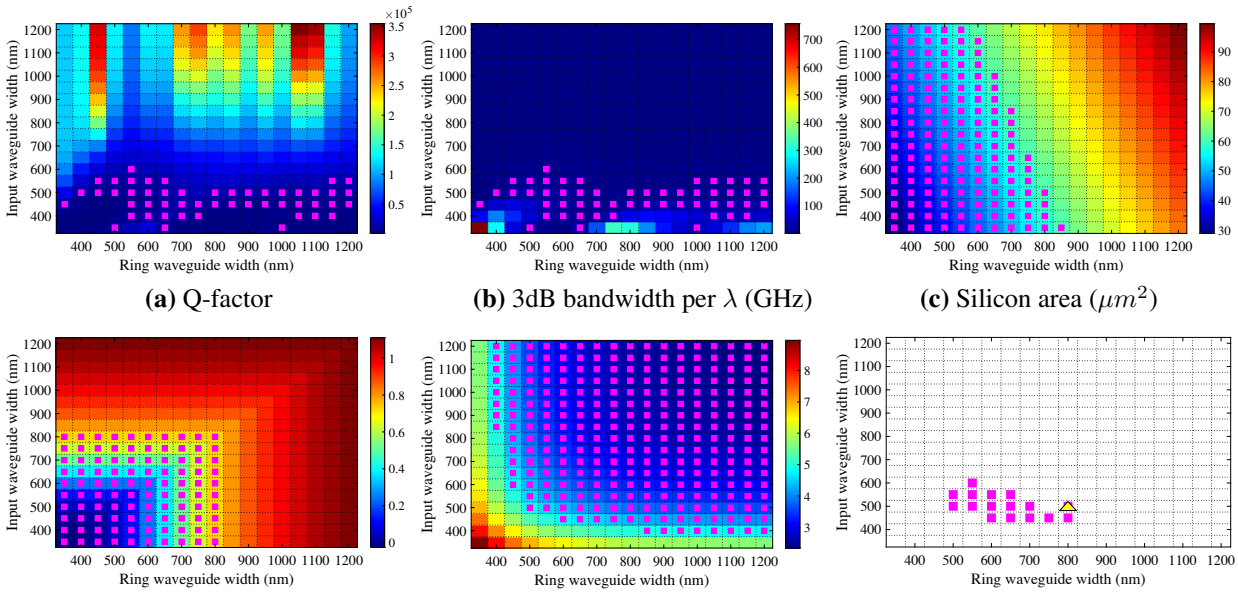
$$s^* \leftarrow O(s \in S, \text{Objective} : \min(C) \text{ s.t. } \Omega), \tag{3.12}$$

where S is a set of all the possible input and ring waveguide widths, defined by the designer. We use an exhaustive search approach, shown in Algorithm 1, to address the optimization in (3.12). Note that such an exhaustive search is made possible thanks to the high computational efficiency of the proposed models in Section 3.3. To provide a designer with more flexibility when, for example, a total resonate-wavelength shift smaller than T_M is still acceptable (e.g., power budget allows for a tuning range up to T_M), the search can also find a set of input and ring waveguide widths (s) that satisfies the constraints in (3.11) with $T_{\Delta\lambda_R} \leq T_M$. If this is desired, the parts underlined in Algorithm 1 must be considered. Note that we do not explore R and g in Algorithm 1 as the impact of R on $T_{\Delta\lambda_R}$ is negligible (see Fig. 3.2(a)) and g does not impact $T_{\Delta\lambda_R}$. Yet, leveraging the analytical models in Section 3.3, the optimization search can be easily extended to include the impact of these parameters on the constraints defined in (3.11).

As an example for the MRR design optimization, we consider a passive (using strip waveguides) and an active (using ridge waveguides) MRR design problem with the design parameters in Table 3.1 and $g = 100$ and 200 nm for the passive and the active MRR, respectively. For brevity, we focus on the TE mode analysis. Similar to Section 3.4, we consider σ values listed in Table 3.1. Note that all the parameters in this section are considered as an example, but the proposed method can be applied to any MRR design problem under FPVs. Leveraging the analytical models in Sections 3.3 and 3.4, we analyze the Q-factor, 3dB bandwidth, and total resonant-wavelength shift in the MRRs while sweeping both the input and ring waveguide widths from $w_{i\text{-min}} = w_{r\text{-min}} = 350$ nm to $w_{i\text{-max}} = w_{r\text{-max}} = 1200$ nm (see lines 2–3 in Algorithm 1). Moreover, we analyze the total silicon-area overhead and the second-order TE mode optical-confinement strength (see our discussion above) in the MRRs.

Figs. 3.8(a) and 3.9(a) show heatmaps for the Q-factor in the passive and active MRR, respectively. A low Q-factor in an MRR can increase the crosstalk noise and power penalty, and a very high Q-factor can put burden on the signal modulation (e.g., by enforcing the step size of the input signal to be smaller, which is limited by the tunable laser), and calls for a precise tuning mechanism in PICs [69]. As shown in Section 3.3, MRR 3dB bandwidth and Q-factor are correlated. Assuming desired MRR 3dB bandwidth to be between 10 GHz and 50 GHz (see below), the MRR Q-factor should be larger than 3800 and smaller than 19000 (i.e., $Q_m = 3800$ and $Q_M = 19000$ in (3.11)). Accordingly, the design points (i.e., w_i and w_r) in Figs. 3.8(a) and 3.9(a) are specified using magenta squares. Figs. 3.8(b) and 3.9(b) show heatmaps for the 3dB bandwidth per wavelength in the passive and active MRR, respectively. A narrow 3dB bandwidth will result in heavy and undesired truncation of an optical signal spectrum, thus causing distortion, while a large 3dB bandwidth will result in higher crosstalk noise and power penalty. Similar to [44] and assuming a minimum signaling rate of 10 Gbps per optical channel (λ) for a wavelength-division multiplexing based link, we assume the desired bandwidth to be greater than 10 GHz and smaller than 50 GHz ($B_m = 10$ GHz and $B_M = 50$ GHz in (3.11)), based on which the design points in Figs. 3.8(b) and 3.9(b) are determined.

Figs. 3.8(c) and 3.9(c) show heatmaps for the silicon area consumption in the passive and active MRR, respectively. The silicon area in an MRR, which affects the fabrication cost, increases as w_i and w_r increase. Given the critical dimensions of an MRR add-drop filter, we define the silicon area as the sum of silicon-surface area on the input and drop waveguides and that on the ring. For simplicity, we consider the same silicon-area model for passive and active MRRs. As an example, we consider $w_i = w_r = 400$ nm—which corresponds to the total silicon area of 33.1 μm^2 for the MRR case study considered in this section—as a baseline for silicon-area consumption. Accordingly, as w_i and w_r increase, we consider an upper limit for the resulting silicon area as twice the baseline (i.e., $A_M = 2\times 33.1$ μm^2 in (3.11)), as indicated by the design points in Figs. 3.8(c) and 3.9(c).

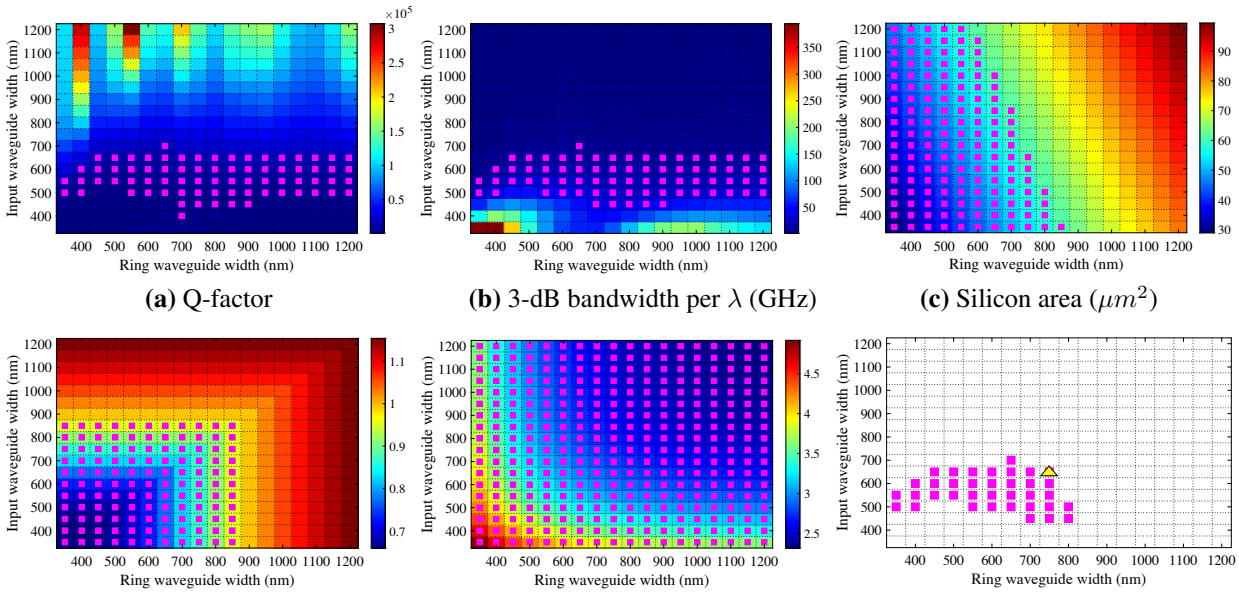


(a) Q-factor **(b)** 3dB bandwidth per λ (GHz) **(c)** Silicon area (μm^2)

(d) 2nd-order TE mode confinement **(e)** Total resonant shift ($T_{\Delta\lambda_R}$) (nm) **(f)** Optimal design points

Figure 3.8: Performance of a passive MRR designed using the parameters in Table 3.1 and with $g = 100$ nm where both the input and ring waveguide widths change from 350 to 1200 nm. The desired design points are selected and shown with magenta squares and the optimal MRR design region in (f) satisfies all the requirements in (a)–(e). Moreover, the yellow triangle in (f) shows the design point at which the total resonant-wavelength shift is minimum ($T_{\Delta\lambda_R} = 3.8$ nm).

Increasing the waveguide width will excite higher-order modes in MRRs (i.e., multi-mode waveguides). While this could be desired for applications such as mode-division multiplexing [70], PICs are often designed for single-mode operation. Note that higher-order mode excitation in our proposed MRR design can be avoided by engineering the MRR structure (e.g., adiabatically increasing the ring waveguide width [53]). Moreover, when increasing w_i and w_r, we ensure that the optical-confinement strength of the first-order TE mode (fundamental mode) is stronger than that of the second-order TE mode in both the passive and active MRRs. Figs. 3.8(d) and 3.9(d) show the optical-confinement strength of the second-order TE mode—Δn_{2nd} in (3.11) calculated based on line 6 in Algorithm 1—in the passive and active MRR, respectively. We consider 0.76 (for the passive MRR) and 0.99 (for the active MRR) as the optical-confinement strength of the first-order TE mode (Δn_{1st} in (3.11)) in the MRRs, calculated similar to line 6 in Algorithm 1 when $w_i = w_r = 400$ nm, the same baseline considered for the silicon area, and for $n_{\text{eff}1st}$. Note that Δn_{1st} increases as the waveguide with increases, hence Δn_{1st} at $w_i = w_r = 400$ nm (i.e.,



(a) Q-factor **(b)** 3-dB bandwidth per λ (GHz) **(c)** Silicon area (μm^2)

(d) 2nd-order TE mode confinement **(e)** Total resonant shift ($T_{\Delta\lambda_R}$) (nm) **(f)** Optimal design points

Figure 3.9: Performance of an active MRR designed using the parameters in Table 3.1 and with $g = 200$ nm where both the input and ring waveguide widths change from 350 to 1200 nm. The desired design points are selected and shown with magenta squares and the optimal MRR design region in (f) satisfies all the requirements in (a)–(e). Moreover, the yellow triangle in (f) shows the design point at which the total resonant-wavelength shift is minimum ($T_{\Delta\lambda_R} = 2.8$ nm).

the baseline MRR) is considered in (3.11). The design points selected in Figs. 3.8(d) and 3.9(d) ensure that the optical-confinement strength of the second-order TE mode is always weaker that of the first-order mode in the baseline (i.e., $\Delta n_{2nd} < 0.76$ and $\Delta n_{2nd} < 0.99$ for, respectively, the passive and active MRR in (3.11)).

Figs. 3.8(e) and 3.9(e) show heatmaps for the total resonant-wavelength shift in the passive and active MRR, respectively. As shown in Algorithm 1 and discussed above, a designer can consider a maximum tolerable total resonant-wavelength shift per MRR (T_M), and hence find a set of input and ring waveguide widths that corresponds to $T_{\Delta\lambda_R} \leq T_M$. Considering thermal tuning, state-of-the-art integrated heaters can consume as low as 27.5 mW/FSR [71] for resonance tuning in MRR add-drop filters. Assuming a maximum tuning power budget corresponding to FSR/2 nm per MRR (i.e., $T_M =$ FSR/2 nm) and a similar tuning efficiency in passive and active MRRs, we select the design points in Figs. 3.8(e) and 3.9(e) where the tuning power consumption per MRR is less or equal to FSR/2 nm (i.e., 13.75 mW). This corresponds to the resonant shift of ≈5 and ≈6 nm

for the passive and active MRR, respectively. Note that our assumption here (27.5 mW/FSR and T_M = FSR/2 nm) is considered as an example and it can be simply updated based on the tuning power budget in a system.

Figs. 3.8(f) and 3.9(f) show the search results in Algorithm 1 for the passive and active MRR, respectively. When T_M is considered (see above), the figures show the design points that satisfy the performance constraints discussed in this section for the considered examples of the passive and active MRRs. A designer can choose any design points (i.e., w_i and w_r) highlighted by magenta square in these figures to realize an MRR design with high tolerance to FPVs while achieving specific 3dB bandwidth and Q-factor, preserving silicon-area consumption, and alleviating higher-order mode excitation. When T_M is not considered, the search returns a single w_i and w_r per MRR (yellow triangles in Figs. 3.8(f) and 3.9(f)) that minimizes the total resonant-wavelength shift while satisfying all the constraints discussed in this section. Thanks to the low complexity of our proposed analytical models, the proposed design optimization can be easily integrated into an automated MRR design-space exploration and optimization tool.

3.6 Experimental Analysis of Adiabatic Silicon Photonic Microring Resonators under Process Variations[2]

This section presents a comprehensive experimental analysis of adiabatic MRRs to demonstrate their performance and robustness under FPVs. In addition, we discuss the design and analysis of adiabatic MRRs to facilitate their employment in PICs. Our experimental study includes the fabrication of 268 conventional (i.e., when $w = w'$ in Fig. 3.10(a)) and 289 adiabatic MRRs using electron-beam (E-Beam) lithography, and automatic characterization of all the MRRs to compare and study their tolerance to FPVs, and uniformity and performance in terms of Q-factor, extinction ratio, and free-spectral range. Characterization results show significant uniformity in the response of adiabatic MRRs and across all the performance metrics, compared to conventional MRRs.

[2]A. Mirza, R. E. Gloekler, J. Thompson, S. Pasricha, and M. Nikdsast, "Experimental Analysis of Adiabatic Silicon Photonic Microring Resonators under Process Variations," IEEE Photonics Technology Letters (PTL), 2024.

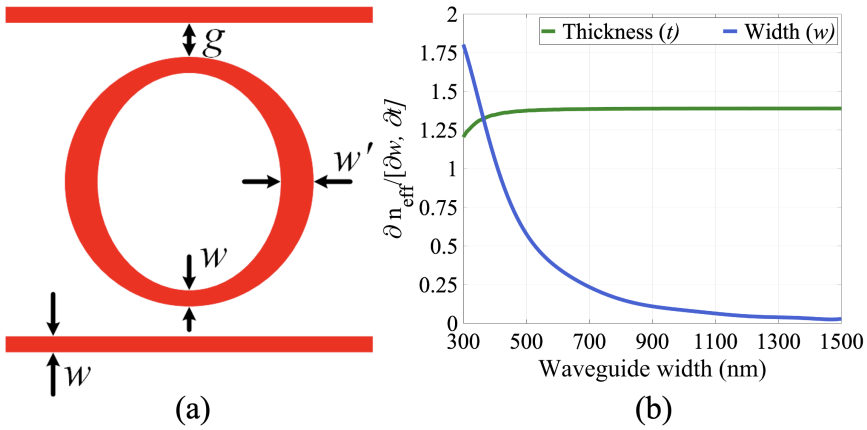


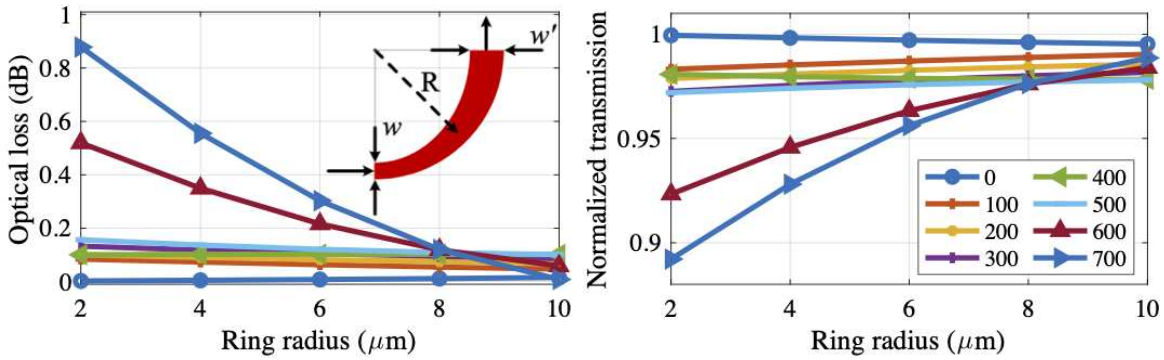
Figure 3.10: (a) Adiabatic MRR with its design parameters ($w' > w$). (b) Rate of changes in the effective index (n_{eff}) w.r.t. the variations in the waveguide width (w) and thickness (t) as the waveguide width increases from 300 nm to 1500 nm (x axis).

3.6.1 Adiabatic MRR Design and Analysis

Fig. 3.10(b) shows that with increasing the waveguide width, the rate of changes in the effective index (n_{eff}) w.r.t. the variations in the waveguide width (w) decreases, while such a rate remains almost the same under waveguide thickness (t) variations [3]. This implies that by increasing the waveguide width in an MRR, one should expect higher tolerance under FPVs in the device. However, uniformly increasing the waveguide width (i.e., increasing $w = w'$ in Fig. 3.10(a) together) in an MRR will result in undesired optical mode distortion and excitation, as indicated in our prior work [3]. This necessitates an adiabatic MRR design using curved, tapered waveguides within the ring (as further discussed later in this section) by gradually increasing the ring waveguide's width from the input (i.e., w) to the center (i.e., w'), as shown in Fig. 3.10(a), to avoid optical mode distortions and excitation of higher order modes in the MRR. Note that in an adiabatic MRR design, the coupling region between the input/drop waveguide and the ring is the same as that in a conventional MRR, all following the same waveguide width (i.e., w, where $w < w'$).

The resonant wavelength (λ_R^a) in an adiabatic MRR can be modeled based on its equivalent effective index (n_{eff}^a) as:

$$\lambda_R^a = \left(\frac{2\pi R}{m}\right) n_{eff}^a, \tag{3.13}$$



(a) Optical power loss **(b)** Transmission efficiency

Figure 3.11: (a) Optical loss and (b) transmission efficiency in a tapered ring with different radii and input–output waveguide widths. Results are based on simulating the curved, tapered quarter ring (inset in (a)) with the radius of R and nonidentical input–output waveguide widths (w and w') using Lumerical FDTD. The legend shows the difference in waveguide width (i.e., 0 nm $\leq |w' - w| \leq$ 700 nm).

where R is the radius of the MRR and m is an integer that denotes the order of the resonance. Also, n^a_{eff} can be approximated by taking the average of effective indices when considering w and w' waveguide widths: i.e., $n^a_{\text{eff}} = \frac{1}{2}\left(n_{\text{eff}}(w) + n_{\text{eff}}(w')\right)$. Accordingly, λ^a_R in an adiabatic MRR is impacted by the critical dimensions of the MRR and its radius, slight variations (e.g., due to FPVs) in which will deviate λ^a_R, which is known as the resonance-wavelength shift ($\delta\lambda^a_R$). In order to model $\delta\lambda^a_R$ in an adiabatic MRR, a compact model can be defined based on:

$$\delta\lambda^a_R = \left(\frac{\lambda^a_R |n^a_{\text{eff}} - n^{a\dagger}_{\text{eff}}|}{n^a_{\text{g}}}\right) + \left(\frac{\lambda^a_R R^{\dagger}}{R}\right)\frac{n^a_{\text{eff}}}{n^a_{\text{g}}}, \tag{3.14}$$

where $n^{a\dagger}_{\text{eff}}$ and $R^{\dagger}$ are the effective index and radius of the MRR that have changed due to FPVs, and n^a_{g} is the ideal group index in the MRR, which can be calculated using n^a_{eff}. According to (3.14) and Fig. 3.10(b), compared to conventional MRRs, adiabatic MRRs should benefit from reduced $\delta\lambda^a_R$ by carefully designing w and w' (see Section 3.6.2).

To realize a robust, adiabatic MRR, one needs to use curved, tapered waveguides within the ring structure (see Fig. 3.10(a)). We simulated different curved, tapered waveguides using Lumerical finite-difference time-domain (FDTD) simulation [72] to analyze fundamental TE-mode transmission efficiency when light traverses a tapered ring with nonidentical input (w) and output (w') waveguide widths (see the inset in Fig. 3.11(a)). Fig. 3.11(a) shows the optical losses in this

Table 3.3: MRR test structure design parameters (see Fig. 3.10).

	Radius (R)	**Gap (g)**	**Width (w)**	**Width (w')**
MRR1	9.935 μm	100 nm	450 nm	450 nm
MRR2	9.935 μm	100 nm	450 nm	820 nm

structure accounting for the bending loss, propagation loss, and mode-mismatch loss due to using different waveguide widths under different ring radii (see the x-axis). The transmission efficiency is shown in Fig. 3.11(b) in which transmission at $|w' - w| = 0$ accounts for losses in a conventional ring. As can be seen in Fig. 3.11, larger radii is required when $|w' - w|$ increases to compensate for inefficiencies in the ring and improve transmission efficiency in the tapered ring structure.

3.6.2 Experimental Results and Discussion

By utilizing the MRR design-space exploration framework developed in our prior work [3]—to which we added resonance-wavelength shift model in adiabatic MRRs (i.e., (3.14)) and results from Fig. 3.8 —we designed two TE-polarized MRR test structures namely MRR1 (conventional MRR) and MRR2 (adiabatic MRR), based on add-drop structures, as shown in Fig. 3.12. The design parameters of each MRR are listed in Table 3.3, optimized in a way to design MRR1 and MRR2 with the same overall size so that they experience similar FPVs while keeping their nominal resonant wavelengths as close as possible (i.e., 1550 nm in MRR1 and 1546 nm in MRR2). Note that the resonant wavelength difference between MRR1 and MRR2 does not hinder our analysis in this chapter. Moreover, 268 identical copies of MRR1 and 289 identical copies of MRR2 were strategically placed across a 10 $\times$ 10 mm chip fabricated using a standard E-Beam multi-project wafer (MPW) process at Applied Nanotools Inc. [73], where each pair of MRR1 and MRR2 were placed as close as possible (see Fig. 3.12(c)) to ensure that they experience similar FPVs.

We tested the chip using an automated testing station from Maple Leaf Photonics that has a capability to test a single die as large as 25 $\times$ 25 mm. In addition to the testing station, the test setup consists mainly of four parts: a laser source, a polarization controller set to TE, a fiber array attached to the station, and a photodetector. The chip stage was fixed and the fiber arm provided automatic movement in the X, Y, Z, yaw, and pitch directions. These movements are critical for

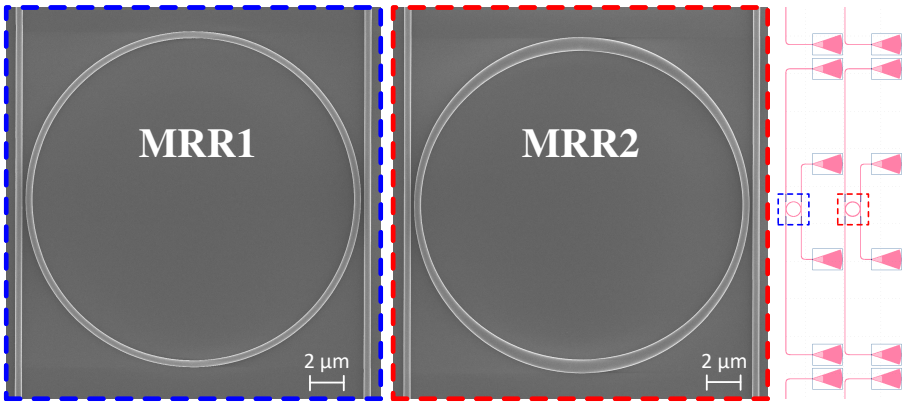
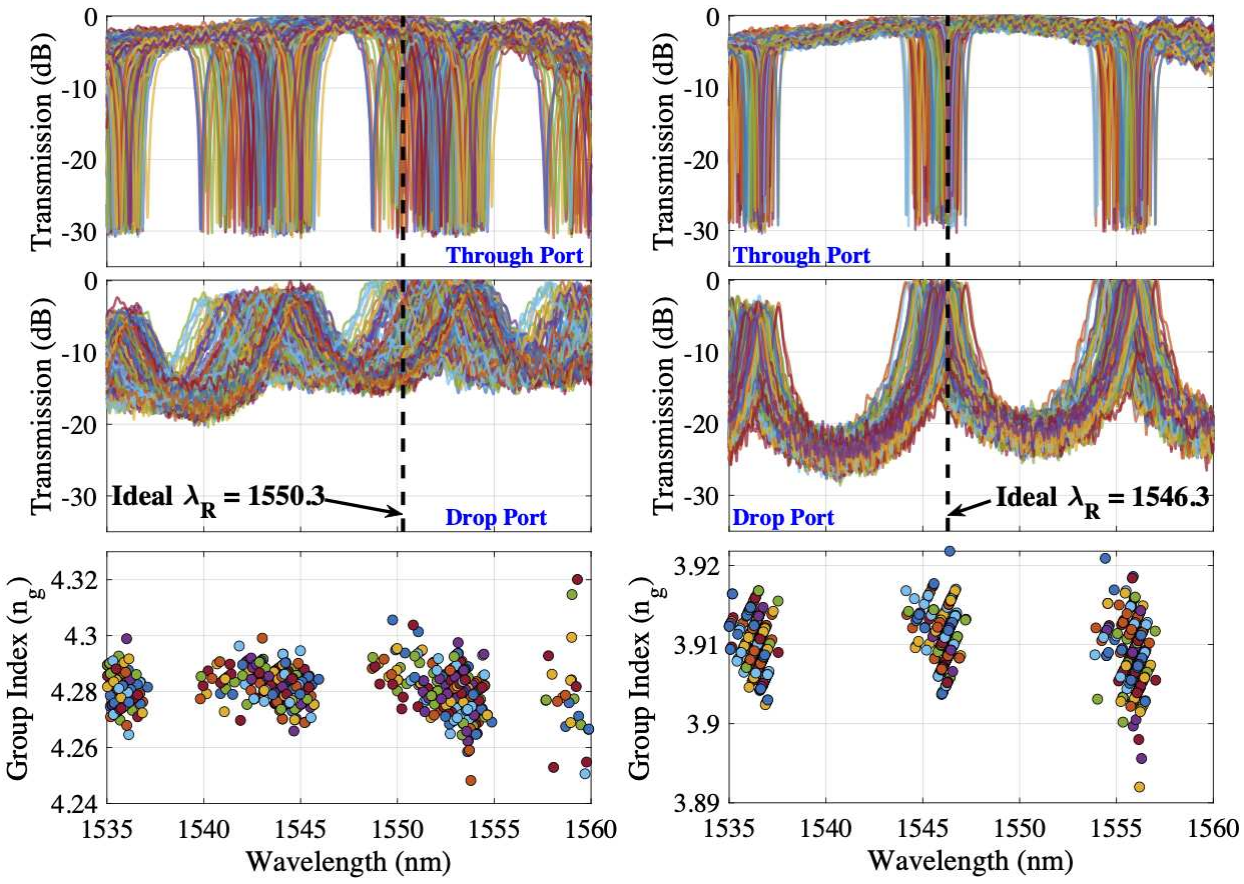


Figure 3.12: Scanning electron microscopic (SEM) images of fabricated (a) conventional and (b) adiabatic MRRs taken from part of the chip. (c) The unit cell of the fabricated MRRs.

efficient alignment of the fiber array with the chip to perform low-loss measurements. Coupling to the GCs on the chip was achieved using a single-mode fiber array with a 127 μm pitch. For precise alignment of the chip with the fiber array, we utilized two digital microscopic cameras integrated into our testing station. The laser source, produced by Agilent HP 8164A, operates within the C-band, spanning a wavelength range of 1460 nm to 1580 nm and offering a resolution of 0.1 pm. Light detection from the chip was performed with the Keysight 81635A dual optical photodetector, which covers a wavelength range from 800 nm to 1650 nm with a power spectrum from $+10$ to -80 dBm. All connections within the setup were established using single-mode fibers. Moreover, the temperature of the chip stage was maintained at 300 K to eliminate the impact of thermal variations during the measurements. The input power was set to 7.5 dBm, for which we experienced a total loss of 25.2 dB in the output. The loss due to grating couplers was 17.6 dB (i.e., 8.8 dB each), and the loss for each device was estimated to be $\approx$1 dB (see Table 3.4). Accordingly, the total unaccounted losses due to the test and alignment was $\approx$6.6 dB.

We begin by comparing the total resonance-wavelength shift between the conventional and adiabatic MRRs. Note that it is crucial to observe the same resonant mode when analyzing the resonant shift, given the fact that the resonance-wavelength shift may exceed the free-spectral range (FSR) in an MRR. To verify we are examining the correct resonant mode, we compute the group index (n_{g}) for every resonant wavelength by using $FSR = \frac{\lambda_R^2}{n_{\mathrm{g}} \cdot L}$, in which λ_R denotes



(a) Conventional MRR (MRR1) **(b)** Adiabatic MRR (MRR2)

Figure 3.13: Through-port and drop-port response in (a) conventional and (b) adiabatic MRRs. Figures show the resonant wavelength versus corresponding group index (n_g) to find responses that belong to the same resonant mode. Black-dotted lines show the ideal (nominal) resonant wavelengths.

the resonant wavelength and L is the MRR's round trip length. Note that FSR can be directly estimated from measured through-/drop-port response. Based on this method, initially proposed in [45], the MRR responses belonging to the same resonant mode will have close group indices, hence can be identified by observing the resonant wavelength versus group index plot.

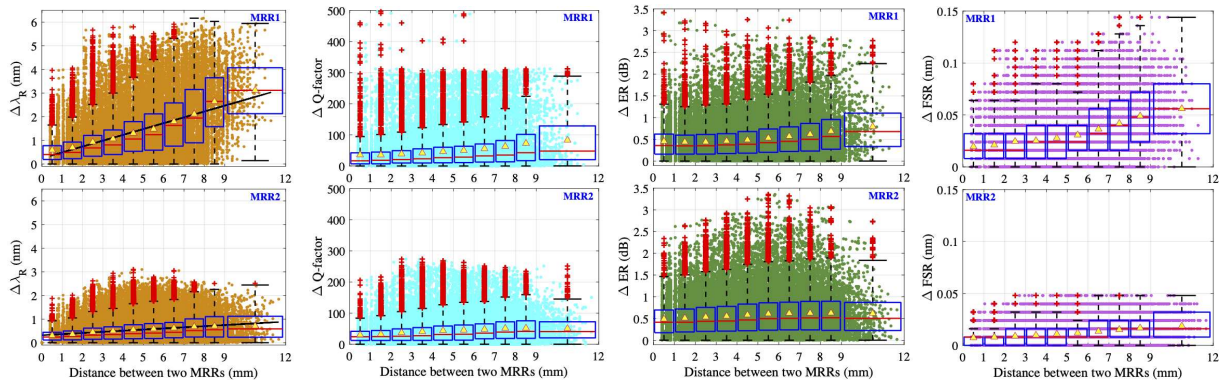
Figs. 3.13(a) and 3.13(b) show the measured through- and drop-port responses for the conventional (MRR1) and adiabatic (MRR2) MRRs, respectively. For each captured resonant wavelength, we also calculated the corresponding group index, the result for which is shown in the bottom plots in Fig. 3.13 where clusters of data points show resonant wavelengths that correspond to the same resonant mode. Considering this analysis, our observation is that the resonance-wavelength shift measured does not surpass the FSR in any MRR.

Table 3.4: Characterized device performance (Avg.: Average, SD: Standard Deviation, λ_R: Resonant Wavelength, ER: Extinction Ratio).

	MRR1 (Conventional)		**MRR2 (Adiabatic)**	
	Through	**Drop**	**Through**	**Drop**
Avg. λ_R	1552.8 nm		1546.1 nm	
SD λ_R ($\propto \delta\lambda_R$)	1.3 nm		0.5 nm	
Avg. Q-factor	3567	590	10067	790
Avg. ER	27.7 dB	12.8 dB	25 dB	21.8 dB
Crosstalk	-22.6 dB	-12.3 dB	-21.2 dB	-19.6 dB
Drop loss	1.3 dB		0.8 dB	

Considering the through- and drop-port responses in Fig. 3.13, the adiabatic MRRs' responses align closer to the *ideal response* at 1546.3 nm by up to 70% on average, compared to the conventional MRRs whose *ideal response* is at 1550.3 nm. For a more complete experimental comparison between MRR1 and MRR2, Table 3.4 shows our measurement results across multiple performance metrics. The average responses for each MRR (close to 1550 nm resonant mode) calculated considering all the responses captured are 1552.8 nm and 1546.1 nm for MRR1 and MRR2, respectively, while the corresponding standard deviations (proportional to the measured resonance-wavelength shifts—$\delta\lambda_R$) for MRR1 and MRR2 are, respectively, 1.3 nm and 0.5 nm, showing significant uniformity in the responses of adiabatic MRRs compared to the conventional ones (see Avg. λ_R and SD λ_R in Table 3.4). Table II also compares experimentally measured Q-factor, extinction ratio (ER), crosstalk, and drop loss between MRR1 and MRR2. The drop loss in each device is estimated based on the Q-factor and attenuation-factor analysis in each MRR. Also, the crosstalk was measured based on the minimum and maximum transmissions observed at the through and drop ports. As can be seen, our experimental results show that the adiabatic design results in a better performance in general.

To further analyze the resonance-wavelength shift across all the devices on the chip, we compared the resonance-wavelength difference ($\Delta\lambda_R$) between each pair of MRRs—of the same type: either conventional or adiabatic—positioned at different locations on the chip. The result of this analysis in shown in Fig. 3.14(a). As can be seen, both conventional and adiabatic MRRs positioned in proximity exhibit more similar responses (i.e., smaller $\Delta\lambda_R$) compared to those spaced



(a) Resonance-wavelength difference **(b)** Q-factor difference **(c)** Extinction ratio (ER) difference **(d)** FSR difference

Figure 3.14: Comparison of various performance metrics represented with different colors between each pair of MRRs placed at different positions on the chip (x axis). Yellow triangles show the average within each box, to which a linear fit (shown as black line) is depicted in (a). For each box, the red plus signs represent the outliers that fall outside the typical range and the red lines show the median within each box.

farther apart. In addition, $\Delta\lambda_R$ escalates as the distance between the MRRs increases. This observation confirms a strong linear correlation (see the black lines—fit to the average in each boxplot—in Fig. 3.14(a)) between MRR placement and resonance uniformity, as showed also by [45]. Another important observation from Fig. 3.14(a) is that, with increasing distance between each pair of MRR (x-axis in the figure), the average $\Delta\lambda_R$ in MRR2 (adiabatic) is much smaller than that in MRR1 (conventional), exhibiting 50% reduction in $\Delta\lambda_R$ on average. More importantly, the resonance-wavelength difference among adiabatic MRRs across the entire chip, even when two of them are placed as far as $>$9 mm, stays within $<$1 nm.

We also analyzed the difference in Q-factor, ER, and FSR for each pair of MRRs placed at different locations on the chip, results of which are shown in Figs. 3.14(b), 3.14(c), and 3.14(d), respectively. The Q-factor was estimated by performing a Lorentzian fit to the captured responses and analyzing the full-width half-maximum (FWHM), and correspondingly the Q-factor. Compared to the conventional MRRs, we observe that for all Q-factor, ER, and FSR, the adiabatic MRRs' performance is more uniform and almost independent of the MRRs' placement, when considering each MRR pair and their placement distance (x-axis in the figures).

3.7 Conclusion

With various advantages, silicon photonic microring resonators are often presented as the workhorse of emerging optical interconnects and photonic integrated circuits. In this chapter, we presented a comprehensive design-space exploration and optimization of MRRs under different fabrication-process variations. We consider variations in the waveguide width, SOI thickness, slab thickness (etching depth), and MRR radius in both passive and active MRRs. We present computationally efficient analytical models tailored to capture the impact of physical-level variations on MRR device-level performance. Leveraging these models, we exhaustively explore the design space of MRRs to enable an optimal MRR design with high tolerance to FPVs and desired Quality factor and 3dB bandwidth performance, all of which can be determined during design-time. We further described our experimental analysis that compares fabricated conventional and adiabatic MRR designs. Our findings consistently demonstrate the superior attributes of adiabatic MRRs over their conventional counterparts which is a biproduct of efficient analytical models. Notably, adiabatic MRRs exhibit a remarkable 70% alignment with ideal resonant wavelengths for which they are designed. Moreover, compared to conventional MRRs, adiabatic MRRs show significant uniformity in terms of frequency response, Q-factor, extinction ratio, and FSR. Such significant uniformity in adiabatic MRRs make them promising designs for many PIC applications, both in Datacom and computation, where high inter-device matching is required. Furthermore, the high uniformity helps simplify device placement and routing, as well as tuning by allowing for collectively tuning all of MRRs (as compared to individual tuning), hence reducing tuning complexity and overall power overhead. In the next chapter, we shall dwell into a few applications where optimized rings can make a huge impact on the performance of a system.

Chapter 4

Applications of FPV-aware Optimized Photonic Devices

4.1 Introduction

Dense wavelength-division multiplexing (DWDM) is a common solution to boost the bandwidth performance in silicon photonic (SiPh) integrated circuits, where microring resonator (MRR) filters have been widely employed for optical channel (wavelength) demultiplexing. Such densely integrated circuits are possible due to the high refractive-index contrast in silicon-on-insulator (SOI) platforms. However, this same contrast makes SiPh devices susceptible to inevitable fabrication-process variations (FPVs), where nanometer-scale deviations in critical dimensions can considerably impact device functionality. For example, a mere 1 nm variation in an MRR filter's width or thickness can shift its wavelength response by several nanometers. Such a wavelength drift necessitates active tuning (compensation) for inter-device matching in today's DWDM systems with channel spacing typically less than 1 nm. Active compensation, however, is complex and resource-intensive, demanding efficient design solutions to mitigate the effects of PVs at the design stage in a fabless ecosystem. For the first application in Section 4.2, we propose a design optimization framework to improve inter-device matching (i.e., channel spacing accuracy) in MRR-based wavelength demultiplexers under different PVs. In particular, our optimization framework considers actual layout information and fundamental variations in SOI thickness and waveguide width present on different length scales (i.e., correlations in variations). By minimizing deviations in the channel spacing in MRR-based demultiplexers, we can compensate for wavelength shifts by collectively tuning all the MRRs, hence simplifying the tuning and improving its efficiency.

Silicon photonics is a promising technology to enable ultra-high bandwidth, low-latency, and energy-efficient communication solutions. CMOS-compatible photonic interconnects have already

replaced metallic ones for light-speed data transmission at almost every level of computing, and are now actively being considered for chip-scale integration. Remarkably, it is also possible to use optical components to perform computation, e.g., matrix-vector multiplication. Thus, it is now possible to conceive of a new class of DNN accelerators that employ photonic interconnects and photonic integrated circuits (PICs) for low-latency and energy-efficient data transport and computation. The operational bandwidth of such photonic accelerators can approach the photodetection rate (hundreds of GHz), which is significantly higher than electronic systems today that operate at few GHz. For the second application in Section 4.3, we introduce *CrossLight*, a novel silicon photonic neural network accelerator that addresses the challenges highlighted above through a cross-layer design approach. By cross-layer, we refer to the design paradigm that involves considering multiple layers in the hardware-software design stack together, for a more holistic optimization effort.

Domain-specific neural network accelerators have garnered attention due to their superior energy efficiency and inference performance compared to CPUs and GPUs. These accelerators are particularly well-suited for resource-constrained embedded systems. However, implementing complex neural network models on these accelerators still poses challenges in terms of energy and memory consumption, coupled with high inference time overhead. Binarized neural networks (BNNs), employing single-bit weights, offer an efficient solution for implementing and deploying neural network models on accelerators. In Section 4.4, we introduce *ROBIN*, a novel optical-domain BNN accelerator, which intelligently integrates heterogeneous microring resonator optical devices with complementary capabilities to efficiently implement the key functionalities in BNNs. We perform detailed FPV analyses at the optical device level, explore efficient corrective tuning for these devices, and also integrate circuit-level optimization to counter thermal variations. As a result, our proposed *ROBIN* architecture possesses the desirable traits of being robust, energy-efficient, low latency, and high throughput, when executing BNN models. For the final application, in Section 4.5 we, for the first time, model and explore the impact of FPVs in the waveguide width and silicon-on-insulator (SOI) thickness in coherent SPNNs that use Mach–Zehnder Inter-

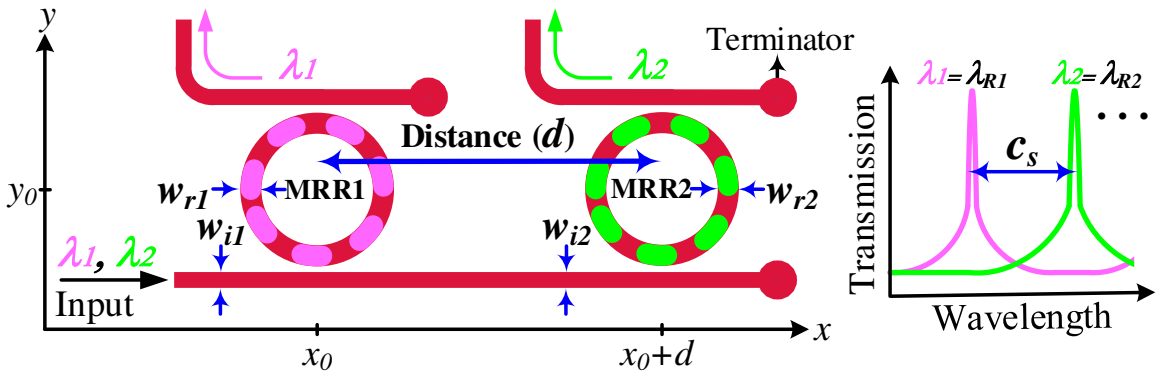


Figure 4.1: A two-channel passive wavelength-selective MRR-based demulitplexer with two MRRs placed at a distance d and with a channel spacing c_s, designed based on Table 2.1. Here, we assume $\lambda_{R1} = 1550$ nm and $\lambda_{R2} = 1553$ nm (radius in MRR2 is slightly different). Therefore, $c_s = 3$ nm.

Table 4.1: Different parameters used to generate FPV maps

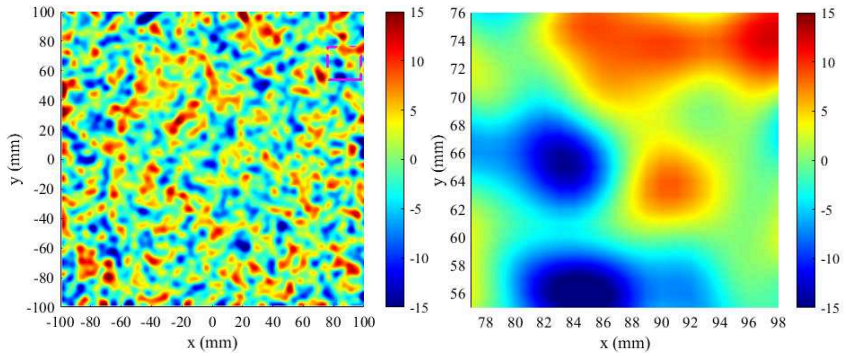
Design Parameter	Correlation Length	Standard Deviation
Waveguide width	$l_w = 4.5$ mm	Center: $\sigma_w = 4.2$ nm Edges: $\sigma_w = 5.5$ nm
SOI thickness	$l_t = 4.5$ mm	Center: $\sigma_t = 0.7$ nm Edges: $\sigma_t = 2.2$ nm
MRR radius	$l_R = 4.5$ mm	Center: $\sigma_R = 0.5$ nm Edges: $\sigma_R = 1$ nm

ferometers (MZIs). Leveraging such models, we propose a novel variation-aware, design-time optimization solution to improve MZI tolerance to different FPVs in SPNNs.

4.2 A Wavelength-Selective MRR-Based Demultiplexer[1]

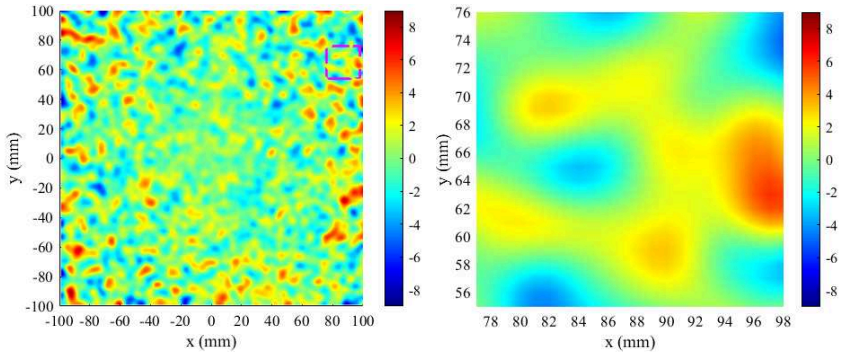
Leveraging the proposed MRR design-space exploration and optimization in Sections 3.4 and 3.5, in this section we improve the inter-device matching (i.e., channel-spacing accuracy) in a study of a passive wavelength-selective MRR-based demultiplexer, shown in Fig. 4.1, under different FPVs. In addition, we develop virtual FPV wafer maps to account for actual layout information and fundamental variations in the waveguide width, SOI thickness, and radius, which are present on different length scales (i.e., correlations in variations).

[1]A. Mirza, F. Sunny, P. Walsh, K. Hassan, S. Pasricha, and M. Nikdast, "Silicon photonic microring resonators: A comprehensive design-space exploration and optimization under fabrication-process variations," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), vol. 41, no. 10, pp. 3359–3372, 2022.



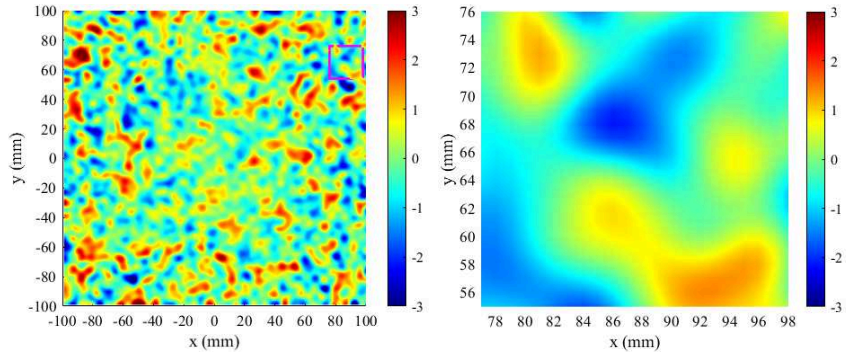
(a) Waveguide width, wafer

(b) Waveguide width, die



(c) SOI thickness, wafer

(d) SOI thickness, die



(e) MRR radius, wafer

(f) MRR radius, die

Figure 4.2: Virtual FPV wafer maps ((a), (c), and (e)) and interpolated die maps ((b), (d), and (f)) that are correlated and mimic radial-variation effects. The maps are generated using the parameters in Table 4.1 with a mean of zero.

We start by developing FPV wafer maps using a similar method proposed in [33] where an uncorrelated random distribution map with specific mean (μ) and standard deviation (σ) is first generated. Then, we convolve the resulting map with a Gaussian filter and specific correlation length (l) to obtain correlated FPV wafer maps. Moreover, we further enhance the correlated FPV maps by incorporating radial-variation effects: i.e., the non-uniformity increases as moving from the wafer center to the wafer edges, as reported also in other works [53, 55]; thus the wafer map center should have the least variations. To capture radial-variation effects, we first characterized the waveguide width and SOI thickness variations (i.e., σ_w and σ_t) at the center of several 200 mm wafers, and then repeated the same at multiple points while moving towards the edges of the wafers. The standard deviations were then averaged over the points within the same distance to the wafer center. For instance, Table 4.1 shows the standard deviations averaged at the center and edges of several 200 mm wafers that we characterized in collaboration with CEA-Leti (σ_w and σ_t only). Accordingly, we enhance our virtual wafer-map models with a variable standard deviation that increases almost linearly as moving from the center towards the wafer edges.

Leveraging the aforementioned method, we develop waveguide width, SOI thickness, and MRR radius virtual FPV maps with means of zero ($\mu_w = \mu_t = \mu_R = 0$) and correlation lengths (l_w, l_t, and l_R) and standard deviations (σ_w, σ_t, and σ_R) listed in Table 4.1. In this table, $\sigma_{w,t}$ are analyzed through experimentally characterizing several 200 mm wafers at CEA-Leti, $l_{w,t}$ are taken from [33], and σ_R and l_R are considered as an example. Moreover, the table only shows the $\sigma_{t,w,R}$ at the wafer center and edges. Figs. 4.2(a), 4.2(c), and 4.2(e) show the resulting waveguide width, SOI thickness, and radius correlated wafer maps, respectively. Also, from each wafer map, we select a die with a size of 22×22 mm^2, which is then interpolated as shown in Figs. 4.2(b), 4.2(d), and 4.2(f).

The channel spacing (c_s) in the two-channel MRR-based demultiplexer in Fig. 4.1 can be defined as the optical frequency space between the two MRRs' consecutive resonant wavelengths which can be given by $c_s = |\lambda_{R2} - \lambda_{R1}|$, where $\lambda_{R2} > \lambda_{R1}$. As shown by Fig. 4.2, FPVs are present on different length scales and are correlated, hence we consider not only the MRR layout

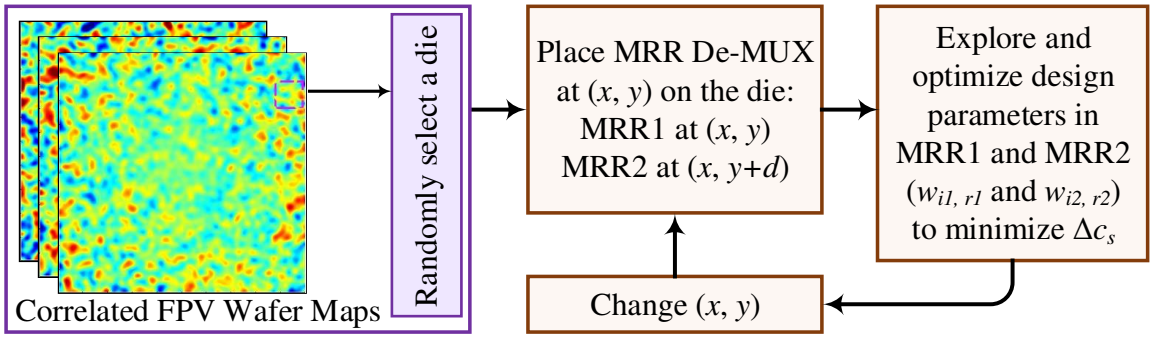


Figure 4.3: An overview of the channel-spacing accuracy optimization for the MRR demultiplexer in Fig. 4.1.

design parameters that are affected by different variations but also the positioning of each MRR on a die—FPVs on each MRR can be different—and the distance between the two MRRs (d in Fig. 4.1). Employing (3.9), we can model the deviated channel spacing (c_s') in the two-channel MRR-based demultiplexer under different FPVs as:

$$c_s' = c_s + \Delta c_s, \tag{4.1}$$

$$= c_s + \left| T_{\Delta\lambda_{R2}}(x+d, y) - T_{\Delta\lambda_{R1}}(x, y) \right|, \tag{4.2}$$

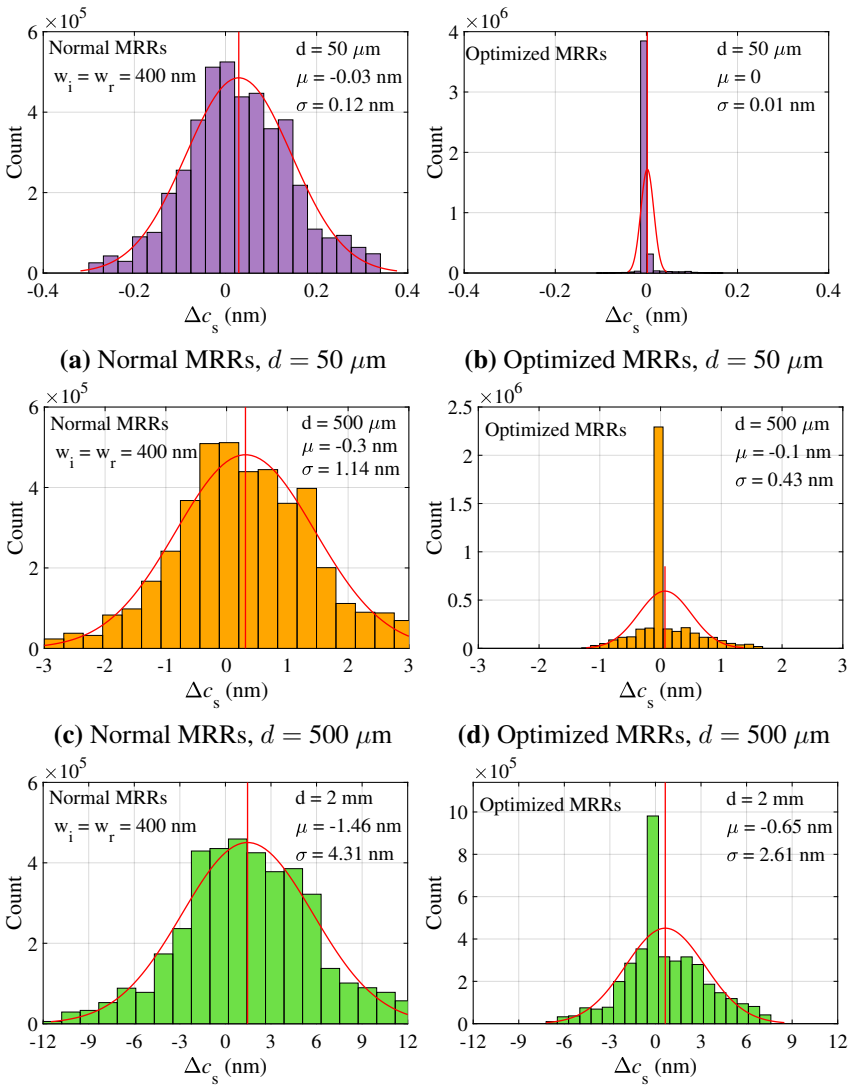
where Δc_s denotes variations in the channel spacing. Also, $T\Delta\lambda_{Ri}(x, y)$ is the total resonant-wavelength shift in MRRi located at position (x, y) on a die. We assume the FPVs on the input/drop and ring waveguides in the same MRR to be the same, but FPVs can be different in the two MRRs.

The channel-spacing accuracy under different FPVs can be improved by minimizing the channel-spacing variations (i.e., Δc_s in (4.1)). Therefore, an optimization search similar to the one in (3.12) can be formed to design an MRR-based demultiplexer with high tolerance to FPVs, and hence high channel-spacing accuracy. Leveraging our proposed MRR design optimization discussed in Section 3.5, this can be achieved by exploring and optimizing the input/drop and ring waveguide widths in MRR1 and MRR2 (i.e., $w_{i1/i2}$ and $w_{r1/r2}$ in Fig. 4.1) while considering the specific FPV profile experienced by each MRR. While we focus on the channel-spacing accuracy in this section,

one can easily add other objectives (e.g., 3dB bandwidth and Q-factor) to the design-optimization problem, similar to the one proposed and addressed in Section 3.5.

Considering (4.2), the channel-spacing accuracy can be improved by applying $T_{\Delta\lambda_{R1}} \rightarrow 0$ and $T_{\Delta\lambda_{R2}} \rightarrow 0$, or $T\Delta\lambda_{R2/1} \rightarrow T\Delta\lambda_{R1/2}$, both minimizing Δc_s. Employing the FPV die maps in Fig. 4.2 (22×22 mm^2) and our MRR design optimization in Section 3.5, we optimize the design parameters in MRR1 and MRR2—i.e., $w_{i1/r1}$ and $w_{i2/r2}$; other parameters are based on those in Table 3.1—while uniformly positioning these MRRs at every location on the die, and analyzing channel-spacing variations (Δc_s) in the demultiplexer (see Fig. 4.3). We consider different scenarios where the MRRs are in proximity ($d = 50$ μm) and when they are placed apart on the die ($d = 500$ μm and $d = 2$ mm). Based on the optimal design region specified in Fig. 3.8(f), we set the input waveguides to be 450 nm wide in both MRR1 and MRR2 ($w_{i1,i2} = 450$ nm), and explore ring waveguide width ($w_{r1,r2}$) from 570 to 820 nm to minimize channel-spacing variations. Therefore, the resulting MRR designs will satisfy all the performance requirements discussed in Fig. 3.8.

Fig. 4.4 indicates the channel-spacing variations (Δc_s) in the demultiplexer with normal (unoptimized) (Figs. 4.4(a), 4.4(c), and 4.4(e)) and optimized (Figs. 4.4(b), 4.4(d), and 4.4(f)) MRRs when $d = 50$ μm, 500 μm, and 2 mm, and for 4×10^6 design samples. In the normal MRR design, $w_{i1,r1} = w_{i2,r2} = 400$ nm. As can be seen, without MRR optimization, the channel-spacing variation is high and further increases as d increases (see Figs. 4.4(a), 4.4(c), and 4.4(e)). As shown in Figs. 4.4(b), 4.4(d), and 4.4(f), our MRR design optimization helps maintain the channel-spacing accuracy in ≈98% of design samples within 0.05 nm when $d = 50$ μm, ≈80% of design samples within 0.5 nm when $d = 500$ μm, and ≈60% of design samples within 2 nm when $d = 2$ mm. This clearly shows the effectiveness of our MRR design optimization. When $d = 50$ μm, both MRRs experience variations on a smaller length scale, hence intuitively inter-device matching is already higher and the optimization chooses MRRs of mostly equal width to improve channel-spacing accuracy. However, when d increases to 500 μm and 2 mm, MRRs experience much different variations and on a larger length scale, hence inter-device matching is more challenging. Never-



(a) Normal MRRs, $d = 50$ μm **(b)** Optimized MRRs, $d = 50$ μm

(c) Normal MRRs, $d = 500$ μm **(d)** Optimized MRRs, $d = 500$ μm

(e) Normal MRRs, $d = 2$ mm **(f)** Optimized MRRs, $d = 2$ mm

Figure 4.4: Statistical analysis of channel-spacing variations (Δc_s) in the demultiplexer in Fig. 4.1 while considering normal and optimized MRRs and different distances (d) between the two MRRs. In the normal MRR design, $w_{i1,r1} = w_{i2,r2} = 400$ nm. The optimized MRR design is based on the procedure in Fig. 4.3 with $w_{i1,i2} = 450$ nm and $w_{r1,r2} \in [570, 820]$ nm. The legends show the mean (μ) and standard deviation (σ) of the normal-distribution fit of each histogram. The nominal channel spacing in Fig. 4.1 is 3 nm.

theless, the optimization succeeds to efficiently improve the channel-spacing accuracy even when d is large. Note that the optimization results in this section are in agreement with our design-space exploration results in Section 3.4. With such an optimized channel-spacing accuracy in MRR-based demultiplexers—enabled by our proposed MRR design optimization—one can compensate for wavelength shifts through collectively tuning all the MRRs, hence simplifying the circuit tuning and enhances its efficiency. The results presented in this chapter show the promise of our proposed design-space exploration and optimization to improve MRR robustness in optical interconnects and emerging noncoherent artificial intelligence (AI) accelerators [74].

4.3 CrossLight: A Cross-Layer Optimized Silicon Photonic Neural Network

Many emerging applications such as self-driving cars, autonomous robotics, fake news detection, pandemic growth and trend prediction, and real-time language translation are increasingly being powered by sophisticated machine learning models. With researchers creating deeper and more complex deep neural network (DNN) architectures, including multi-layer perceptron (MLP) and convolution neural network (CNN) architectures, the underlying hardware platform must consistently deliver better performance while satisfying strict power dissipation limits. This endeavor to achieve higher performance-per-watt has driven hardware architects to design custom accelerators for deep learning, e.g., Google's TPU [75] and Intel's Movidius [76], with much higher performance-per-watt than conventional CPUs and GPUs.

Unfortunately, electronic accelerator architectures face fundamental limits in the post Moore's law era where processing capabilities are no longer improving as they did over the past several decades [77]. In particular, moving data electronically on metallic wires in these accelerators creates a major bandwidth and energy bottleneck [78]. Silicon photonics is a promising technology to enable ultra-high bandwidth, low-latency, and energy-efficient communication solutions [79]. CMOS-compatible photonic interconnects have already replaced metallic ones for light-speed data

transmission at almost every level of computing, and are now actively being considered for chip-scale integration [42].

Remarkably, it is also possible to use optical components to perform computation, e.g., matrix-vector multiplication [80]. Thus, it is now possible to conceive of a new class of DNN accelerators that employ photonic interconnects and photonic integrated circuits (PICs) built with on-chip waveguides, electro-optic modulators, photodetectors, and lasers for low-latency and energy-efficient optical domain data transport and computation. Not only can such photonics-based accelerators address the fan-in and fan-out problems with linear algebra processors, but their operational bandwidth can approach the photodetection rate (typically in the hundreds of GHz), which is orders of magnitude higher than electronic systems today that operate at a clock rate of a few GHz [37].

Despite the above benefits, a number of obstacles must be overcome before viable photonic DNN accelerators can be realized. Fabrication-process and thermal variations can adversely impact the robustness of photonic accelerator designs by introducing undesirable crosstalk noise, optical phase shifts, resonance drifts, tuning overheads, and photo-detection current mismatches. For example, experimental studies have shown that micro-ring resonator (MRR) devices used in chip-scale photonic interconnects can experience significant resonant drifts (e.g., 9 nm reported in [20]) within a wafer due to process variations. This matters because even a 0.25 nm drift can cause the bit-error-rate (BER) of photonic data traversal to degrade from 10-12 to 10-6. Moreover, thermal crosstalk in silicon photonic devices such as MRRs can limit the achievable precision (i.e., resolution) of weight and bias parameters to a few bits, which can significantly reduce DNN model accuracy. Common tuning circuits that rely on thermo-optic phase-change effects to control photonic devices, e.g., when imprinting activations or weights on optical signals, also place a limit on the achievable throughput and parallelism in photonic accelerators. Lastly, at the architecture level, there is a need for a scalable, adaptive, and low-cost computation and communication fabric that can handle the demands of diverse MLP and CNN models.

In this section of the chapter, we introduce *CrossLight*, novel silicon photonic neural network accelerator that addresses the challenges highlighted above through a cross-layer design approach.

By cross-layer, we refer to the design paradigm that involves considering multiple layers in the hardware-software design stack together, for a more holistic optimization of the photonic accelerator. *CrossLight* involves device-level engineering for resilience to fabrication-process variations and thermal crosstalk, circuit-level tuning enhancements for inference latency reduction, and an optimized architecture-level design that also integrates the device- and circuit-level improvements to enable higher resolution, better energy-efficiency, and improved throughput compared to prior efforts on photonic accelerator design. Our novel contributions in this work include:

- Improved silicon photonic device designs that we fabricated to make our architecture more resilient to fabrication-process variations;
- An enhanced tuning circuit to simultaneously support large thermal-induced resonance shifts and high-speed, low-loss device tuning;
- Consideration of thermal crosstalk mitigation methods to improve the weight resolution achievable by our architecture;
- Improved wavelength reuse and use of matrix decomposition at the architecture-level to increase throughput and energy-efficiency;
- A comprehensive comparison with state-of-the-art accelerators that shows the efficacy of our cross-layer optimized solution.

4.3.1 Background and Related Work

Silicon-photonics based DNN accelerator architectures represent an emerging paradigm that can immensely benefit the landscape of deep learning hardware design [18, 81–84]. A photonic neuron in these architectures is analogous to an artificial neuron and consists of three components: a weighting, a summing, and a nonlinear unit. Noncoherent photonic accelerators, such as [18, 82, 83], typically employ the Broadcast and Weight (B&W) protocol [81] to manipulate optical signal power for setting and updating weights and activations. The B&W protocol is an analog networking protocol that uses wavelength-division multiplexing (WDM), photonic multiplexors,

and photodetectors to combine outputs from photonic neurons in a layer. Coherent photonic accelerators, such as [37, 84], manipulate the electrical field amplitude rather than signal power and typically use only a single wavelength. Weighting occurs with electrical field amplitude attenuation proportional to the weight value, and phase modulation that is proportional to the sign of the weight. The weighted signals are then coherently accumulated with cascaded Y-junction combiners. For both types of accelerators, non-linearity can be implemented with devices such as electro-absorption modulators [37].

Due to the scalability, phase encoding noise, and phase error accumulation limitations of coherent accelerators [85], there is growing interest in designing efficient noncoherent photonic accelerators. In particular, the authors of DEAP-CNN [82] have described a noncoherent neural network accelerator that implements the entirety of the CNN layers using connected convolution units. In these units, the tuned MRRs assume the kernel values by using phase tuning to manipulate the energy in their resonant wavelengths. Holylight [83] is another noncoherent architecture that uses microdisks (instead of MRRs) for its lower area and power consumption. It utilizes a "whispering gallery mode" resonance for microdisk operation, which unfortunately is inherently lossy due to a phenomenon called tunneling ray attenuation [86]. More generally, these noncoherent architectures suffer from susceptibility to process variations and thermal crosstalk, which are not addressed in these architectures. Microsecond-granularity thermo-optic tuning latencies further reduce the speed and efficiency of optical computing [71]. We address these shortcomings as part of our proposed cross-layer optimized noncoherent photonic accelerator architecture in this work.

4.3.2 Noncoherent Photonic Computation Overview

As mentioned earlier, noncoherent photonic accelerators typically utilize the Broadcast and Weight (B&W) photonic neuron configuration with multiple wavelengths. Fig. 4.5 shows an example of this B&W configuration with n neurons in a layer where the colored-dotted box represents a single neuron. Each input to a neuron is imprinted onto a unique wavelength (λ_i) emitted by a laser

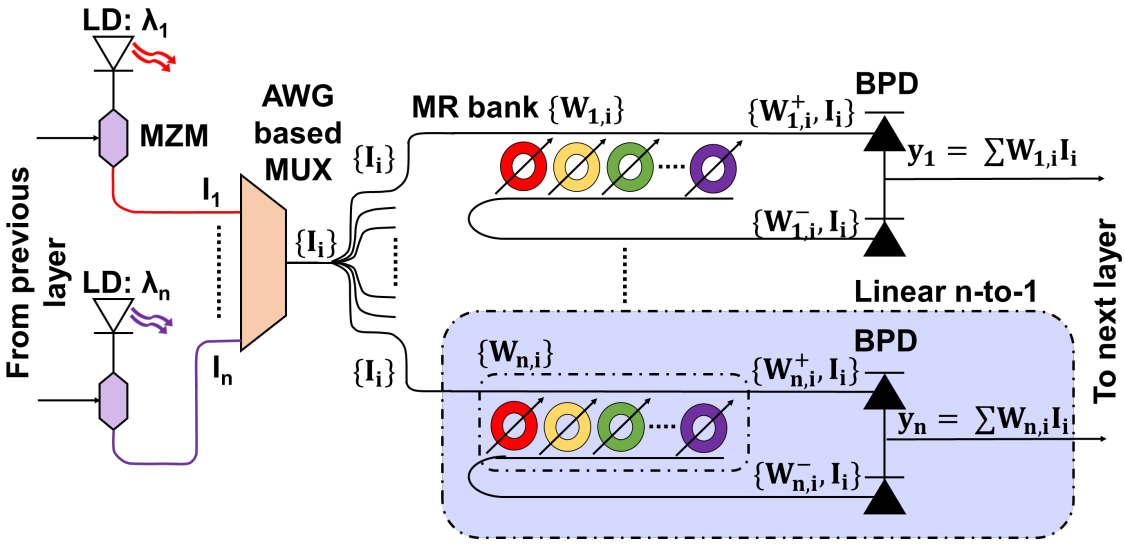


Figure 4.5: Noncoherent Broadcast-and-weight (B&W) based photonic neuron.

diode (LD) using a Mach–Zehnder modulator (MZM). The wavelengths are multiplexed (MUXed) into a single waveguide using arrayed waveguide grating (AWG), and split into n branches that are each weighted with a micro-ring resonator (MRR) bank that alters optical signal power proportional to weight values. A balanced photodetector (BPD) performs summation across positive and negative weight arms at each branch. Optoelectronic devices such as electro-absorption modulators (not shown for brevity) introduce non-linearity after the multiplication and summation operations.

MRRs are the fundamental components that impact the efficiency of this configuration. Weights (and biases) are altered by tuning MRRs so that the losses experienced by wavelengths—on which activations have been imprinted—can be modified to realize matrix-vector multiplication. MRR-weight banks have groups of these tunable MRRs, each of which can be tuned to drain energy from a specific resonant wavelength so that the intensity of the wavelength reflects a specific value (after it has passed near the MRR). As an example of performing computation in the optical domain, consider the case where an activation value of 0.8 must be weighted by a value of 0.5 as part of a matrix-vector multiplication in a DNN model inference phase. Let us assume that the red wavelength (λ_1) is imprinted with the activation value of 0.8 by using the MZM in Fig. 4.5 (alternatively, MRRs can be used for the same goal, where an MRR will be tuned in such a way that 20% of the input optical signal intensity is dropped as the wave traverses the MRR). When λ_1 passes

through an MRR bank, e.g., the one in the dotted-blue box in Fig. 4.5, the MRR in resonance with λ_1 can be tuned to drop 50% of the input signal intensity. Thus, as λ_1 passes this MRR, we will obtain 50% of the input intensity at the through port, which is 0.4 (=0.8$\times$0.5). The BPD shown in Fig. 4.5 then converts the optical signal intensity from that wavelength (and other wavelengths) into an electrical signal that represents an accumulated single value.

An MRR is essentially an on-chip resonator which is said to be in resonance when an optical wavelength on the input port matches with the resonant wavelength of the MRR, generating a Lorentzian-shaped signal at the through port. Fig. 4.6 shows an example of an all-pass MRR and its output optical spectrum. The extinction ratio (ER) and free-spectral range (FSR) are two primary characteristics of an MRR. These depend on several physical properties in the MRR, including its width, thickness, radius, and the gap between the input and ring waveguide [12]. Changing any of these properties changes the effective index (n_{eff}) of the MRR, which in turn causes a change in the output optical spectrum. For reliable operation of MRRs, it is crucial to maintain the central wavelength at the output optical spectrum. However, MRRs are sensitive to fabrication-process variations (FPVs) and variations in surrounding temperature. These cause the central wavelength of the MRR to deviate from its original position, causing a drift in the MRR resonant wavelength ($\Delta\lambda_{MRR}$) [46]. Such a drift (due to FPV or thermal variations) can be compensated using thermo-optic (TO) or electro-optic (EO) tuning mechanisms. Both of these have their own advantages and disadvantages. EO tuning is faster ($\approx$ns range) and consumes lower power ($\approx$4 μW/nm) but with a smaller tuning range [48]. In contrast, TO tuning has a larger tunability range, but consumes higher power ($\approx$27 mW/FSR) and has higher (μs range) latency [71].

A large number of MRRs must be used at the architecture-level to support complex MLP and CNN model executions. As the number of MRRs increase, so does the length of the waveguide which hosts the banks. Unfortunately, this leads to an increase in the total optical signal propagation, modulation, and through losses experienced, which in turn increases the laser power required to drive the optical signals through the weight banks, so that they can be detected error-free at the photodetector. An excessive number of parallel arms with MRR weight banks (the dotted box

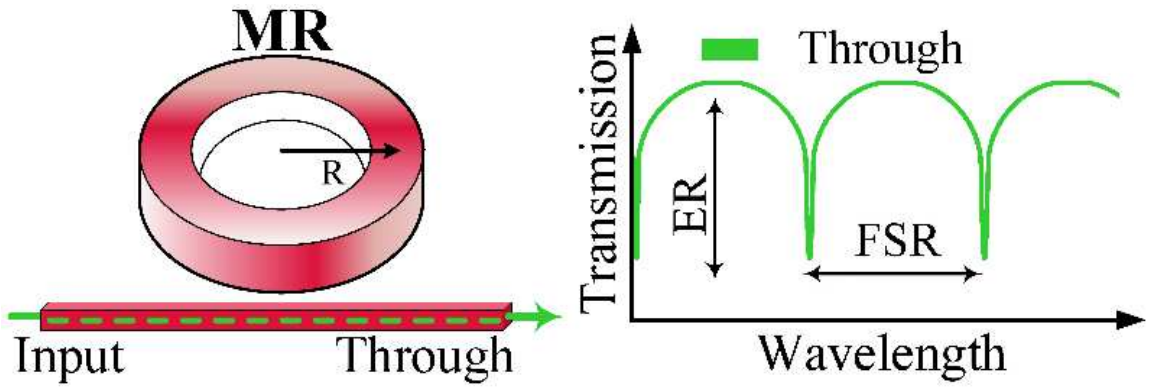


Figure 4.6: An all-pass MRR with output spectral characteristics at the through port with extinction ratio (ER) and free spectral range (FSR) specified in the figure.

in Fig. 4.5 represents one arm working in parallel with other arms) also increases optical splitter losses. Moreover, without considering crosstalk mitigation strategies (as is the case with previously proposed photonic accelerators), there is increased crosstalk noise in the optical signals, which drives down the weight resolution of the architecture.

In summary, to design efficient photonic accelerators, there is a need for (i) improved MRR device design to better tolerate variations and crosstalk; (ii) efficient MRR tuning circuits to quickly and reliably imprint activation and parameter values; and (iii) a scalable architecture design that minimizes optical signal losses. Our novel CrossLight photonic accelerator design addresses all of these concerns and is discussed next.

4.3.3 Crosslight Architecture

Fig. 4.7 shows a high-level overview of our *CrossLight* noncoherent silicon photonic neural network accelerator. The photonic substrate performs vector dot product (VDP) operations using silicon photonic MRR devices, and summation using optoelectronic photodetector (PD) devices over multiple wavelengths. An electronic control unit is required for the control of photonic devices, and for communication with a global memory to obtain the parameter values, mapping of the vectors, and for partial sum buffering. We use digital to analog converter (DAC) arrays to convert buffered signals into analog tuning signals for MRRs. Analog to digital converter (ADC) arrays are used to map the output analog signals generated by PDs to digital values that are sent back for

post-processing and buffering. We break down the discussion of this accelerator into three parts (subsections A-C), corresponding to the contributions at the device, tuning circuit, and architecture levels, as discussed next.

4.3.3.1 MRR device engineering and fabrication

Process variations are inevitable in CMOS-compatible silicon photonic fabrications, causing undesirable changes in resonant wavelength of MRR devices ($\Delta\lambda_{MRR}$). We fabricated a 1.5×0.6 mm^2 chip with high-resolution Electron Beam (EBeam) lithography and performed a comprehensive design-space exploration of MRRs to compensate for FPVs while improving MRR device insertion loss and Q-factor. In this exploration, we varied the input and ring waveguide widths to find an MRR device design that was tolerant to FPVs. We found that in an MRR design of any radii and gap, when the input waveguide is 400 nm wide and the ring waveguide is 800 nm wide at room temperature (300 K), the undesired $\Delta\lambda_{MRR}$ due to FPVs can be reduced from 7.1 to 2.1 nm (70% reduction). *This is a significant result, as these engineered MRRs require less compensation for FPV-induced resonant wavelength shifts, which can reduce the power consumption of architectures using such MRRs*.

Unfortunately, even with such optimized MRR designs, the impact of FPVs is not completely eliminated, and there is still a need to compensate for FPVs. Thermal variations are another major factor to cause changes in MRR n_{eff} which also leads to undesirable $\Delta\lambda_{MRR}$. Thermo-optic (TO) tuners are used to compensate for such deviations in $\Delta\lambda_{MRR}$. These TO tuners use microheaters to change the temperature in the proximity of an MRR device, which then alters the $n_{\text{eff}}$of the MRR, changing the device resonant wavelength, and correcting the $\Delta\lambda_{MRR}$. Unfortunately, high temperatures from such heaters can cause thermal energy dissipation, creating thermal crosstalk across MRR devices placed close to each other. One can avoid such thermal crosstalk by placing devices at an appropriate distance from each other, typically 120 μm to 200 μm (depending on the number of MRR devices in proximity within an MRR bank). But such a large spacing hurts area efficiency and also increases waveguide length, which increases propagation losses and its

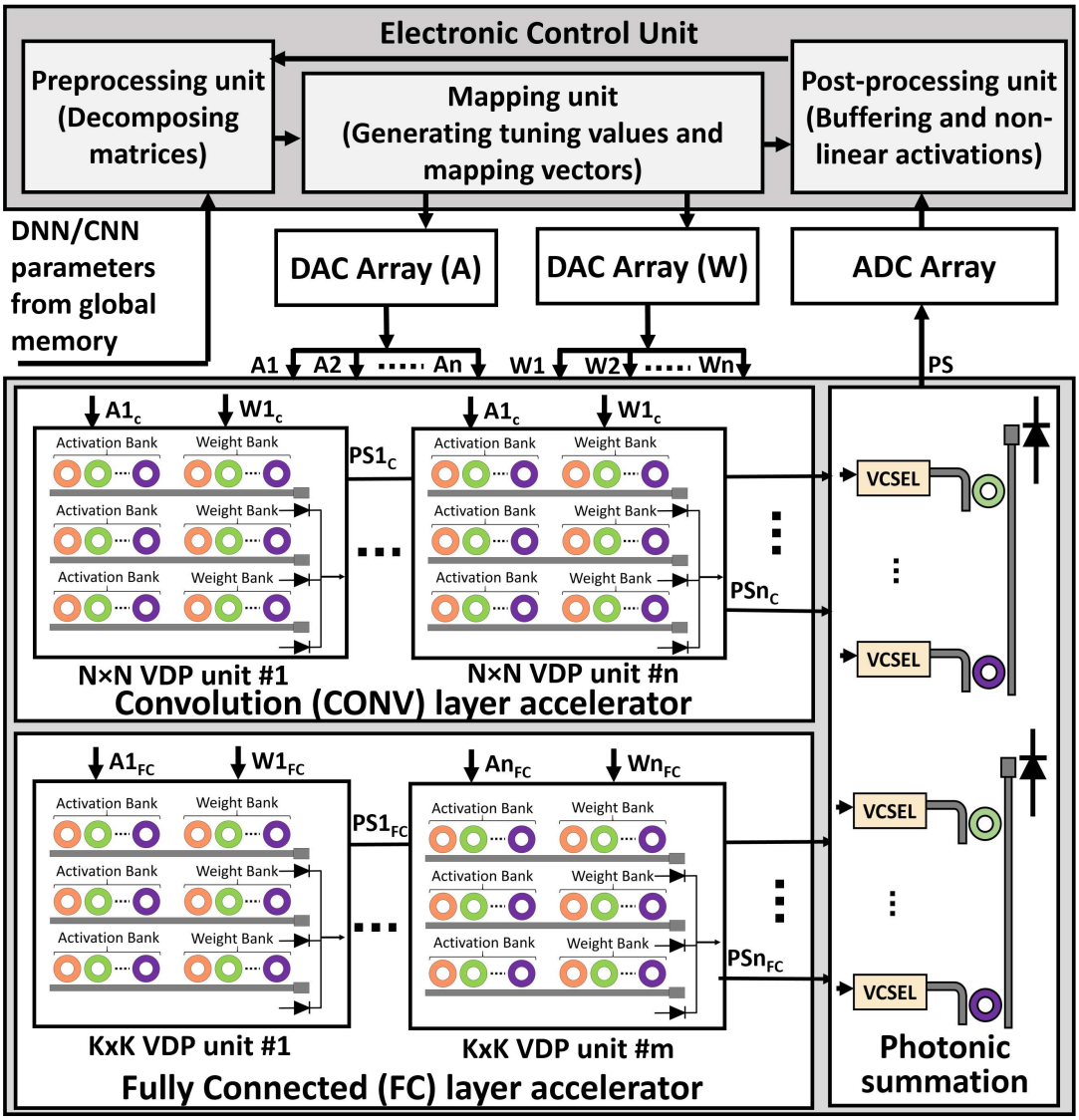


Figure 4.7: An all-pass MRR with output spectral characteristics at the through port with extinction ratio (ER) and free spectral range (FSR) specified in the figure.

associated laser power overhead. We propose to address this challenge at the circuit level, as discussed next.

4.3.3.2 Tuning circuit design

To reduce thermal crosstalk, we must reduce the reliance on TO tuning, an approach that is used in all prior photonic neural network accelerators, but one that entails high overheads. We propose to use a hybrid tuning circuit where both thermo-optic (TO) and electro-optic (EO) tuning are used to compensate for $\Delta\lambda_{MRR}$. Such a tuning approach has previously been proposed in [87] for silicon photonic Mach–Zehnder Interferometers with low insertion loss. Such an approach can be easily transferred to an optimized MRR for hybrid tuning in our architecture. The hybrid tuning approach supports faster operation of MRRs with fast EO tuning to compensate for small $\Delta\lambda_{MRR}$ shifts and, when necessary, using TO tuning when large $\Delta\lambda_{MRR}$ shifts need to be compensated.

To further reduce the power overhead of TO tuning in this hybrid approach, we adapt a method called Thermal Eigen Decomposition (TED), which was first proposed in [88]. Using TED, we can collectively tune all the MRRs in an MRR bank to compensate for large $\Delta\lambda_{MRR}$ shifts. By doing so, we can cancel the effect of thermal crosstalk (i.e., an undesired phase change) in MRRs with much lower power consumption. The TO tuning power can be calculated by the amount of phase shift necessary to apply to the MRRs in order for them to be at their desired resonant wavelength. The extent of phase crosstalk ratio (due to thermal crosstalk) as a function of the distance between an MRR pair is shown in Fig. 4.8, for our fabricated MRR devices. The results are based on detailed analysis with a commercial 3D heat transport simulation EDA tool for silicon photonic devices (Lumerical HEAT [89]). It can be seen from the orange line that as the distance between an MRR pair increases, the amount of phase crosstalk reduces exponentially. Such a trend has also been observed in [90]. To find a balance between tuning power savings while having reduced crosstalk, we perform a sensitivity analysis based on the distance between two adjacent MRRs in our architecture. We placed the optimized MRRs (described in the previous section) in such a manner that maximum tuning power is saved when they are close to each other while compensating for thermal crosstalk. Results from our analysis (the solid-blue line in Fig. 4.8) indicate that placing

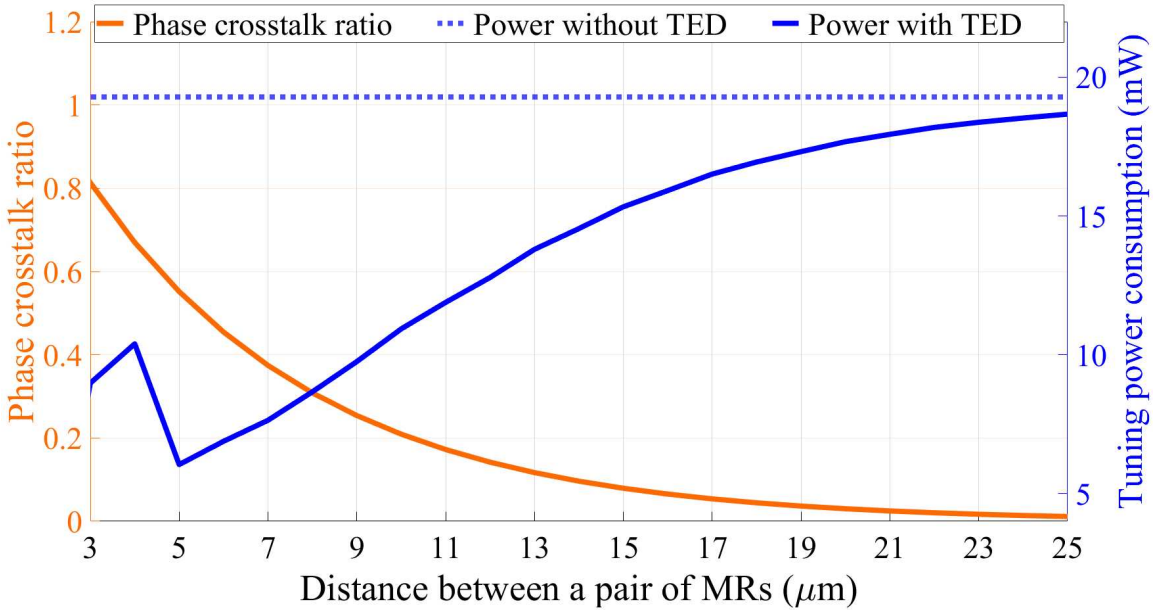


Figure 4.8: Phase crosstalk ratio and tuning power consumption in a block of 10 fabricated MRRs with variable distance between adjacent pair of MRRs.

each MRR pair at a distance of 5 μm is optimal, as increasing or decreasing such a distance causes an increase in power consumption of individual TO heaters in the MRRs. Fig. 4.8 also shows the tuning power required without using the TED approach (blue dotted line), which can be seen to be notably higher.

The workflow of our circuit-level hybrid tuning approach can be summarized as follows. When the accelerator is first booted at runtime, a one-time compensation for design-time FPVs is applied using TO tuning. The extent of compensation for crosstalk is calculated offline during the test phase, where the required phase shift in each of the MRRs is calculated, and once the system is online, the respective phase shift values are applied to cancel the impact of thermal crosstalk. Subsequently, we apply EO tuning due to its extremely low latency to represent vector elements in each vector operation with MRRs (discussed in more detail in the next section). If large shifts in temperature are observed at runtime, we can perform a one-time calibration with TO tuning to compensate for it. In our analysis, runtime TO tuning would be required rarely beyond its first use after the initial bootup of the photonic accelerator platform.

4.3.3.3 Architecture design

The optimized MRR devices, layouts, and tuning circuits are utilized within optical vector dot product (VDP) units, which are shown in Fig. 4.7. We use banks (groups) of MRRs to imprint both activations and weights onto the optical signal. At the architecture level, we compose multiples of VDP units into two architectural sub-components: one to support convolution (CONV) layer acceleration and the other to support fully connected (FC) layer acceleration. We focus on these two types of layers as they are the most widely used and consume the most significant amount of latency and power in computational platforms that execute DNNs. In contrast, other layer types (e.g., pooling, batch normalization) can be implemented very efficiently in the electronic domain. Note also that we focus on inference acceleration, as done in all photonic DNN accelerators, and almost all electronic DNN accelerators.

4.3.3.3.1 Decomposing vector operations in CONV/FC layers To map CONV and FC layers from DNN models to our accelerator, we first need to decompose large vector sizes into smaller ones. In CONV layers, a filter performs convolution on a patch (e.g., 2×2 elements) of the activation matrix in a channel to generate an element of the output matrix. The operation can be represented as follows:

$$K \bigotimes A = Y \tag{4.3}$$

For a 2×2 filter kernel and weight matrices, (4.3) can be expressed as:

$$\begin{bmatrix} k_1 & k_2 \\ k_3 & k_4 \end{bmatrix} \bigotimes \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix} = k_1a_1 + k_2a_2 + k_3a_3 + k_4a_4 \tag{4.4}$$

Rewriting (4.4) as a vector dot product, we have:

$$\begin{bmatrix} k_1 & k_2 & k_3 & k_4 \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} = k_1a_1 + k_2a_2 + k_3a_3 + k_4a_4 \tag{4.5}$$

Once we are able to represent the operation as a vector dot product, it is easy to see how it can be decomposed into partial sums. For example:

$$\begin{bmatrix} k_1 & k_2 \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = k_1a_1 + k_2a_2 + k_3a_3 + k_4a_4 = PS_1 \tag{4.6}$$

$$\begin{bmatrix} k_3 & k_4 \end{bmatrix} \cdot \begin{bmatrix} a_3 \\ a_4 \end{bmatrix} = k_1a_1 + k_2a_2 + k_3a_3 + k_4a_4 = PS_2 \tag{4.7}$$

$$PS_1 + PS_2 = Y \tag{4.8}$$

In FC layers, typically much larger dimension vector multiplication operations are performed between input activations and weight matrices:

$$AW = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \begin{bmatrix} w_1 & w_2 & \dots & w_n \end{bmatrix} \tag{4.9}$$

$$AW = \begin{bmatrix} a_1 \cdot w_1 & +a_1 \cdot w_2 \dots & +a_1 \cdot w_n \\ a_2 \cdot w_1 & +a_2 \cdot w_2 \dots & +a_2 \cdot w_n \\ \vdots & \vdots & \vdots \\ a_n \cdot w_1 & +a_n \cdot w_2 \dots & +a_n \cdot w_n \\ \end{bmatrix} \tag{4.10}$$

In (4.9), a_1 to a_n represent a column vector of activations (A) and w_1 to w_n represent a row vector of weights (W). The resulting vector is a summation of dot products of vector elements (4.10). Much like with CONV layers, these can be decomposed into lower dimensional dot products.

4.3.3.3.2 Vector dot product (VDP) unit design We separated the implementation of CONV and FC layers in *CrossLight* due to the vastly different orders of vector dot product computations required to implement each layer. For instance, typical CONV layer kernel sizes vary from 2×2 to 5×5, whereas in FC layers it is not uncommon to have 100 or more neurons (requiring 100×100 or higher order multiplication). State-of-the-art photonic DNN accelerators, e.g., [82], only consider the scales involved at the CONV layer, and either only support CONV layer acceleration in the optical domain, or use the same CONV layer implementation to accelerate FC layers. This will lead to increased latencies and reduced throughput as the larger vectors involved with FC layer calculation must be divided up into much smaller chunks, in the order of the filter kernel size of the CONV layer.

For improved efficiency, we separately support the unique scale and requirements of vector dot products involved in CONV layers and FC layers. For CONV layer acceleration, we consider

n VDP units, with each unit supporting an $N \times N$ dot product. For FC layer acceleration, we consider m units, with each unit supporting a $K \times K$ dot product. Here $n > m$ and $K > N$, as per the requirements of each of the distinct layers. In each of the VDP units, the original vector dimensions are decomposed into N or K dimensional vectors, as discussed above. We performed an exploration to determine the optimal values for N, K, n, and m. The results of this exploration study are presented in Section 4.3.4.

4.3.3.3.3 Optical wavelength reuse in VDP units Prior work on photonic DNN accelerator design typically considers a separate wavelength to represent each individual element of a vector. This approach leads to an increase in the total number of lasers needed in the laser bank (as the size of the vectors increases) which in turn increases power consumption. Beyond employing the decomposition approach discussed above, we also consider wavelength reuse per VDP unit to minimize laser power. In this approach, within VDP units, the N or K dimensional vectors are further decomposed into smaller sized vectors for which dot products can be performed using MRRs in parallel, in each arm of the VDP unit. The same wavelengths can then be reused across arms within a VDP to reduce the number of unique wavelengths required from the laser. PDs perform summation of the element-wise products to generate partial sums from decomposed vector dot products. The partial sums from the decomposed operations are then converted back to the photonic domain by VCSELs (bottom right of Fig. 4.7), multiplexed into a single waveguide, and accumulated using another PD, before being sent for buffering. Thus, our approach leads to an increase in the number of PDs compared to other accelerators but significantly reduces both the number of MRRs per waveguide and the overall laser power consumption.

In each arm within a VDP unit, we used a maximum of 15 MRRs per bank for a total of 30 MRRs per arm, to support up to a 15×15 vector dot product. The choice of MRRs per arm considers not only the thermal crosstalk and layout spacing issues (discussed earlier), and the benefits of wavelength reuse (discussed in previous para), but also the fact that optical splitter losses become non-negligible as the number of MRRs per arm increases, which in turn increases

Table 4.2: Models and datasets considered for evaluation

Model no.	CONV layers	FC layers	Parameters	Datasets
1	2	2	60,074	Sign MNIST
2	4	2	890,410	CIFAR10
3	7	2	3,201,080	STL10
4	8	4	38,951,745	Omniglot

Table 4.3: Parameters considered for analysis of photonic accelerators

Devices	Latency	Power
EO Tuning [48]	20ns	4μmW/nm
TO Tuning [71]	4 μs	27.5 mW/FSR
VCSEL [93]	10 ns	0.66 mW
TIA [94]	0.15 ns	7.2 mW
Photodetector [95]	5.8 ps	2.8 mW

laser power requirements. Thus, the selection of MRRs per arm within a VDP unit was carefully adjusted to balance parallelism within/across arms, and laser power overheads.

4.3.4 Evaluation and Simulation Results

4.3.4.1 Simulation setup

To evaluate the effectiveness of our *CrossLight* accelerator, we conducted several simulation studies. These studies were complemented by our MRR-device fabrication and optimization efforts on real chips, as discussed in Section 4.3.3. We considered the four DNN models shown in Table 4.2 for execution on the accelerator. Model 1 is Lenet5 [91] and models 2 and 3 are custom CNNs with both FC and CONV layers. Model 4 is a Siamese CNN utilizing one-shot learning. The datasets used to train these models are also shown in the table. We designed a custom *CrossLight* accelerator simulator in Python to estimate its performance and power/energy. We used Tensorflow 2.3 along with Qkeras [92], for analyzing DNN model accuracy across different parameter resolutions.

We compared *CrossLight* with the DEAP-CNN [82] and Holylight [83] photonic DNN accelerators from prior work. Table 4.3 shows the optoelectronic parameters considered for this simulation-based analysis. We considered photonic signal losses due to various factors: signal

propagation (1 dB/cm [42]), splitter loss (0.13 dB [96]), combiner loss (0.9 dB [97]), MRR through loss (0.02 dB [98]), MRR modulation loss (0.72 dB [99]), microdisk loss (1.22 dB [100]), EO tuning loss (6 dB/cm [48]), and TO tuning loss (1 dB/cm [71]). We also considered the 1-to-56-Gb/s ADC/DAC-based transceivers from recent work [101]. To calculate laser power consumption, we use the following laser power model:

$$P_{\text{laser}} - S_{detector} \geq P_{\text{photo\_loss}} + 10 \times \log_{10} N_{\lambda}, \tag{4.11}$$

where P_{laser} is laser power in dBm, $S_{detector}$ is the PD sensitivity in dBm, and $P_{\text{photo\_loss}}$ is the total photonic loss encountered by the optical signal, due to all of the factors discussed above.

4.3.4.2 Results: *CrossLight* resolution analysis

We first present an analysis of the resolution that can be achieved with *CrossLight*. We consider how the optical signals from MRRs impact each other due to their spectral proximity, also known as inter-channel crosstalk. For this, we use the equations from [102]:

$$\varphi(i,j) = \frac{\delta^2}{(\lambda_i - \lambda_j)^2 + \delta^2} \tag{4.12}$$

In (4.12), $\varphi(i,j)$ describes the noise content from the j^{th} MRR present in the signal from the i^{th} MRR. As the noise content increases, the resolution achievable with *CrossLight* will decrease. Also, $((\lambda_i - \lambda_j))$ is the difference between the resonant wavelengths of i^{th} MRR and j^{th} MRR, while $\delta = \lambda_i/2Q$ denotes the 3dB bandwidth of the MRRs, with Q being the quality factor (Q-factor) of the MRR being considered. The noise power component can thus be calculated as:

$$P_{\text{noise}} = \sum_{i}^{n-1} \varphi(i,j) P_{\text{in}}[i] \tag{4.13}$$

For unit input power intensity, resolution can then be computed as:

$$\text{Resolution} = \frac{1}{max|P_{\text{noise}}|} \tag{4.14}$$

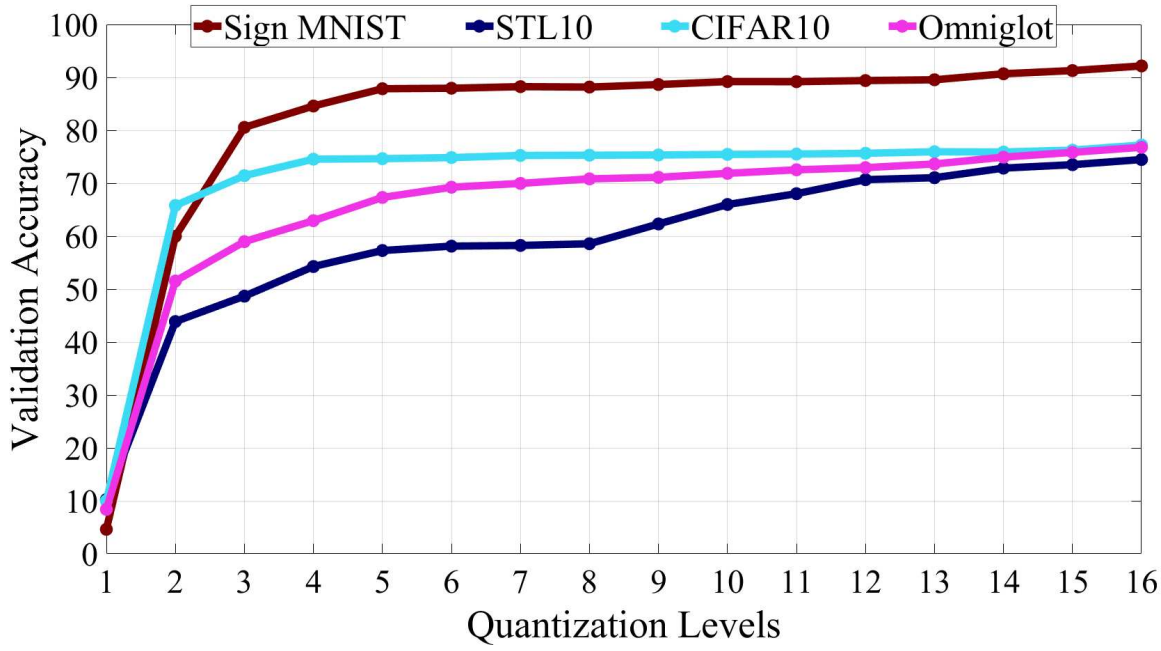


Figure 4.9: Inference accuracy of the four DNN models considered, across quantization (resolution) range from 1 bit to 16 bits (for both weights and activations).

From this analysis, we found that with the FSR value of 18 nm and the Q value of ≈8000 in our optimized MRR designs, and the wavelength reuse strategy in *CrossLight*, which allows us to have large $(\lambda_i - \lambda_j)$ values (>1 nm), our MRR banks will be able to achieve a resolution of 16 bits for up to 15 MRRs per bank (Section 4.3.3.3.2). This is much higher than the resolution achievable by many photonic accelerators. For instance, DEAP-CNN can only achieve a resolution of 4 bits, whereas Holylight can only achieve a 2-bit resolution per microdisk (they however combine 8 microdisks to achieve an overall 16-bit resolution). Higher resolution ensures better accuracy in inference, which can be critical in some applications. Fig. 4.9 shows the impact of varying the resolution across the weights and activations from 1 bit to 16 bits (we used quantization-aware training to maximize accuracy), for the four DNN models considered (Table 4.2). It can be observed that model inference accuracy is sensitive to the resolution of weight and activation parameters. Models such as the one for STL10 are particularly sensitive to the resolution. Thus, the high resolution afforded by *CrossLight* can allow achieving higher accuracies than other photonic DNN accelerators, such as DEAP-CNN.

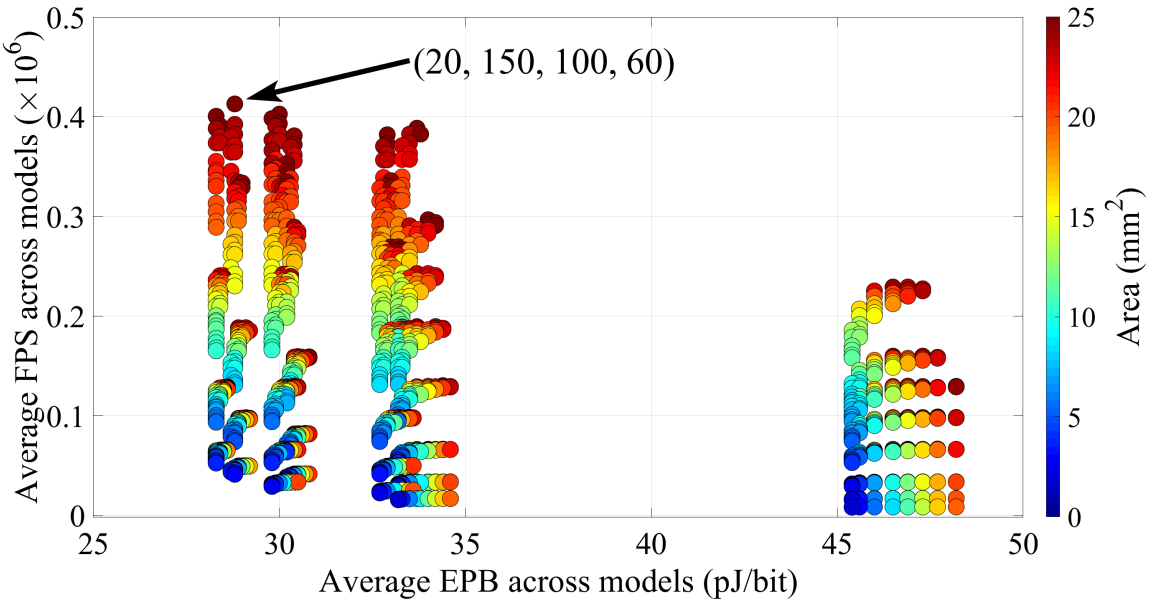


Figure 4.10: Scatterplot of average FPS vs. average EPB vs. area of various *CrossLight* configurations. The configuration with highest FPS/EPB (and FPS) is highlighted.

4.3.4.3 Results: *CrossLight* sensitivity analysis

We performed a sensitivity analysis by varying the number of VDP units in the CONV layer accelerator (n) and FC layer accelerator (m), along with the complexity of the VDP units (N and K, respectively). Fig. 4.10 shows the frames per second (FPS; a measure of inference performance) vs. energy per bit (EPB) vs. area of various configurations of *CrossLight*. We selected the best configuration as the one that had the highest value of FPS/EPB. In terms of (N, K, n, m), the values of the four parameters for this configuration are (20, 150, 100, 60). This configuration also ended up being the one with the highest FPS value, but had a higher area overhead than other configurations. Nonetheless, this area is comparable to that of other photonic accelerators. We used this configuration for comparisons with prior work, as discussed next.

4.3.4.4 Results: Comparison with state-of-the-art accelerators

We compared our *CrossLight* accelerator against two well-known photonic accelerators: DEAP-CNN and Holylight, within a reasonable area constraint for all accelerators (16-25 mm^2). We present results for four variants of the *CrossLight* architecture: 1) *Cross_base* utilizes conventional MRR designs (without FPV resilience) and traditional TO tuning; 2) *Cross_opt* utilizes the opti-

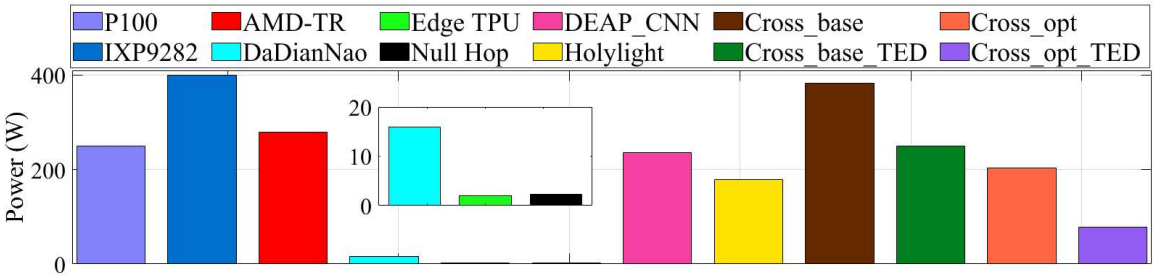


Figure 4.11: Power consumption comparison among variants of *CrossLight* vs. photonic accelerators (DEAP-CNN, Holylight), and electronic accelerator platforms (P100, Xeon Platinum 9282, Threadripper 3970x, DaDianNao, EdgeTPU, Null Hop)

mized MRR designs from Section IV.A, and traditional TO tuning; 3) *Cross_base_TED* utilizes the conventional MRR designs with the hybrid TED-based tuning approach from Section 4.3.3.2; and 4) *Cross_opt_TED* utilizes the optimized MRR designs and the hybrid TED-based tuning approach.

Fig. 4.11 shows the power consumption comparison across the four *CrossLight* variants and the two photonic accelerators from prior work. We also include comparison numbers for electronic platforms: three deep learning accelerators (DaDianNao, Null Hop, and EdgeTPU), a GPU (Nvidia Tesla P100), and CPUs (Intel Xeon Platinum 9282 denoted as IXP9282, and AMD Threadripper 3970x denoted as AMD-TR) [103]. The difference in power values between the *CrossLight* variants arises due to the optimization approaches adopted in each of the variant. The variants which considered conventional MRR design instead of the optimized designs have larger power consumption for compensating for FPV. This value becomes non-trivial as the number of MRRs increase, and thus having reduced tuning power requirement per MRR (in *Cross_opt* and *Cross_opt_TED*) becomes a significant advantage. Using the TED based hybrid tuning approach provides further significant power benefits for *Cross_opt_TED* over Cross_opt, which uses conventional TO tuning. *Cross_opt_TED* can be seen to have lower power consumption than both photonic accelerators, as well as the CPU and GPU platforms, although this power is higher than that of the edge/mobile electronic accelerators.

Fig. 4.12 shows a comparison of energy-per-bit (EPB) across all of the photonic accelerators, for the four DNN models. On average, our best *CrossLight* configuration (*Cross_opt_TED*) has

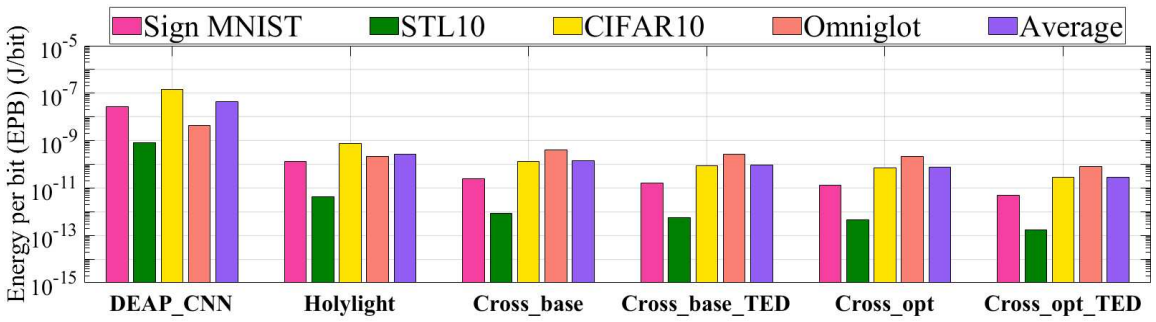


Figure 4.12: Comparison of EPB values of the photonic DNN accelerators)

1544× and 9.5× lower EPB compared to DEAP-CNN and Holylight, respectively. The reason for *CrossLight*'s lower EPB is because we comprehensively took into consideration various losses and crosstalk that a photonic DNN accelerator would experience, and put in place novel approaches at the device, circuit, and architecture layers to counteract their impact in *CrossLight*. The utilization of TED-based thermal crosstalk management allows us to have MRRs placed much closer together, which in turn reduces propagation losses. In addition, *CrossLight* considers a combination of TO and EO tuning which enables the reduction of power and EPB as well. The use of EO tuning in our hybrid tuning approach also provides the advantage of lower latencies, which is apparent in the EPB values.

Table 4.4 summarizes the average values of EPB (in pJ/bit) and performance-per-watt (in kilo-FPS/Watt) of the photonic accelerators as well as the electronic accelerators considered in this work. It can be observed that the best *CrossLight* configuration (*Cross_opt_TED*) achieves significantly lower EPB and higher performance-per-watt values than all of the accelerators considered. Specifically, against Holylight, which is the best out of the two photonic DNN accelerators considered, *CrossLight* achieves 9.5× lower energy-per-bit and 15.9× higher performance-per-watt. Our work demonstrates the effectiveness of cross-layer design of deep learning accelerators with the emerging silicon photonics technology. With the growing maturity of silicon photonic device fabrication in CMOS-compatible processes, it is expected that the energy costs of device tuning, losses, and laser power overheads will go further down, making an even stronger case for considering optical-domain accelerators for deep learning inference.

Table 4.4: Average EPB and kiloFPS/Watt values across accelerators

Accelerator	Avg. EPB(pJ/bit)	Avg. kiloFPS/watt
P100	971.31	24.9
IXP 9282	5099.68	2.39
AMD-TR	5831.18	2.09
DaDianNao	58.33	0.65
Edge TPU	697.37	17.53
Null Hop	2727.43	4.48
DEAP_CNN	44453.88	0.07
Holylight	274.13	3.3
Cross_base	142.35	10.78
Cross_base_TED	92.64	16.54
Cross_opt	75.58	20.25
Cross_opt_TED	28.78	52.59

4.3.5 Conclusion

In this section of the chapter, we presented a novel cross-layer optimized photonic neural network accelerator called *CrossLight*. Utilizing silicon photonic device-level fabrication-driven optimizations along with circuit-level and architecture-level optimizations, we demonstrated 9.5× lower energy-per-bit and 15.9× higher performance-per-watt compared to state-of-the-art photonic DNN accelerators. *CrossLight* also shows improvements in these metrics over several CPU, GPU, and custom electronic accelerator platforms considered in our analysis. *CrossLight* shows the promise of cross-layer optimization strategies in countering various challenges such as crosstalk, fabrication-process variations, high laser power, and excessive tuning power. The results presented in this chapter demonstrate the promise of photonic DNN accelerators in addressing the need for energy-efficient and high performance-per-watt DNN acceleration.

4.4 ROBIN: A Robust Optical Binary Neural Network Accelerator

In this section, we introduce robust optical binary neural network (ROBIN), a novel optical-domain Binarized Neural Network (BNN) accelerator that addresses the challenges highlighted

above by optimizing electro-optic components across the device, circuit, and architecture layers. *ROBIN* combines novel device- and circuit-level techniques to achieve more efficient fabrication-process-variation (FPV) correction in optical devices, which helps with reducing energy and improving accuracy in BNNs that utilize these devices. Additionally, circuit- level tuning enhancements for inference latency reduction, and an optimized architecture-level design help improve performance and also energy consumption compared to the state-of-the-art. Our novel contributions in this work include:

- The design of a novel optical-domain BNN accelerator architecture that is robust to FPVs and thermal variations, and utilizes efficient wavelength reuse and a modular structure to enable high throughput and energy-efficient execution across BNN models;
- A novel integration of heterogeneous optical microring resonator (MRR) devices; we also conduct design space exploration for these MRR designs to determine device characteristics for efficient BNN execution;
- An enhanced tuning circuit to simultaneously support large thermal-induced resonance shifts and high-speed, low-loss device tuning to compensate for FPVs;
- A comprehensive comparison with state-of-the-art BNN and non-BNN accelerator platforms from the optical and electronic domains, to demonstrate the potential of our BNN accelerator platform.

The rest of the section in this chapter is organized as follows: Section 4.4.1 briefly explores the related works in the field of BNN acceleration. Section 4.4.2 gives a brief overview of non-coherent optical computation for photonic accelerators similar to ours. Section 4.4.3 provides an overview of BNNs and the partially binarized approach we have adopted for better accuracy in models. Section 4.4.4 describes the *ROBIN* architecture and our optimization efforts in tuning circuits, photonic devices, and photonic system level. Details of the experiments conducted, simulation setup, and the obtained results are provided in Section 4.4.5. Finally, Section 4.4.6 presents some concluding remarks and directions for future work.

4.4.1 Related Work

Silicon-photonic-based DNN accelerator architectures are becoming increasingly prominent with significant interest from both academic and industrial research communities [104]. This growth in interest can be attributed to the benefits of photonic acceleration over electronic acceleration, as discussed in the previous section. Optical DNN accelerator architectures can be broadly classified into two types: coherent architectures and non-coherent architectures. Coherent architectures use a single wavelength to operate and imprint weight/activation parameters onto the electrical field amplitude of the light wave [37, 84]. These architectures mainly use on-chip optical interferometer devices like Mach-Zehnder Interferometers (MZIs). For imprinting the parameters, optical phase- change mechanisms can be introduced to MZI devices. These mechanisms use heating or carrier injection to change the refractive index in the MZI structure. Weighting occurs with electrical field amplitude attenuation proportional to the weight value, and phase modulation that is proportional to the sign of the weight. The weighted signals are then accumulated with cascaded optical com- biners, through coherent interference. Here the term coherent refers to the physical property of the wave, where it is possible for the wave to interfere constructively or destructively, on the same wavelength. Non-coherent architectures, such as [18, 82, 83], use multiple wavelengths, where each wavelength can be used to perform an individual neuron operation. These architectures are referred to as non-coherent architectures as they use different optical wavelengths, the interaction among which can be non-coherent. A large number of neurons can be represented simultaneously in non-coherent architectures by using wavelength-division multiplexing (WDM) or dense WDM (DWDM). In these architectures, parameters are imprinted on to the signal amplitude directly, and to manipulate individual wavelengths, wavelength-selective devices such as microring resonators (MRRs) or microdisks are used. The optical signal power is controlled, for imprinting parameter values, by controlling the optical loss in these devices through tuning mechanisms (Section 4.4.4.1). The Broadcast and Weight (B&W) protocol [81] is typically employed for setting and updating the weight and activation values. The *ROBIN* architecture we present in this chapter is a noncoherent architecture, i.e., it uses multiple wavelengths that are routed to photonic

computation units in waveguides using WDM in accordance with the B&W protocol. The growing interest in noncoherent architectures can be attributed to the limitations in scalability, phase encoding noise, and phase error accumulation in coherent architectures [85, 104].

For optical DNN acceleration using noncoherent mechanics, [82] introduced a photonic accelerator for CNNs where all the layers of CNN models are implemented using connected photonic convolution units. In these units, MRRs are used to tune wavelengths to desired kernel values through phase tuning. Another such work, in [83], utilizes microdisks instead of MRRs due to the lower area and power consumption they offer. But microdisks use 'whispering gallery mode' resonance which is inherently lossy due to the tunneling ray attenuation phenomenon [86]; which reduces reliability and energy-efficiency with microdisks. There are very few works which focus on implementations of BNN accelerators with optical components. The work in [105] proposed an MR-based accelerator for discretized neural network acceleration, with an encoding scheme to enable positive and negative product considerations. The authors in [106] leveraged microdisks for implementing an accelerator with a design similar to [83]. This work considered an accelerator for fully binarized neural networks, i.e., both weights and activations and considered to be single-bit parameters. Be- cause of this simplification, [106] was able to utilize energy-efficient photonic XOR and population count operations instead of conventional multiply and accumulate operations. The work also made use of photonic non-volatile memory and claimed operating frequencies of up to 50 GHz. All of these existing works on noncoherent optical-domain DNN/BNN acceleration have several short- comings. They suffer from susceptibility to fabrication-process variations (FPVs) and also thermal crosstalk, which are not addressed in these architectures. Microsecond-granularity thermo-optic tuning latencies further can reduce the speed and efficiency of optical computing [71], which is also not considered when analyzing accelerator performance. We address these crucial shortcomings as part of our *ROBIN* optical-domain BNN accelerator architecture in this work.

In this chapter, we aim to ensure the robustness of the architecture against process and thermal variations by using MRR design-space exploration and photonic tuning-circuit optimizations,

which will be further explained in Section 4.4.3. We also utilize the broadband capabilities of the key photonic device in our work, microring resonators (MRRs), to perform batch normalization folding, which moves batch normalization operations from the electrical domain to the photonic domain. Section 4.4.4.3 further details the modular architectural design aiming at ensuring wavelength reuse, to reduce VCSEL usage and splitter losses and waveguide length reduction. We also explore how the architecture performs in the presence of FPVs and how we may further reduce energy con- sumption in terms of device tuning in this scenario, in Section 4.4.5.2.

4.4.2 Overview of Noncoherent Optical Computation

Noncoherent optical accelerators leverage the low-latency and energy-efficient optical computation for multiply and accumulate (MAC) operations, which consumes substantial computational power and incurs high latencies in electronic accelerators. These accelerators typically utilize the B&W protocol with multiple wavelengths. Fig. 4.13(a) (from [2]) gives an overview of a B&W-based optical MAC unit. The figure depicts a recurrent MAC unit which is employed repeatedly to compute different layers of a neural network model. The layer parameters such as weights or activations can be imprinted on to the wavelengths using the MRRs that are tuned to modify the optical signal intensity to represent those values. The MRRs are placed in MRR banks where multiple parameters can be imprinted onto wavelengths simultaneously. In the MRR banks, each MRR is tuned to a specific optical wavelength and can be used to alter the optical characteristics of the wavelength (thereby changing its intensity) to represent the imprinted parameter. There can be separate wavelengths which carry positive and negative parameters, as discussed in [105]; these parameters are summed using balanced photodetectors (BPDs), as shown in Fig. 4.13(a).

The output from the MAC unit is passed on to a Mach-Zehnder Modulator (MZM) which tunes the output from a designated laser diode (LD) to this output. Multiple MZMs and LDs are used to generate the outputs from multiple MAC units; these are collected and multiplexed using an arrayed waveguide grating (AWG) based optical multiplexer (MUX). The output from the MUX is passed back into the MAC units, through splitters, now embedded with parameters from the next

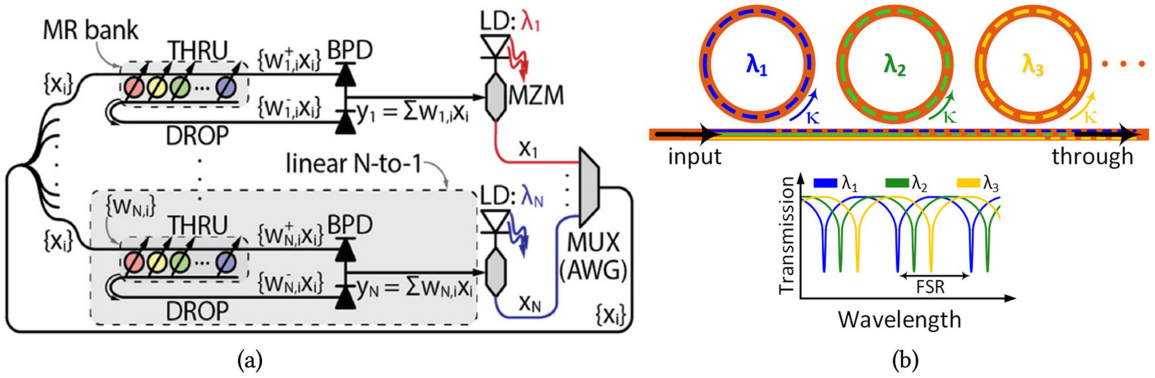


Figure 4.13: (a) A recurrent noncoherent B&W MAC based design [2]; (b) An MRR bank consisting of MRRs with individual resonant wavelength (λ_i) coupled to the MRRs at cross-over coupling (κ) and the output spectrum, showing free spectral range (FSR).

layer. Devices such as electro-optic modulators (not depicted) may be used to implement non-linearities after the MAC operation. Unfortunately, the static nature of the hardware limits the size of the neural network model that can be accelerated using such a configuration. The number of splitters being used can also cause increased optical losses and thus higher laser power requirement to compensate for the losses as the size of an accelerator using this B&W configuration increases. MRRs and other on-chip optical resonators such as microdisks are crucial components in such noncoherent MAC configurations, as they impact the reliability and efficiency of the operation performed. Fig. 4.13(b) depicts an MRR bank and its output spectrum along with the free spectral range (FSR). Factors such as fabrication-process variations (FPVs) and thermal variations which impact the MRR critical dimensions and hence the effective refractive index (n_{eff}) of the device can cause a drift in the resonant wavelength ($\Delta\lambda_{MRR}$) [46]. This drift can introduce errors into optical computation and is thus usually compensated for (i.e., corrected) with TO or EO tuning circuits. While EO offers faster tuning (ns range) and consumes lesser power (4 μm/nm), it also has a smaller tuning range [48]. TO tuning, on the other hand, consumes higher power (27 mW/FSR) and has higher tuning latency (μs range) [71], but offers a larger tuning range. Because of the larger correction capacity, TO is often preferred over EO despite its higher latency and power consumption. Therefore, as the number of MRRs increases—when considering larger CNN or MLP models—the tuning power consumption also increases. This also creates increased wavelength

requirements per waveguide and calls for longer waveguides to host the MRRs, causing increased laser power consumption to supply the wavelengths and to compensate for the propagation losses in the longer waveguides. Also, more MRRs and more wavelengths increase optical crosstalk, and also introduce thermal crosstalk due to the larger number of TO tuners employed. To counteract these challenges, and ensure better weight resolution, crosstalk mitigation strategies must also be considered.

To design an effective optical-domain BNN accelerator, all of these considerations must be taken into account. This in turn highlights the need for (i) better device optimizations to tolerate variations; (ii) efficient and low-latency tuning mechanisms; (iii) a scalable architecture design, which is optimized for energy efficiency, area, and throughput. Our work aims to address all of these concerns for an efficient BNN accelerator implementation in the optical domain.

4.4.3 Binarized Neural Network

BNNs [107] are types of DNNs (or CNNs) where both weights and activation parameters only use binary values, and the binary values are utilized during both inference and backpropagation training. The binary nature of weights in BNNs makes them resilient to small perturbations which can usually lead to gross classification errors in DNNs. Inspired by the seminal work on efficiently training BNNs [107], recent efforts either explore how BNN accuracy can be improved, apply BNNs to different application domains, or explore how BNNs can be implemented efficiently in hardware to leverage their low computation power and memory requirements in resource constrained environments.

BNNs utilize the sign function to convert real valued weights to +1 or -1. But this typically leads to complications in training as the gradient for the sign function always results in a zero. The work in [108] introduced a heuristic called straight through estimator (STE) to circumvent this issue. STEs approximate the gradient by bypassing the gradient of the layer, by turning it into an identity function. The gradient thus obtained is used for updating real valued weights, using an optimization strategy such as Adam or stochastic gradient descent (SGD). This process is utilized

for activation parameters as well. Also, the use of batch normalization (BN) layers in BNNs has been shown to lead to several benefits [109]. The gain (γ) and bias (β) terms of the BN layer not only help condition the values during training, which speeds up BNN training, but also helps to improve accuracy in BNNs.

Another approach for increasing inference accuracy in BNNs is to employ partial binarization, where some of the layers are not completely binarized. The last layer is usually not binarized to avoid severe loss in accuracy. With detailed analysis of the model, critical layers can be identified and can be kept at higher precision, for better accuracy, at the cost of increased resource (computation, memory) utilization. To determine the appropriate activation parameter precision, which is required to determine the digital-to-analog converter (DAC) resolution in our accelerator architecture, we conducted a BNN accuracy analysis, where weights were restricted to binary (1-bit) values, but the bit precision level of the activations was altered from 1-bit to 16-bits. During BNN training, we ensured that we only binarize weights during the forward and backward propagations but not during the parameter update step, because keeping good precision weights during the updates is necessary for SGD to work at all (as parameter changes are usually tiny during gradient descent). After training, all weights were in binary format, while the precision of input activations was varied. Fig. 4.14 shows the results of varying activation precision across four different models and their datasets (described later in Section 4.4.5.1). We observed that the accuracy had notable change initially as activations bits were increased, but this gain in accuracy soon saturated. Based on the results, we consider 1-bit weights with 4-bit activations, and thus use 4-bit DACs in our architecture.

4.4.4 *ROBIN* Architecture

In this section, we describe the various optimization considerations at device, circuit, and architecture level used for designing the *ROBIN* architecture.

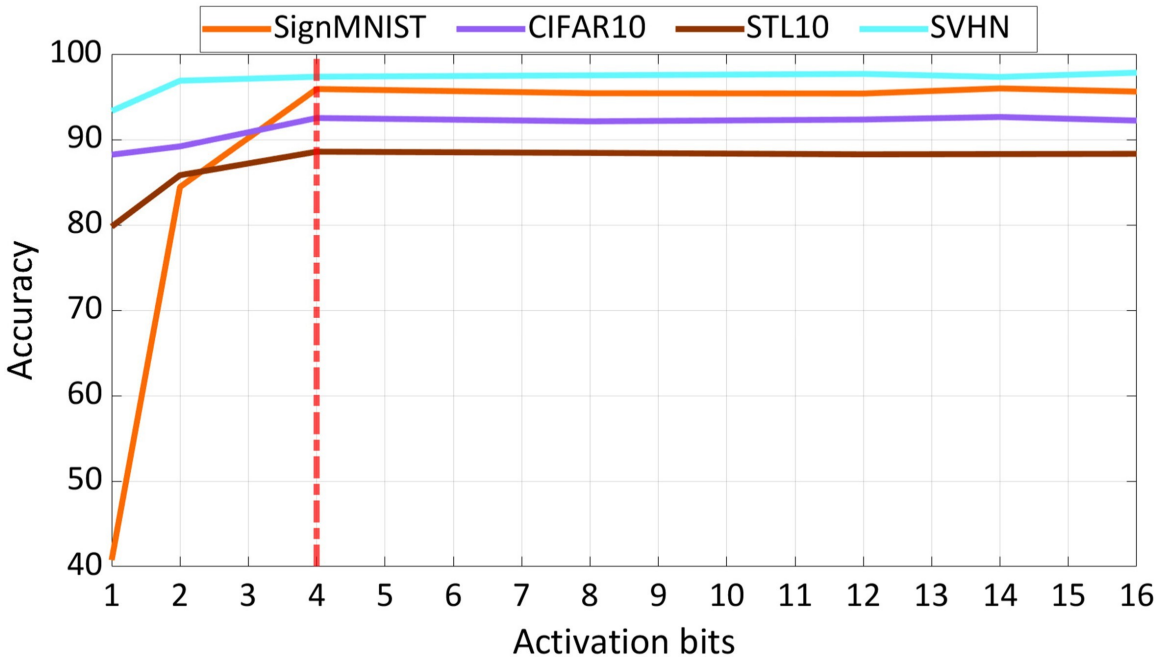


Figure 4.14: The accuracy sensitivity study conducted by varying activation parameter precision (number of bits). Weights are kept as binary values in all cases. The study was performed across four different models and their datasets (described later in Section 4.4.5.1

4.4.4.1 Tuning Circuit Design

A tuning circuit design is essential for fast and accurate operation of MRRs in our BNN accelerator. Errors due to fabrication-process variations (FPVs) can be significantly reduced by using an appropriate MRR tuning circuit. The tuning circuit employed can be either thermo-optic (TO) or electro-optic (EO) tuning circuits. Thermo-optic(TO)-based tuning mechanisms use microheaters to change the temperature in the proximity of a microring resonator (MRR), which then alters the effective index (n_{eff}) of the MRR. This in turn changes the device characteristics such as resonant wavelength (λ_{MRR}). Such a change in resonant wavelength ($\Delta\lambda_{MRR}$) can help compensate for fabrication-process and thermal variations in MRRs. The electro-optic (EO)-based tuning mechanisms in an MRR is based on the depletion and injection of carriers on a PN diode. EO tuning is faster (ns range) and consumes lower power (4 μW/nm) [48] when compared to TO tuning (27 mW/FSR) [71], where FSR is the free-spectral range. However, only small shifts in an MR's resonant wavelength can be compensated using this mechanism (i.e., EO has a limited correction range). TO tuning is preferred to compensate for large shifts in MR's resonant wavelength. How-

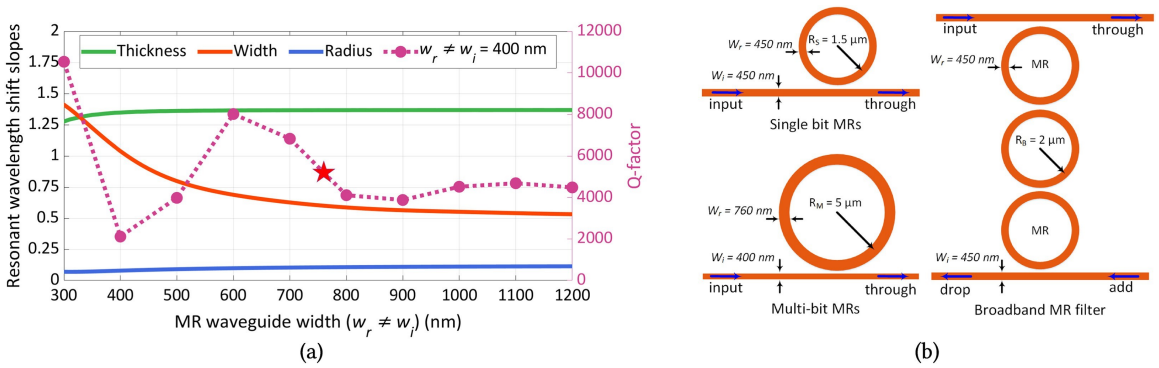


Figure 4.15: Tuning power compensation in a block of 10 MRRs placed with and without considering thermal eigen- mode decomposition (TED) for different MRR radius. The orange line represents phase crosstalk ratio variation with distance between MRRs.

ever, one has to compromise on latency (μs range) and power consumption, which is higher than for EO tuning. To reduce our reliance on TO tuning, which entails high overheads, we explore the possibility of a hybrid tuning mechanism, where both TO and EO tuning are used to compensate for $\Delta\lambda_{MRR}$. Such a tuning method has been proposed earlier [87] and can be easily transferred to an optimized MRR for hybrid tuning in our architecture. Such a mechanism would significantly reduce the overhead caused just by TO tuning.

To reduce the power overhead of TO tuning in such a hybrid approach, we adapt a method called thermal eigenmode decomposition (TED), which was first proposed in [88] that involves collectively tuning all the MRRs in an MRR bank. Such a tuning method proves to be more beneficial than individually tuning MRRs. By doing so we can cancel the effect of crosstalk (i.e., undesired phase shift) in MRRs with much lower power consumption. The amount of phase crosstalk induced from one MRR on another MR, placed adjacent to each other, can be modelled using the trend in Fig. 4.16 (orange line). In this figure, as the distance between two devices (MRRs) increases, the amount of phase crosstalk between them reduces. Correspondingly, as an example we calculate the tuning power compensation for an MRR bank consisting of 10 MRRs and different radii placed at a distance (d) from each other. A few important trends to observe from Fig. 4.16 are (i) as the radius of an MRR increases, tuning power compensation for $\Delta\lambda_{MRR}$ increases;(ii) Without TED (collective tuning of MRRs), the tuning power consumption is high, indicating that

each MRR would require more power to compensate for respective shifts in resonant wavelength ($\Delta\lambda_{MRR}$); (iii) By employing TED, we see a significant reduction in tuning power consumption: 51% (radius of 1.5 μm) and 41% (radius of 5 μm) when MRRs are placed at a distance of 5 μm and 7 μm apart from each other, respectively. Though placing MRRs further close to each other would yield better compensation in power, one must take into account the placement and routing of tuning circuit for each MRR in an MRR bank. Additional power reduction can be obtained by performing device level optimizations. Designing MRRs tolerant to FPVs would reduce the total power used to compensate for fabrication variations.

4.4.4.2 Device-Level Optimization

We explore different MRR designs to accommodate different needs in our *ROBIN* architecture such as multi-bit precision for activation values, single-bit precision for weight value representation, and batch normalization.

4.4.4.2.1 Fabrication-Process Variation Resilience FPVs cause undesirable changes in device critical dimensions (e.g., width and thickness), which cause resonant wavelength shifts ($\Delta\lambda_{MRR}$). To address $\Delta\lambda_{MRR}$, we explore the impact of change in device parameters such as waveguide width, thickness, gap between input and ring waveguide, and radius using our in-house MRR device-exploration tool. We map the behavior of different changes in the waveguide width, thickness, and radius in MRRs due to FPVs. Fig. 4.15(a) shows one of our design exploration results where we understand and observe the behavior of resonant resonant-wavelength shift slopes due to change variations in the waveguide width, thickness and radius represented by orange, green and blue lines respectively. Resonant wavelength shift slope due to change in waveguide width, thickness, and radius ($\partial\lambda_{MRR}/\partial w, t, R$) can be obtained from (3.10)

Fig. 4.15(a) clearly shows that the impact of resonant-wavelength shift reduces as we increase the waveguide width, whereas the impact of thickness and radius variations remains constant. $\Delta\lambda_{MRR}$ is more sensitive to changes in waveguide width, hence the impact of $\Delta\lambda_{MRR}$ reduces we increase the waveguide width. We employ Lumerical MODE [65], an Eigen mode solver

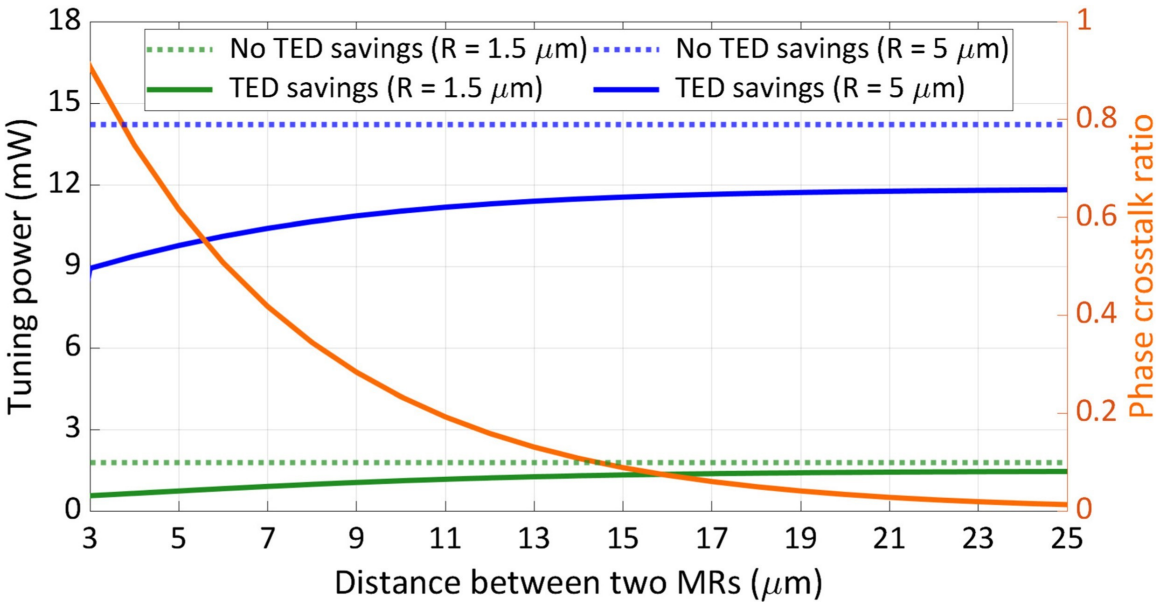


Figure 4.16: (a) Resonant-wavelength shift slopes with respect to changes in waveguide width, thickness, and radius, and corresponding cross-over coupling(κ), when the input waveguide (w_i) is set to 400 nm the marked point represents our selected MRR design; (b) The different MRR designs considered in this work.

to calculate these shifts in resonant wavelengths. One can easily overcome higher-order mode excitation by employing adiabatic designs [65] and waveguide tapers [110] in MRRs with wider waveguides. Such a design translates to lesser tuning-power consumption due to FPVs.

4.4.4.2.2 Multi-Bit Precision MRRs As discussed in Section 4.4.4, increasing the number of bits used to capture activations in a model can boost the model accuracy in our architecture. However, there is not a significant accuracy boost beyond 4-bit activation values, hence we explore MRR designs which can achieve a resolution of 4 bits. To achieve a resolution of 4 bits, we observe how the optical signals from MRRs impact each other due to crosstalk. We consider calculations from [102] to define the amount of noise from one MRR on the other:

$$\varphi(i,j) = \frac{\delta^2}{(\lambda_i - \lambda_j)^2 + \delta^2} \tag{4.15}$$

where, $\varphi(i,j)$ describes the noise content from the j^{th} MRR present in the signal from the i^{th} MR, $((\lambda_i - \lambda_j))$ is the difference between the resonant wavelengths i^{th} MRR and j^{th} MRR, and $\delta = \lambda_i/(2Q - factor)$. Where, Quality factor or Q-factor is a measure of the sharpness of the

resonance relative to the central frequency of a microring resonator (MRR) that impacts the optical channel spacing, crosstalk, bandwidth, and other factors in the MRR [12]. A sharper resonance (i.e., a higher Q-factor) can result in increased susceptibility to noise, as even a small change in the central frequency of the MRR (due to perturbance) can lead to large losses. This effectively limits the achievable resolution of the parameters being represented. Thus, smaller Q-factors are preferred. However, a smaller Q-factor can also lead to larger device dimensions and higher optical crosstalk, which in turn can lead to larger losses and higher tuning power requirements. Q-factor in an MRR is defined as follows:

$$\text{Q-factor} = \frac{\lambda_{MRR}}{\text{FWHM}}, \tag{4.16}$$

where, FWHM is the full width at half maximum of a resonance spectrum which can be defined for an all-pass ring resonator (see Fig. 4.15) as follows:

$$\text{FWHM} = \frac{(1 - ra)\lambda_{MRR}^2}{\pi n_{\text{g}} L\sqrt{ra}}, \tag{4.17}$$

where, r is the self-coupling coefficient and κ is the cross-over coupling coefficient. Also, a is the single-amplitude transmission, including both the propagation loss in the ring and the loss in the couplers, which can be written as $a = e^{-\alpha L}$,where α is power attenuation coefficient. L is round trip length or the circumference of the MRR. In this chapter, we assume a lossless coupler in our designed MRRs, hence $|\kappa|^2 + |r|^2 = 1$. For ideal cases with zero attenuation, a ≈ 1. Based on the above equations, the noise power component can thus be calculated using (4.13), and the resolution can be calculated using (4.14). To achieve a bit resolution of at least 4-bits, we need MRRs with a Q-factor of ≈5000 (from (4.11)) while being tolerant to FPVs. Q-factor is highly sensitive to losses and change in dimensions of MRR. Selecting input waveguide width of 400 nm and ring waveguide width of 760 nm, and radius (RM) of 5 μm as shown in Fig. 4.15(a) (blue line), provides improved tolerance to FPV, desirable Q-factor, and smaller area consumption. Such an MRR design with Q-factor of 5000, allows enough levels of distinction between bits by slightly

changing intensity, helps easily detect optical signal at the output port satisfying the requirement for multi-bit precision of activation values.

4.4.4.2.3 Single-bit MRRs In our architecture, we represent weight values with a single bit, and this requires just two levels of precision with an MRR. An MRR of high Q-factor may be used here, as we don't have to have high resolution here. Compact ring designs with high Q-factor have been proposed in [53, 111]. The work in [111] proposes an MRR design with radius 1.5 μm to achieve a high Q-factor of 46,000 without the consideration of sidewall roughness while maintaining low bending loss $\approx$7 cm^{-1}. Similarly, an adiabatic MRR structure of radius 3 μm is designed in [53] to avoid higher order mode excitation where a high Q-factor of 27,000 is achieved. These works prove our point that such high Q-factor rings can be designed.

We design a ring of radius 1.5 μm, as shown in Fig. 4.15(b),with input waveguide (w_i) and ring waveguide (w_r) width set to 450 nm, to achieve a Q-factor of 25,000 that corresponds to a bit resolution of 1 from (4.11). These designs allow our architecture to save on area and tuning power consumption (). We acknowledge that FPVs are an inevitable part of the fabrication-process. However, since we just need to differentiate between two levels of operations, we do not explore for designs that are tolerant towards FPVs, for single-bit MRRs.

4.4.4.2.4 Broadband MRRs Batch normalization (BN) layers can be considered essential in BNNs as they add complexity to the models, via the gain (γ) and bias (β) terms of the layer. These terms are learned during the training process along with the normalization parameters of the batch mean (μ) and standard deviation (σ). As they are being learned in training, these terms are dynamic, but during inference they have static values. This allows us to have a photonic version of batch normalization folding, where we may tune weights as per the following equation:

$$w_{\text{fold}} = \gamma \cdot \frac{W}{\sqrt{\sigma^2 + \epsilon}} = C_{\text{fold}} \cdot W \tag{4.18}$$

There is a similar equation for bias terms as well, but since BNN models benefit from batch normalization after every layer, these will be normalized out and hence can be ignored. The above constant, C_{fold} is applied to every weight term and hence is a participant in every matrix multiplication operation, i.e.:

$$\text{Input}_{l+1} = f\left(A_l \cdot (w_{\text{fold}})_l\right) = C_{\text{fold}} f(A_l \cdot W_l). \tag{4.19}$$

In (4.19), Input_{l+1} refers to the input to the $(l+1)^{th}$ layer, $f()$ is non-linear activation function, A_l is the activation of l^{th} layer and W_l is the weights from l^{th} layer. This operation can be applied to partial sums as well, and can be implemented using a broadband photonic device with its gain tuned to reflect C_{fold}. The broadband device is preferred as this allows simultaneous gain tuning of all the wavelengths in the waveguide efficiently, both area and energy wise. Hence, the last type of MRRs we consider are broadband MRRs that are needed for batch normalization (BN) layers due to their relevance in BNNs. A large passband can be achieved by cascading several MRRs and properly selecting the design parameters of MRRs [112]. We explore such a higher order MR, or cascaded MRR filter, to achieve a wide passband. [113] explores a possibility for passband widths ranging from 6.25 GHz to a maximum of 3 Thz. This work explores different design parameters of a higher-order filter while evaluating different losses such as insertion, propagation and coupling loss in higher order MRRs. A 0.5 nm resonant wavelength shift of MRR was reported for a fabrication error of 10 nm showing that such a design is tolerant to FPVs. A 3rd-order MRR based switching device with radius of 2 μm shown in Fig. 4.15(b), fits the requirement for broadband MRR. The coupling coefficients at the input (κ_i^2) is 0.53 and coupling at higher order rings is 0.2. The propagation loss of 25 dB/cm has been reported and insertion loss of the two elements in higher order filter are 4.35 dB and 0.36 dB, respectively [112]. Having such a design, one can achieve a flat-top passband with bandwidth width of at least 3 THz. Employing such a broadband MRR can help us apply the non-linear activation on all the available resonant wavelengths in the bank. Having a large bandwidth such as 2.5 Thz allows us to conveniently tune up to 20 different wavelengths.

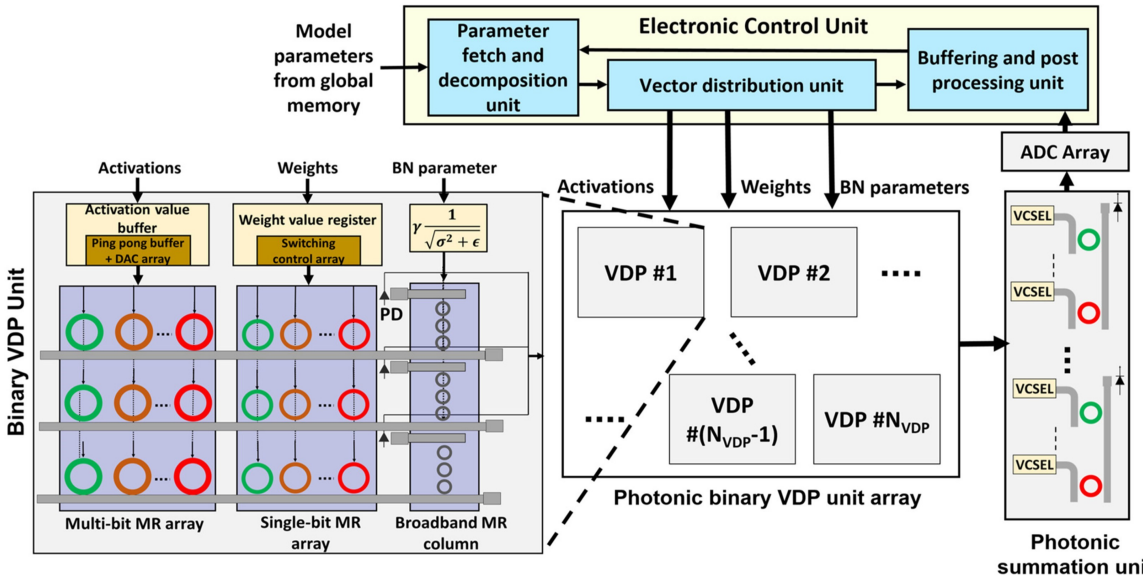


Figure 4.17: An overview of the *ROBIN* architecture, showing the electronic control unit, the photonic vector dot product (VDP) unit array, and the photonic summation unit, along with a detailed view of the VDP unit internal structure.

4.4.4.3 Architecture Design

An overview of the *ROBIN* accelerator architecture is shown in Fig. 4.17. The optical device and tuning circuit optimizations from the previous subsections are utilized within the optical binary vector dot product (VDP) units. We use banks of heterogeneous MRRs to imprint activations, weights, and the BN layer constants onto optical signals. Multiple such VDP units are composed together to form the overall architecture, as shown in the figure, which is then used to accelerate a given BNN model. We utilize a photonic summation unit for summing the partial sum outputs from our VDPs, before passing the partial sums on to the electronic control unit (ECU), as shown in Fig. 4.17. We also rely on the ECU for fetching parameters from the global memory, decomposing them to lower dimensional vectors which can be obtained from 4.3.3.3.2, distributing these vectors among the VDP units, and for implementing non- linear activations functions and pooling layers. We describe the working of the *ROBIN* architecture in more detail in the following subsections.

4.4.4.3.1 Vector dot Product (VDP) Unit Design As discussed in Section 4.3.3.3.2, we consider matrix operations to be decomposed to lower dimensional vector dot products. These vector

dot product operations are executed optically within our VDP units. The heterogeneous MRR designs combined with optical circuit-level optimizations for area and power consumption are utilized to design VDP units (Fig. 4.17) suited for accelerating both CONV and FC layers without compromising on accelerator throughput. For representing weight values, we use high Q-factor, small radius single-bit MRRs described in Section 4.4.4.2.2. The smaller radius contributes to lower tuning power and reduces optical losses along the waveguide. This is possible due to the binarized nature of weight matrices in BNNs. For activation values we consider MRRs with slightly lower Q-factor, for better resolution, as discussed in Section 4.4.4.2.1. For the BN layer we have to simultaneously tune all the wavelengths in the waveguide to the batch normalization constant, and for this we use third order MRR filters, as described in Section 5.2.3. The combination of these heterogeneous designs allows the VDP units to be highly energy efficient. The VDP units also perform BN layer folding, which can be done efficiently in the optical domain, as discussed in Section 4.4.4.2.3. We also make use of electronic buffering in the VDP units to reduce the digital to analog converter (DAC) usage. In particular, we make use of ping-pong buffers, which allow us to use a single DAC array to feed the activation devices in all the waveguides in a VDP unit. As weight values are single-bit values, we can use simple switching circuits to essentially turn the MRRs on or off depending on the value of the weight parameters.

In designing a VDP unit, there are several important parameters that must be carefully considered: number of higher resolution MRRs for activation representation (N_A), number of single-bit MRRs for weight representation (N_W), and number of broadband MRRs (N_B) for batch normalization folding implementation. Thus, the total number of MRRs per waveguide N_{MRR} = N_A + N_W + N_B . The number of required DACs is equal to N_A. The number of waveguides to which we distribute the MRRs is denoted as N_{WG} . The size of the vector represented in the VDP unit is given by $N_{WG} \times N_A$. We divide this vector across multiple waveguides to reduce power consumption, as this allows us to reuse wavelengths and reduce the overall laser power consumption, as discussed next, in Section 5.3.3. For efficient parallelization and to increase the throughput of the accelerator, multiple VDP units work concurrently on parameters from the same layer and

generate partial sums simultaneously. The total VDP unit count used in *ROBIN* is N_{VDP}. Thus, the VDP and architecture design process can be considered as an optimization problem where we try to explore N_{VDP}, N_{WG}, N_A (= N_W), and N_B values while trying to maximize throughput and minimize area and power consumption. We present results of this architecture exploration analysis in Section 4.4.5.3.

4.4.4.3.2 Optical Wavelength Reuse in VDP Units Prior works on optical accelerator design typically considers a separate wavelength to represent each individual element of a vector. This approach leads to an increase in the total number of lasers needed in the laser bank (as the size of the vectors increases) which in turn increases power consumption. Beyond employing the decomposition approach discussed above, we also consider wavelength reuse per VDP unit to minimize laser power. In this approach, within VDP units, the vectors assigned from the electronic control unit (ECU) are further decomposed into smaller sized vectors for which dot products can be performed using MRRs in parallel, in each arm of the VDP unit. The same wavelengths can then be reused across arms within a VDP to reduce the number of unique wavelengths required from the laser. Photodetectors (PDs) perform summation of the element-wise products to generate partial sums from decomposed vector dot products. The partial sums from the decomposed operations are then converted back to the optical domain by VCSELs (bottom right of Fig. 4.17), multiplexed into a single waveguide, and accumulated using another PD, before being sent for buffering. Thus, our approach leads to an increase in the number of PDs compared to other accelerators but significantly reduces both the number of MRRs per waveguide and the overall laser power consumption. The reduction in overall power consumption is also assisted by the fact that PDs do not consume significant power.

In each arm within a VDP unit, we can use a maximum of 15 MRRs per bank for a total of 30 MRRs per arm. The choice of MRRs per arm considers not only the thermal crosstalk and layout spacing issues and the benefits of wavelength reuse (as discussed earlier), but also the fact that optical splitter losses become non-negligible as the number of MRRs per arm increase, which in

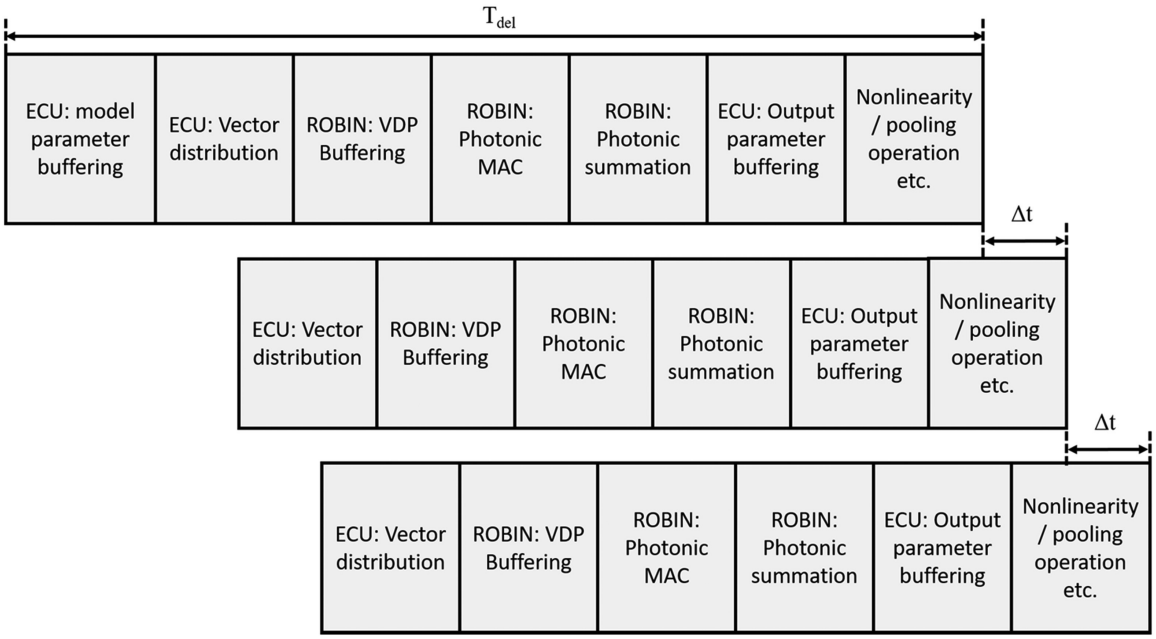


Figure 4.18: Pipelined scheduling of operations during BNN execution on the *ROBIN* accelerator.

turn increases laser power requirements. Thus, the selection of MRRs per arm within a VDP unit must be carefully adjusted to balance parallelism within/across arms, and laser power overheads.

4.4.4.3.3 *ROBIN* Pipeline and Scheduling The pipeline and schedule of operations during BNN model execution on the *ROBIN* accelerator is shown in Fig. 4.18. The electronic control unit (ECU) for the accelerator communicates with the global memory and retrieves the trained weights for the model being accelerated. The weights are stored in SRAM-based buffers. Considering the vec- tor granularity of the VDP units, latency of operation of the photonic core, and the parameter sizes (4-bit activation bits and binary weight parameters), we can calculate the memory bandwidth necessary. From our analyses (presented in Section Section 4.4.5.3), we found that our architecture needs a maximum bandwidth of 93.75 GB/s at the ECU to photonic core interface. This is a reasonable bandwidth assumption for an SRAM-based memory with operating frequency $\geq$ 2.5 GHz and a read width of 250 bits. Previous works, such as [114], have explored similar SRAM systems, but for a much higher bandwidth requirement at 250 GB/s. The lower bandwidth requirement for our system can be attributed to the smaller parameter sizes, while the work in [114] considered 16-bit precision for the neural network parameters. Memory interfaces which exceed

the bandwidth necessary are already available commercially: e.g., NVIDIA Tesla K20M GPUs have 320-bit memory interfaces at 2.6 GHz which can operate every half clock cycle to provide a bandwidth of 208 GB/s.

These weight matrices are decomposed to lower dimensional vectors and are distributed to the VDPs by the ECU's vector decomposition unit. The decomposition operation is described by the left-hand side of Equations (4.22) to (4.24). As described in the equations, the vector decomposition unit converts matrices to vectors (row-wise conversion for weight matrices and column-wise con- version for column matrices), and then those vectors into sub vectors. The size of the sub vectors depends on the granularity of the VDP units. The received vectors are buffered in the VDP units and are fed into the DAC array through a ping-pong buffer so that they can keep the MAC operation running continuously. The partial sums generated are passed on to the photonic summation unit, the output from which is passed on to the ECU. The ECU buffers the sums and calculates inputs that are then passed on to the next layer by subjecting the parameters to non-linearities (activation functions) and performing other layer specific operations, like pooling.

The model parameter buffering stage is not repeated every pipeline operation, but must be repeated as the parameters buffered in the buffers in ECU are depleted (i.e., distributed to VDP units). As such, the total time required by *ROBIN* to perform inference acceleration for a given model can be given as:

$$\textit{Total time of operation} = T_{del} + \Delta t \times X + \textit{(ECU parameter buffering delay)} \times x, \tag{4.20}$$

where,

$$\Delta t = \textit{local buffer operation delay} + \textit{vector distribution delay}, \tag{4.21}$$

$$X = \frac{\textit{Total number of parameters in the model}}{N_w \times N_{VDP}}, \tag{4.22}$$

$$x = \frac{\textit{(Parameters buffered in ECU)}}{N_w \times N_{VDP}}, \tag{4.23}$$

Table 4.5: Models and Datasets used for Evaluations

Model no	CONV layers	FC layers	BN layers	Parameters	Datasets
1	2	2	3	60,642	Sign MNIST
2	6	3	6	1,546,570	CIFAR 10
3	6	3	7	13,570,186	STL10
4	6	2	6	552, 362	SVHN

Comparing our pipeline to the pipeline presented in the previous work on photonic BNN acceleration, [106], we can observe the following differences: (i) ROBIN's pipeline takes into consideration model parameter retrieval from global memory, buffering in the ECU, and how these parameters are utilized in the photonic core. The pipeline in [106] does not include these operations in its pipeline; (ii) ROBIN's pipeline considers both ECU and photonic core operation, whereas the pipeline in [106] is photonic system centric; (iii) *ROBIN* utilizes photonic batch normalization folding which does not require an extra step, whereas in [106] this operation is performed electronically and requires a separate stage in their pipeline.

4.4.5 Results

4.4.5.1 Simulation Setup

We conducted several simulation studies to evaluate the effectiveness of our BNN accelerator. The optimized heterogeneous MRR designs, the tuning circuit optimizations, and architectural level considerations discussed so far were included in our simulation considerations.

We implemented and simulated the *ROBIN* architecture using a custom Python simulator to estimate its performance in terms of power, frames per second (FPS) performance, and energy consumption. For analyzing the inference accuracy across different activation precision and the impact of FPV noise on the inference accuracy, we used Tensorflow 2.3 along with Qkeras [92]. Fig. 4.19 shows the training accuracy versus epoch graph of the models described in Table 4.5, to illustrate the accuracy and loss across the epochs.

We compare *ROBIN* with DEAP-CNN [82] and HolyLight [83], two recent optical DNN accelerators from prior work, along with LightBulb [106], which is an optical BNN accelerator, as

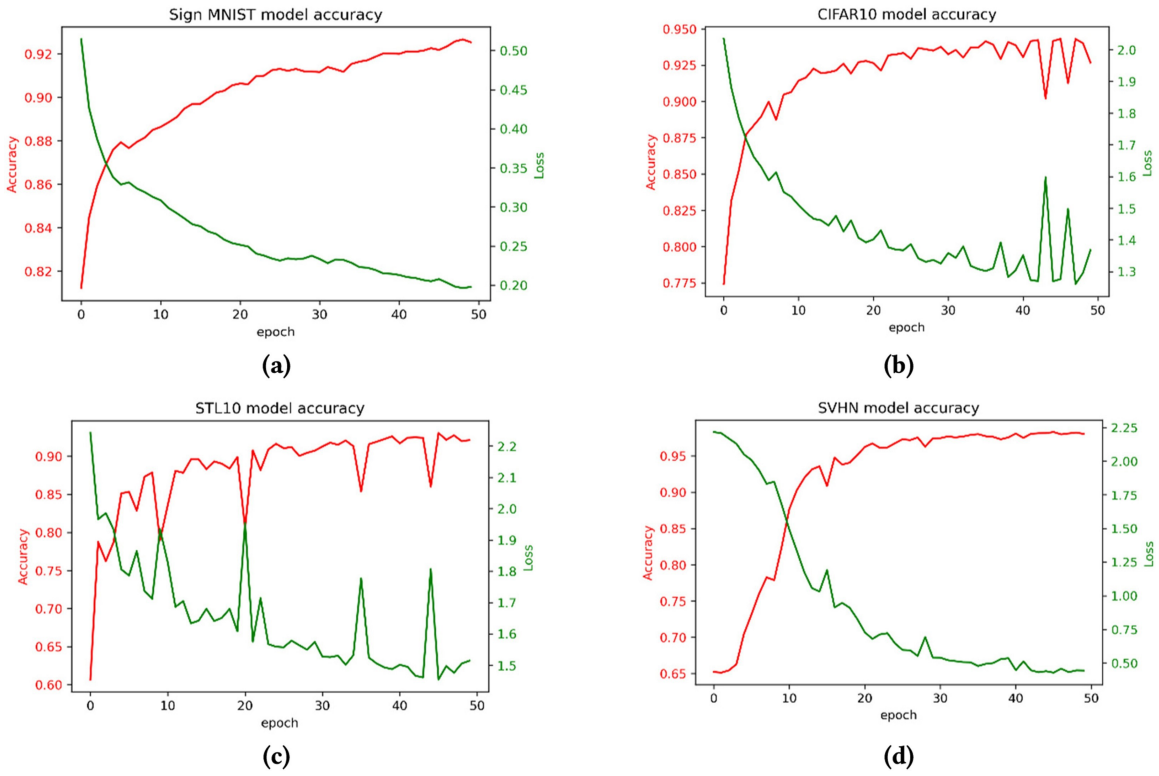


Figure 4.19: The training accuracy vs epoch for the BNN models considered for (a) Sign MNIST, (b) CIFAR10,(c) STL10, and (d) SVHN datasets. (a) shows top-1 accuracy, while (b)-(d) show top-5 accuracy.

Table 4.6: Parameters Considered for Analysis of Photonic Accelerators

Devices	Latency	Power
EO Tuning [48]	20ns	4 μW/nm
TO Tuning [71]	4 μs	27.5 mW/FSR
VCSEL [93]	10 ns	0.66 mW
TIA [94]	0.15 ns	7.2 mW
Photodetctor [95]	5.8 ps	2.8 mW
DAC [116]	0.33 ns	59.7 mW
ADC [115]	24 ns	62 mW

well as numbers reported from several electronic DNN and BNN accelerators. For the optical accelerators, we considered optical signal losses due to various factors: signal propagation loss (1 dB/cm [42]), splitter loss (0.13 dB [96]), combiner loss (0.9 dB [97]), MRR through loss (0.02 dB [98]), MRR modulation loss (0.72 dB [99]), microdisk loss (1.22 dB [100]), EO tuning loss (6 dB/cm [48]), and TO tuning loss (1 dB/cm [71]). We also considered the ADC design from [115] and the 4-bit DAC from [116] in our analyses. The analysis of the optical accelerators (DEAP-CNN [82], HolyLight [83], and LightBulb [106]) follows the modeling methodology we have adopted for *ROBIN*, where we factor in power consumption and delays associated with photonic devices used in these accelerators. A summary of the power and latency considerations for our analyses is given in Table 4.6. These power and latency values were used in our simulations and latency of operation of our architecture. A comparison for inference time on *ROBIN* and a conventional CPU is presented in Section 4.4.5.5. To calculate laser power consumption, we use the following power model given in (4.11):

4.4.5.2 Fabrication-Process Variation Analysis

FPV in optical devices is corrected using TED tuning in our architecture, as discussed in Section 5. At the system level, this tuning leads to significant power consumption overhead, and any avenue to further reduce tuning power consumption becomes important. To this end we conduct an FPV noise injection analysis, where we inject noise, modeled using FPV data, into the MRR devices into our *ROBIN* accelerator, during the inference phase. This experiment was conducted to (i) study the impact of FPV induced noise on BNN models mapped to our accelerator, (ii) deter-

mine how effective TED tuning is in such scenarios, and (iii) uncover any opportunities for further power minimization.

To analyze the impact of FPV on the model and how TED tuning compensates for it, we first consider the effect of FPV on the shift in resonant wavelength ($\Delta\lambda_{MRR}$) in MRRs. Resonant-wavelength shift in an MRR can be modeled from (3.9). We generate virtual FPV maps with a mean (μ) of 0 and standard deviation (σ(w,t,R)) of 4.9 nm, 1.5 nm, and 0.75 nm for waveguide width, thickness, and radius, respectively. These standard deviation values are experimentally obtained based on real fabricated MRR devices through our collaboration with CEA-Leti. Using these values, we are able to derive $\Delta\lambda_{MRR}$ using (3.9). So, the current resonant wavelength (λ'_{MRR}) of the FPV affected MRR becomes:

$$\lambda'_{MRR} = \lambda_{MRR} + \Delta\lambda_{MRR}, \tag{4.24}$$

Due to a shift in λ_{MRR} , the transmission of the wavelength through the MRR is impacted. The inten- sity of the wavelength at the through port is given by the following equation from [12].

$$T = \frac{I_{out}}{I_{in}} = \frac{a^2 - 2rcos\phi + r^2}{1 - 2arcos\phi + ra^2}, \tag{4.25}$$

In (4.25), ϕ = βL, with L being the roundtrip length and β the propagation constant $\beta = 2\pi/\lambda$ of the circulating mode; and r^2 is the self-coupling coefficient of an MRR. A detailed analyses for the calculation of r using super mode theory is presented in [44]. The output intensity from the MRR is important, as for noncoherent MAC units, the parameter values are encoded onto the signal intensity, and a change in expected output can be seen as perturbation or noise source.

The noise injection was modeled using Equations (3.9) and (4.25), where we consider the resonant- wavelength shift ($\Delta\lambda_{MRR}$) in MRRs due to FPV and its impact on the parameters imprinted on the MRRs. From our analysis using the FPV data from our device fabrications with CEA-Leti, and Equation (4.23), we are able to obtain the mean and standard deviation values for $\Delta\lambda_{MRR}$ in a wafer. The values calculated are μ = -0.1461 nm and σ = 24.417 nm. Using these

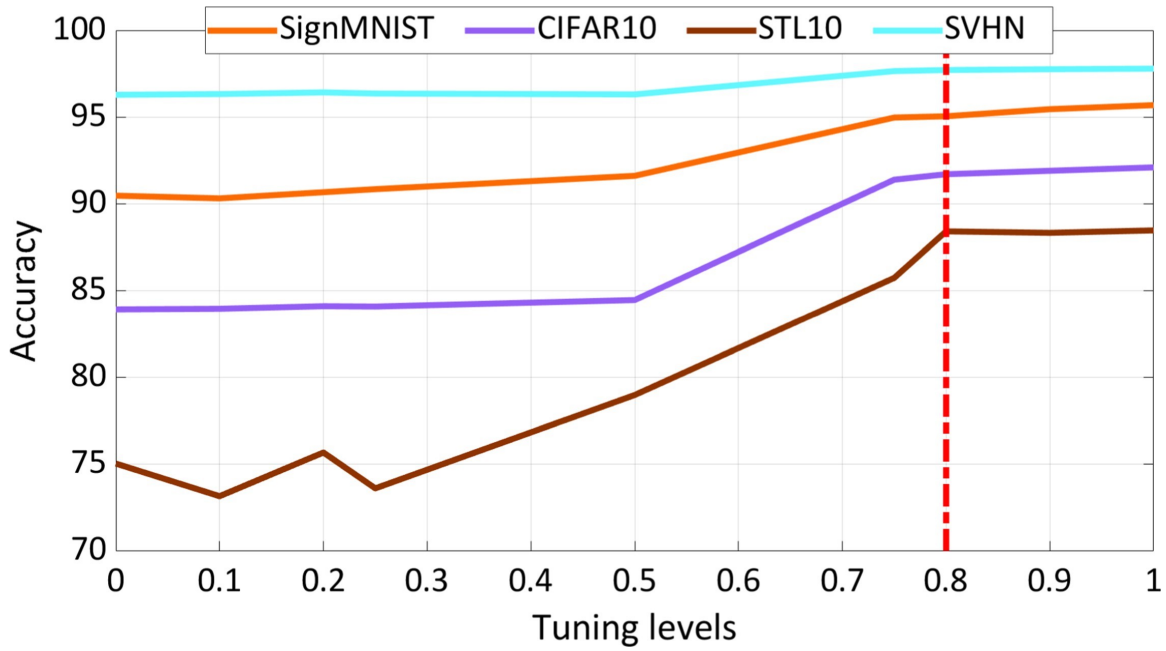


Figure 4.20: Inference accuracy versus level of tuning applied. At 80% tuning, the inference accuracy saturates, rendering further tuning unnecessary, and providing an opportunity to save tuning power.

values, 50 $\Delta\lambda_{MRR}$ maps for the accelerator were generated and then using Equation (4.25) the perturbation to the parameters imprinted on to the devices were modeled. Noise injection to the models was performed at inference time using Tensorflow.

Fig. 4.20 shows the results of this experiment, where we explored the impact of FPV-induced noise in the four BNN models, and the effect of TED tuning for FPV compensation. We expected that the better the devices were tuned, the better the accuracy that would be exhibited by the accelerator. But it was observed that the model's accuracy can be sustained without completely (perfectly) tuning the devices. Fig. 4.20 shows that at 80% FPV correction through tuning, the BNN retains appreciable inference accuracy. Thus, there is not a significant accuracy benefit to tune beyond the 80% level, and staying with the 80% level can allow us to achieve power savings. This reduction in tuning power is factored into our architecture level analysis, which is presented next.

4.4.5.3 *ROBIN* Architecture Optimization Analysis

In this section, we show results of our exploration of the parameters discussed in Section 4.4.4.3.1. As mentioned in Section 4.4.4.3.1, we try to optimize N_{VDP}, N_{WG}, N_A, and N_B to reduce area and power consumption while trying to obtain the best throughput (frames per second or FPS) possible. NB was fixed to be 1 per waveguide, allowing us to have up to 20 wavelengths in the same waveguide with a channel spacing of 1 nm, which in turn allows us to tune all the MRRs simultaneously to the BN layer parameters. We then explored N_{VDP}, N_{WG}, and N_A, with the goal of optimizing power, area, and FPS. The result of this exploration analysis is shown in Fig. 4.21 in the form of a scatter plot. From this analysis we identified two configurations for *ROBIN*, where one is optimized for FPS/Watt, with lowest area and power consumption (energy optimized *ROBIN* or *ROBIN-EO*), and another with the best FPS but with higher area and power consumption (performance optimized *ROBIN* or *ROBIN-PO*). In terms of (N_A, N_{VDP}, N_{WG}), these configurations can be represented as (10, 50, 10) for *ROBIN-EO* and (50, 200, 10) for *ROBIN-PO*. To explore the efficiency of these configurations, they are compared against other optical and electronic DNN/BNN accelerator platforms. Results for these comparisons with other accelerators are presented in the following section.

4.4.5.4 Comparison with State-of-the-art Optical and Electronic DNN/BNN Accelerators

We compared *ROBIN-EO* and *ROBIN-PO* against various electronic and optical neural network acceleration platforms. For optical DNN accelerator platforms, we selected DEAP-CNN [82] and HolyLight [83]. For electronic DNN accelerator platforms, we compare against a GPU (Nvidia Tesla P100), along with several architectures including SIGMA [117], Edge TPU [118], DaDianNao [119], and FPGA implementation of Null Hop [120]. We also compare *ROBIN* against the best-known previous work on photonic BNN accelerators, LightBulb [106]. We also compare *ROBIN* against the best-known previous work on photonic BNN accelerators, LightBulb [106].

When compared to LightBulb, *ROBIN* has the following differences: (i) *ROBIN* is designed to accelerate partially binarized neural networks, as opposed to fully binarized neural networks as in [106], for obtaining better accuracies; (ii) *ROBIN* utilizes photonic batch normalization folding

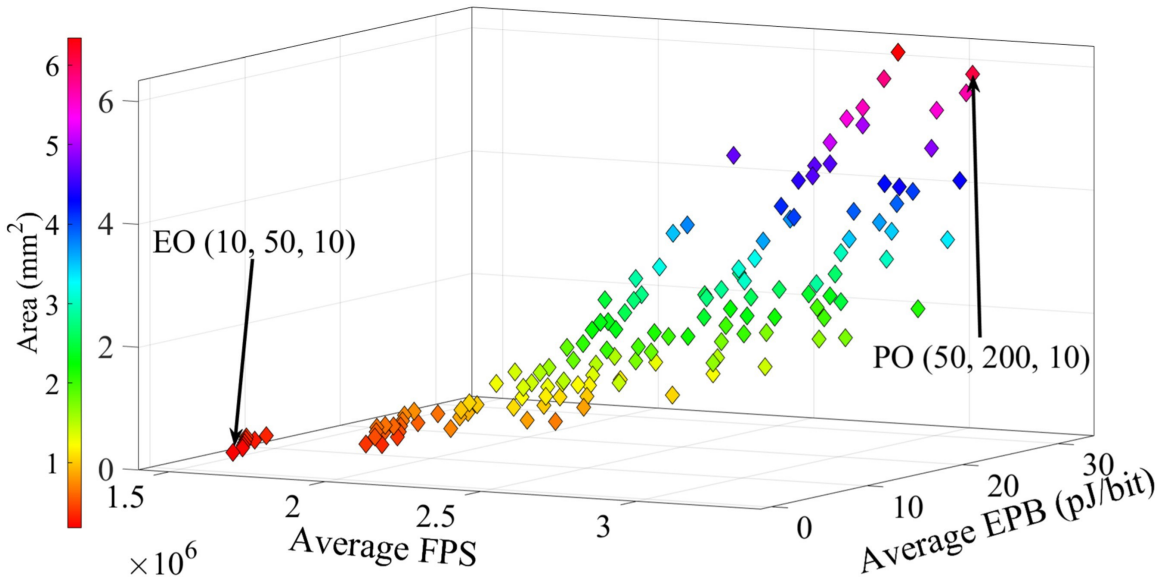


Figure 4.21: Scatterplot of average FPS vs. average EPB vs. area of various *ROBIN* configurations. The configuration with highest FPS/Watt (energy optimized or EO) and the one with best FPS (performance optimized or PO) are specified.

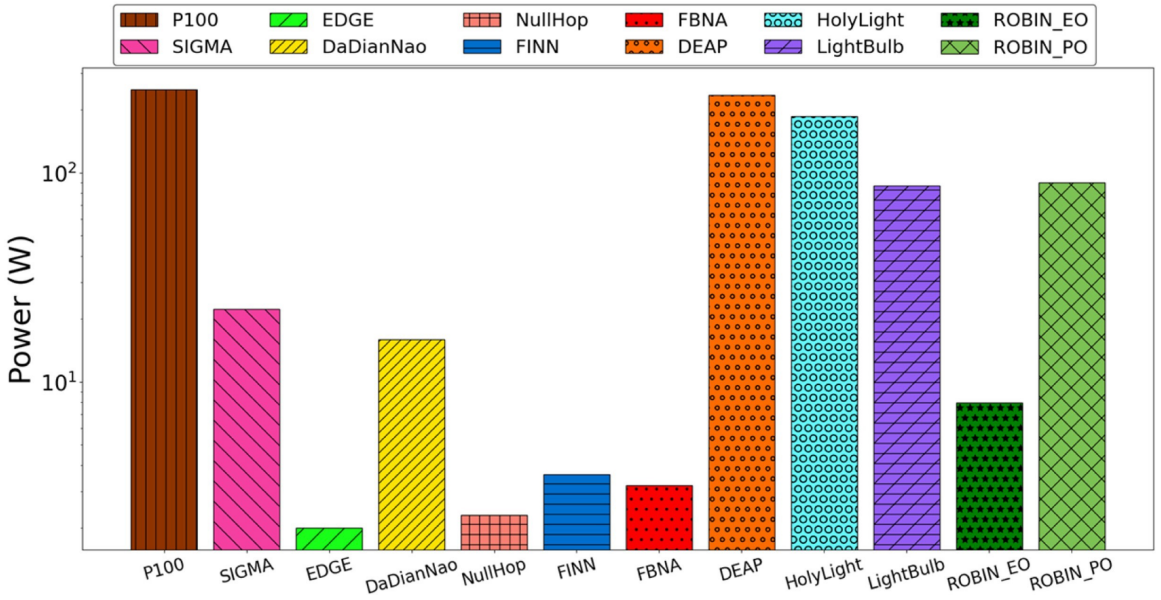


Figure 4.22: Power consumption comparison among variants of *ROBIN* versus other optical accelerators (DEAP- CNN, Holylight, LightBulb), and electronic accelerator platforms (P100, SIGMA, EdgeTPU, DaDianNao, Null Hop, FINN, and FBNA).

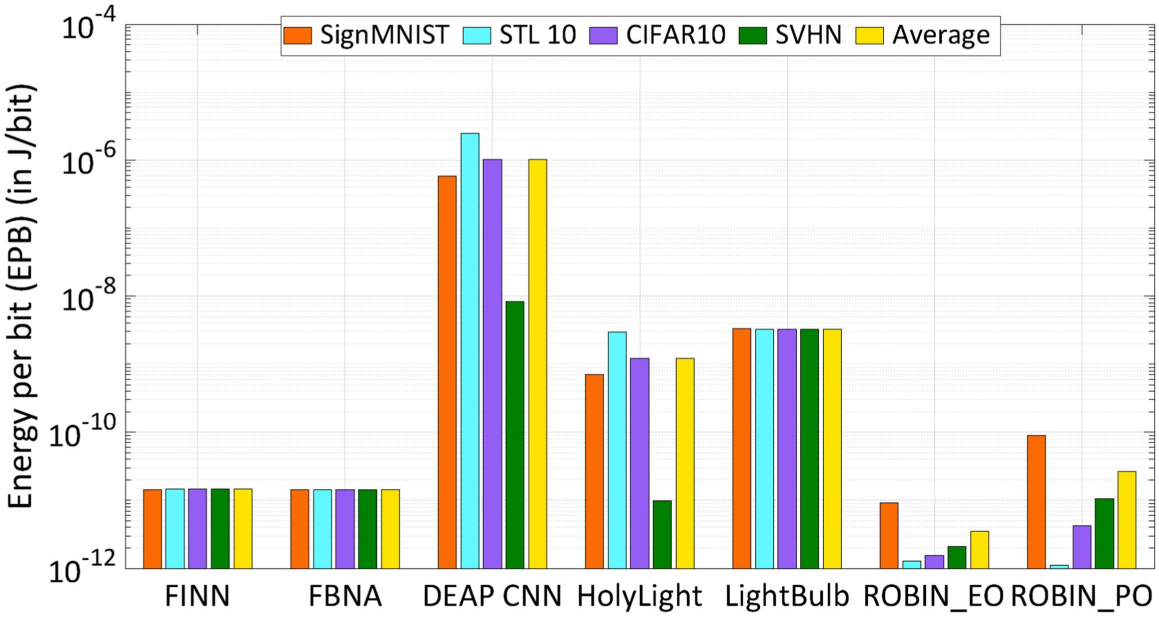


Figure 4.23: EPB comparison between electrical BNN accelerators, optical accelerators, and the *ROBIN* variants.

for faster, energy efficient batch normalization layer operation whereas [106] relies on an electronic implementation of the batch normalization operation; (iii) *ROBIN* has various circuit- and device-level optimizations in place to counteract thermal and process variations, which also ensure high throughput and energy efficient operation; whereas [106] does not take into account thermal and process variations and the necessary tuning latency and energy consumption overheads needed to counter them; (iv) architecture-level optimizations in *ROBIN* ensure lower power consumption in terms of tuning and laser power; these considerations are not part of the architecture proposed in [106]. We also compare against electronic BNN accelerators FBNA [121] and FINN [122]. We used the GOPS and power consumption parameters from [103] and [42] to simulate inference on the electronic platforms.

Fig. 4.22 shows the power comparison across the accelerators from prior work and the two *ROBIN* variants. It can be observed that *ROBIN-PO* has substantially higher power consumption than *ROBIN-EO*, as *ROBIN-PO* is focused on FPS performance rather than energy conservation. *ROBIN-PO* has a much larger vector granularity per VDP unit along with substantially higher VDP unit count to maximize parallelism, when compared to *ROBIN-EO*. The larger unit count and the

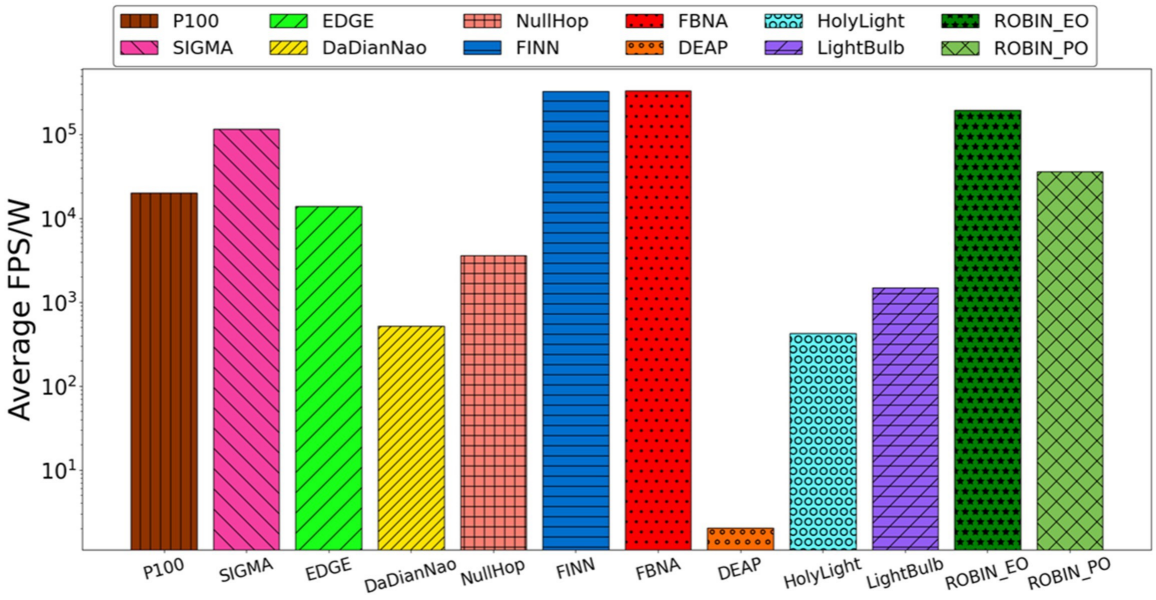


Figure 4.24: Average FPS/Watt among different accelerator platforms, visualized.

waveguide count in *ROBIN-PO* drives its power requirements higher. On the other hand, it can be observed that the energy and area efficient *ROBIN-EO* has comparable power consumption to that of edge and mobile electronic neural network accelerators.

In Fig. 4.23, we compare the energy-per-bit values (EPB) across the various BNN accelerators considered in this work. We can observe that both the *ROBIN* variants perform significantly better than the optical accelerators in comparison. This lower EPB is owing to the meticulous device, circuit, and architecture level optimizations we have considered in our architecture, which takes into account various losses and delays at the architecture level and counteracts them. The heterogeneous MRRs used in *ROBIN* provide energy and area benefits, and the utilization of TED for collectively tuning MRRs provides further energy benefits on top of the 20% reduction we obtained from the analysis in Section 4.4.5.3. TED also allows for closer placement of MRRs, which in turn helps reduce propagation delays. This reduction is also impacted by the faster inputs to DAC arrays enabled by local buffering and ping-pong buffers in the VDP units.

Lastly, in Fig. 4.24 we present the average FPS/Watt comparison between the various accelerator platforms. Both the *ROBIN* variants perform well against the accelerator platforms to which they were compared against. *ROBIN-EO* outperforms all other platforms other than FBNA and

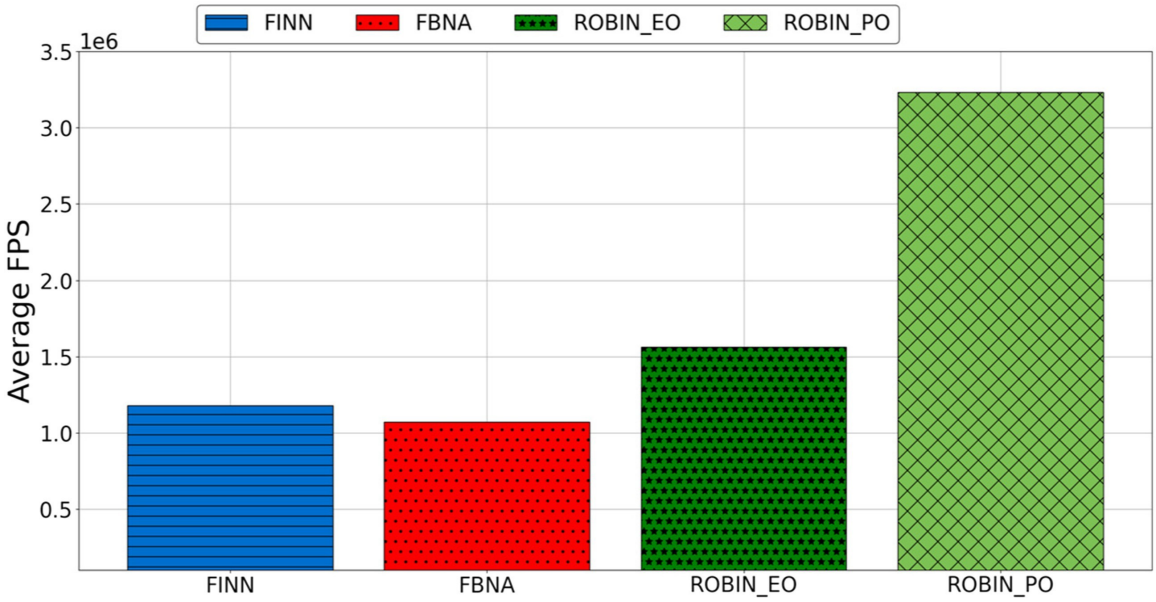


Figure 4.25: FPS comparison between the *ROBIN* variants and the electronic BNN accelerators.

FINN. This is owing to the extremely low power consumption reported by these BNN accelerators. However, the *ROBIN* variants display superior FPS performance with respect to these electronic accelerators, as can be seen in Fig. 4.25.

In summary, *ROBIN* showcases the effectiveness of cross-layer design of BNN accelerators with the emerging silicon photonics technology for energy/area efficient implementations and for performance-oriented designs. Overall, we can see that our energy efficient design (*ROBIN-EO*) exhibits EPB values $\approx$4$\times$ lower than electronic BNN accelerators and $\approx$393$\times$ lower than the photonic BNN accelerator, while the performance-oriented design (*ROBIN-PO*) shows $\approx$3$\times$ and 25$\times$ better FPS than the electronic and photonic BNN accelerators respectively. With the growing maturity of silicon photonic device fabrication in CMOS-compatible processes, it is expected that the energy costs of device tuning, losses, and laser power overheads will go further down, making an even stronger case for considering optical-domain accelerators for deep learning inference.

4.4.5.5 Comparison to CPU Based Inference

To highlight the advantage of dedicated inference acceleration, we have compared the performance of our *ROBIN* architecture against a standard desktop CPU performing inference on these

Table 4.7: Inference time on *ROBIN-PO* and Intel i7 Desktop for the Four Models

Model no.	Parameters	Datasets	Inference time (for one image)	
			ROBIN-PO	**i7-4790**
1	60,642	Sign MNIST	0.0218 μs	0.16ms
2	1,546,570	CIFAR 10	0.28 μs	1.75 ms
3	13,570,186	STL10	2.3 μs	2.5 ms
4	552, 362	SVHN	0.11 μs	1.25 ms

models. The CPU we have considered is an Intel i7-4790, and we have used Tensorflow to analyze the latency for inference. The CPU, i7-4790, is reported to have an average power consumption of approximately 103 W. This power consumption is comparable to the 90 W we report for the *ROBIN-PO* variant. The summary of observations for inference time are shown in Table 4.7. It can be clearly observed that the *ROBIN* accelerator provides several orders of magnitude reduction in inference time for all the four models and datasets, compared to the Intel i7 system.

4.4.6 Conclusion

In this section we proposed *ROBIN*, an optical-domain BNN accelerator which utilizes device-level, circuit-level and architecture-level optimizations to save on energy and area while improving overall throughput. Through our optimization efforts we identified two variants of ROBIN: *ROBIN-EO*, which is optimized for energy and area efficiency, and *ROBIN-PO*, which exhibits higher FPS performance, at the expense of greater power consumption. Our simulation analysis showed that *ROBIN* exhibits significantly better EPB performance than the various state-of-the-art optical neural network accelerators. Owing to significantly lower power consumption reported by the electronic BNN accelerators considered, *ROBIN* variants are not able to obtain better FPS/Watt than them, but upon closer examination both *ROBIN* variants can be seen to have better throughput than the electronic BNN accelerators. These results highlight the promise of our proposed *ROBIN* accelerator for accelerating BNN model execution for resource-constrained platforms. The work described here is focused on BNN acceleration using photonic systems. We considered how photonic systems can be used to accelerate the partially binarized networks, with weights remaining binary, while activations being multi-bit parameters. An immediate consideration for extension

is to employ mixed quantization in the models considered, where different layers have different levels of quantization for their activation parameters. This can enable better accuracy for the considered BNN models. The photonic system- and device-level optimizations are not limited to BNN inference accelerators. These techniques may also be considered for other non-BNN accelerators for DNN/CNNs as well.

4.5 Characterization and Optimization of Coherent MZI-based Nanophotonic Neural Networks under Fabrication Non-Uniformity[3]

The main contribution in the section of this chapter is in developing, to the best of our knowledge, the first comprehensive analysis of the impact of physical-level FPVs on coherent SPNN performance. We consider variations in the waveguide width, silicon-on-insulator (SOI) thickness, and etch depth, which impacts the slab thickness, based on realistic FPV maps developed using different correlation lengths in the variations and experimental measurements. We model the impact of FPVs at the device level for MZI performance and at the network level for arrays of cascaded MZIs in SPNNs. At the system level, we explore the impact of variations on SPNN inferencing accuracy while considering different FPVs and variation correlation lengths. Leveraging our detailed device-level models, we also develop a novel design optimization solution to improve MZI performance in SPNNs under different FPVs. Our simulation results for two example SPNNs (with 1380 and 20,580 phase shifters) under FPVs show that while inferencing accuracy can drop to as low as 7.73% under different variations, using our optimized MZIs can significantly improve the inferencing accuracy (e.g., by up to 72% on average in a large SPNN). Note that this chapter does not consider the impact of thermal crosstalk and FPVs in directional couplers in MZIs.

[3]A. Mirza, A. Shafiee, S. Banerjee, K. Chakrabarty, S. Pasricha, and M. Nikdast, "Characterization and optimization of coherent MZI-based nanophotonic neural networks under fabrication non-uniformity," IEEE Transactions on Nanotechnology (TNANO), vol. 21, pp. 763–771, 2022.

The rest of the section is organised as follows. Section 4.5.1 presents a background on MZIs and SPNNs and a summary of prior related work. In Section 4.5.2, we analyze the impact of FPVs on MZIs (device level) and array of cascaded MZIs (network level) in SPNNs. Section 4.5.3 presents our MZI design optimization solution to design devices with improved tolerance to different FPVs. Section 4.5.4 presents simulation results that highlight how the optimized MZIs can improve the accuracy of the unitary transformation (at the layer level) and the inferencing accuracy (at the system level), in the presence of FPVs. Last, we draw conclusions in Section 4.5.5.

4.5.1 Background and Related Work

This section summarizes fundamentals of MZIs and coherent SPNNs designed based on MZIs. Also, it presents some preliminary models to help understand the impact of FPVs on photonic waveguides, and reviews some prior related work.

4.5.1.1 Mach–Zehnder Interferometer (MZI)

A 2×2 MZI in an SPNN consists of two tunable phase shifters (ϕ and θ) on the upper arm and two 50:50 beam splitters as shown in Fig. 4.26(c). The phase shifters are used to apply configurable phase shifts and obtain varying degrees of interference between the optical signals traversing the MZI arms. In SPNNs, phase shifters are often implemented using thermal micro-heaters for lower optical loss, where the effective index of the underlying waveguide changes with temperature (i.e., due to thermo-optic effect of silicon), hence altering the phase of the optical signal. Note that the proposed analyses in this chapter are independent of the phase-shift mechanism in the MZI. Moreover, 2×2 beam splitters can be designed using directional couplers (DCs) where a fraction of the optical signal (defined as κ) at an input port is transmitted to an output port, and the remaining signal is coupled to the other output port, as shown in the inset of Fig. 4.26(c). The ideal transfer matrix (T_{MZI}) for an MZI with two phase shifters (θ and ϕ) and two 50:50 beam splitters can be

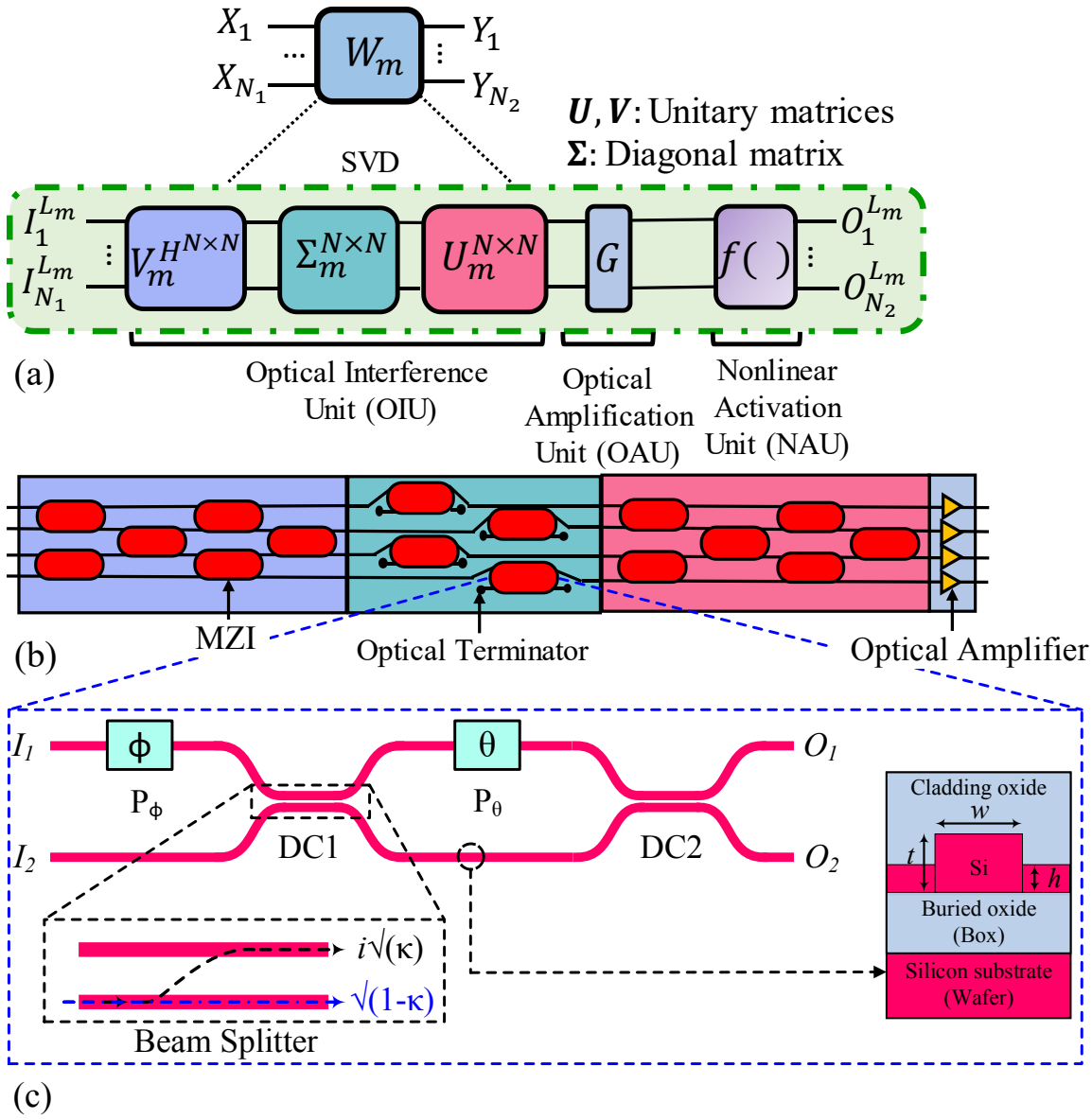


Figure 4.26: (a) Overview of singular value decomposition (SVD) of a weight matrix related to a fully connected layer (L_m) with N_1 as the number of input ports and N_2 as the number of output ports. (b) An optical-interference unit (OIU). (c) A 2×2 MZI structure with two integrated phase shifters (θ and ϕ) and two beam splitters based on directional couplers (DCs).

defined as [123] (Fig. 4.26(c)):

$$T_{MZI}(\theta,\phi)=\begin{bmatrix}\frac{e^{i\phi}}{2}(e^{i\theta}-1) & \frac{i}{2}(e^{i\theta}+1)\\ \frac{ie^{i\phi}}{2}(e^{i\theta}+1) & -\frac{1}{2}(e^{i\theta}-1)\end{bmatrix}. \tag{4.26}$$

4.5.1.2 Coherent SPNNs based on MZIs

Compared to non-coherent SPNNs, coherent SPNNs—considered in this chapter—use a single wavelength and MZI devices, in which the adjusted phase shifts in the phase shifters denote the dynamic weight parameters. Fig. 4.26(a) shows an example of a coherent SPNN. A fully connected layer in a deep neural network (L_m) can be realized with n_m neurons. Each layer performs a linear matrix-vector multiplication and accumulation (MAC) and passes the outputs to the next layer. The output vector of L_m can be mathematically modeled as $O_m^{n_m\times 1}=f_m(W_m\times O_{m-1}^{n_{m-1}\times 1})$. Here, f_m is a non-linear activation function (performed by non-linear activation unit (NAU) in Fig. 4.26(a)) of L_m, and W_m is the corresponding weight matrix of L_m. Given a weight matrix W_m, which can be obtained by training the network and mapped to MZIs using singular value decomposition (SVD), each weight matrix W_m can be written as $W_m=U_m^{n_m\times n_m}\Sigma_m^{n_m\times n_m}V_m^{H,n_m\times n_m}$. Here, $U_m^{n_m\times n_m}$ and $V_m^{H,n_m\times n_m}$ are the unitary matrices with dimension of $n_m\times n_m$, and $\Sigma_m^{n_m\times n_m}$ is a diagonal matrix with dimension of $n_m\times n_m$. Also, $V_m^{H,n_m\times n_m}$ is the Hermitian-transpose of $V_m^{n_m\times n_m}$. A unitary matrix can be realized by using a cascaded array of 2×2 MZIs. As shown in Figs. 4.26(a) and 4.26(b), this unit is called the optical-interference unit (OIU) and is responsible for performing the MAC operation.

Several approaches have been proposed to design the architecture of MZI arrays to perform MAC operations in the optical domain [124–126]. Out of these, the Clements design, due its symmetric nature, has a low optical loss and footprint. Therefore, we use the Clements design [125] to transform each unitary matrix ($U_m^{n_m\times n_m}$ and $V_m^{H,n_m\times n_m}$) to a cascaded MZI array with a specific phase setting per MZI in the network. In the Clements method, each unitary matrix will be mapped to a cascaded MZI array with a total number of $\frac{N(N-1)}{2}$ MZIs, where N is the size of the designated unitary matrix. Note that the diagonal matrix ($\Sigma_m^{n_m\times n_m}$) can be realized with an array of MZIs with

Table 4.8: Parameters used to generate FPV maps.

Design Parameter	Correlation Length (l)	Standard Deviation (σ)
Waveguide width	1 mm and 100 μm	$\sigma_w = 5$ nm
SOI thickness	1 mm and 100 μm	$\sigma_t = 2$ nm
Slab thickness	1 mm and 100 μm	$\sigma_h = 2.5$ nm

one input and one output terminated. The optical-amplification unit (OAU) in Fig. 4.26(a), which is required to obtain arbitrary diagonal matrix, can be realized using semiconductor optical amplifiers (SOAs).

4.5.1.3 Fabrication-Process Variations (FPVs)

FPVs in silicon photonics originate in optical-lithography process imperfections, contributing to changes in the waveguide width, SOI thickness (dominated by the host wafer), and slab thickness (in case of a ridge waveguide). Such changes deviate the effective index (n_{eff}) in a waveguide and in turn the propagation constant (β) which determines the optical phase of the signal traversing the waveguide. Considering Fig. 4.26(c) and as an example, the effective index (n_{eff}) in a ridge waveguide depends on the optical wavelength and the critical dimensions of the waveguide [46], i.e., width (w), SOI thickness (t), and slab thickness (h) in Fig. 4.26(c). Note that $h = 0$ for a strip waveguide and this relationship can be extracted from (3.1). Here, β can be defined as $\beta = \psi/L$, where ψ is the single pass phase-shift induced in a waveguide of length L. Leveraging (3.1), propagation constant changes ($\Delta\beta$) in a ridge waveguide under FPVs is given by:

$$\Delta\beta = \frac{2\pi}{\lambda}\left(\frac{\partial n_{\text{eff}}}{\partial w}\rho_w + \frac{\partial n_{\text{eff}}}{\partial t}\rho_t + \frac{\partial n_{\text{eff}}}{\partial h}\rho_h\right). \tag{4.27}$$

Here, ρ_w, ρ_t, and ρ_h are the variations in the waveguide width, SOI thickness, and slab thickness, respectively. When using a strip waveguide, $\frac{\partial n_{\text{eff}}}{\partial h} = \rho_h = 0$. Note that, we do not consider the variations in L (see Section 4.5.3).

4.5.1.4 Related Work on FPV Analysis in SPNNs

Our prior work in [29] studied the impact of random phase noise due to FPVs and thermal crosstalk at the system level in SPNNs by developing a framework that identifies critical components in the network. In [123], imprecisions were introduced in SPNNs after software training such that pre-fabrication training tends to be more scalable in terms of network size and volume. This helps designing precise and cost effective MZIs, when compared to re-configurable ones, which can be exploited for AI applications to perform matrix multiplication. A method was presented in [127] to counter the impact of both FPVs and thermal effects using modified cost functions during training with added benefits of post-fabrication hardware calibration. The impact of FPVs can also be reduced by minimizing the tuned phase angles in an SPNN; this can be done by leveraging the non-uniqueness of SVD as it was shown in [128]. The work in [129] proposed a circuit-level hardware error correction solution for SPNNs in which local error correction requires characterization of each phase shifter and passive splitter in the photonic circuit while relying on detectors in the output to calibrate the parameters.

The aforementioned methods focus on mitigating deviations in SPNNs *post-fabrication* by either post-fabrication training methods to compensate for any additionally introduced phase noises, which might impact the network accuracy [127], or calibrating each noisy component, where additional error-detection phases are required and the complexity can increase as the SPNN scales up [129]. In this work, we focus on exploring and optimizing the physical design of MZI devices in coherent SPNNs under FPVs *prior to fabrication*, hence improving the network tolerance under different FPVs. We show that by exploring and optimizing MZI device physical-level design, we can improve the relative-variation distance (RVD) in SPNNs, which quantifies the deviation between the intended unitary matrix and the deviated unitary matrix, to enhance the overall inferencing accuracy.

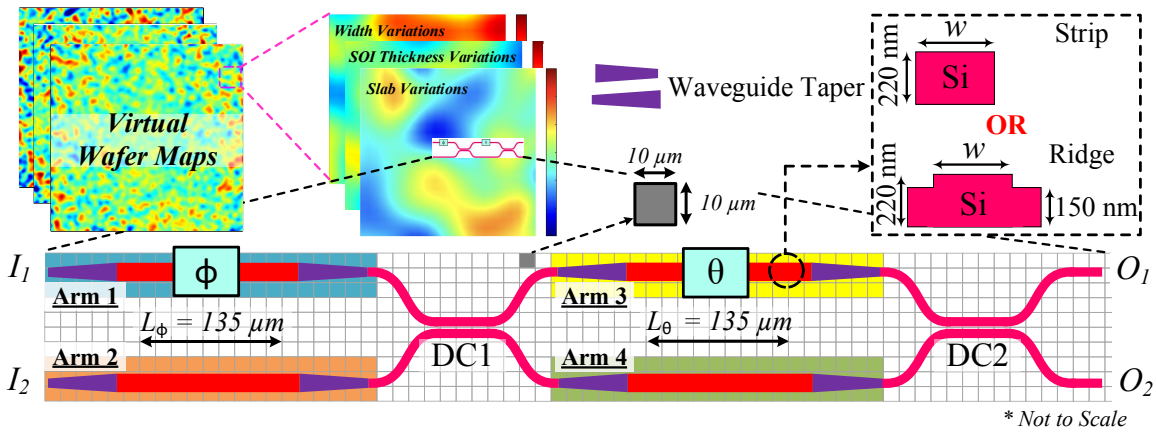


Figure 4.27: An MZI device structure with waveguide tapers mapped to FPV maps (top), based on [3], with a mesh size of 10 μm. The MZI can use strip waveguides or ridge waveguides, both with the SOI thickens of 220 nm and varying waveguide width (w) on each arm. The design of slab thickness (150 nm) in the ridge waveguide is discussed in Section IV. Note that variation-free directional couplers (DCs) are considered.

4.5.2 Modeling FPVs in Coherent SPNNs

This section presents a detailed bottom-up analysis of the impact of FPVs in the waveguide width, SOI thickness, and slab thickness at the device level (i.e., MZI devices) and network level (OIU in Fig. 4.26(b)) in coherent SPNNs. We show the impact of FPVs on SPNN inferencing accuracy (system level) in Section V.

4.5.2.1 Device Level: MZI Performance under FPVs

To study the impact of FPVs on MZI devices, we should first model FPVs in silicon photonics and explore how MZI devices experience such FPVs. In our prior work [3], we have developed realistic wafer variation maps that model radial-variation effects and correlation among different variations—both of which are critical to realistically model FPVs in silicon photonics—in SOI wafers. Such maps were developed based on mean, standard deviation (σ), and correlation lengths (l) experimentally characterized in collaboration with CEA-Leti in [3]. Table 4.8 summarizes different parameters considered to generate FPV maps for our calculations, which are based on analyzing experimental data from characterizing actual 200-mm wafers at CEA-Leti. Fig. 4.27-top shows examples of wafer and die maps generated using our in-house FPV wafer map simulator with a resolution of 10 μm $\times$10 μm (i.e., the mesh size in the map—see Fig. 4.27).

MZIs are bulky devices (e.g., ≈340 μm in length [130]), and hence every section of the device will experience slightly different FPVs (see Fig. 4.27). Such a difference changes with the correlation length in the variations (e.g., long-range versus short-range correlated variations). As a result, it is critical to analyze the impact of the non-uniformity of FPVs in MZI devices. Considering the MZI shown in Fig. 4.27 with its four arms labeled (Arm1–Arm4), we average the width, SOI thickness, and slab thickness (when using a ridge waveguide in the MZI) variations observed on each MZI arm, separately over the section colored on each arm.

Considering the MZI in Fig. 4.27, the optical signals traversing the two opposite arms of the device—before the input DC (i.e., Arms 1 and 2) and output DC (i.e., Arms 3 and 4)—should only experience the desired phase change ϕ or θ, adjusted after the network training. Note that the phase shifters are integrated on the top of the silicon waveguides. However, due to the impact of non-uniform FPVs on each individual arm, the optical signals experience some undesired phase changes. Assuming that each arm's length ($L_1 = L_2 = L_\phi$ and $L_3 = L_4 = L_\theta$) does not undergo any variations, the optical phase difference between the two optical signals traversing the opposite arms (Arms 1–2 and Arms 3–4) and interfering at the input of DC1 ($\Delta\Phi_{DC1}$) and DC2 ($\Delta\Phi_{DC2}$) is:

$$\Delta\Phi_{DC1} = \phi + \left|\Delta\beta_1 L_1 - \Delta\beta_2 L_2\right|, \tag{4.28}$$

$$\Delta\Phi_{DC2} = \theta + \left|\Delta\beta_3 L_3 - \Delta\beta_4 L_4\right|. \tag{4.29}$$

Here, $\Delta\beta_1$, $\Delta\beta_2$, $\Delta\beta_3$, and $\Delta\beta_4$ are, respectively, the propagation constant changes on MZI's Arms 1–4, which can be calculated using (4.27).

Leveraging (4.27),(4.28), and (4.29), and assuming variation-free DCs, the MZI transfer matrix in (4.26) can be updated to take into consideration the impact of FPVs on MZI arms, resulting in undesired optical phase noises:

$$T'_{MZI}(\theta,\phi) = \begin{pmatrix} \frac{\sqrt{2}}{2}e^{i(\theta+\Delta\beta_3 L_3)} & i\frac{\sqrt{2}}{2}e^{i\Delta\beta_4 L_4} \\ i\frac{\sqrt{2}}{2}e^{i(\theta+\Delta\beta_3 L_3)} & \frac{\sqrt{2}}{2}e^{i\Delta\beta_4 L_4} \end{pmatrix} \cdot \begin{pmatrix} \frac{\sqrt{2}}{2}e^{i(\phi+\Delta\beta_1 L_1)} & i\frac{\sqrt{2}}{2}e^{i\Delta\beta_2 L_2} \\ i\frac{\sqrt{2}}{2}e^{i(\phi+\Delta\beta_1 L_1)} & \frac{\sqrt{2}}{2}e^{i\Delta\beta_2 L_2} \end{pmatrix}, \tag{4.30}$$

where, as shown in Fig. 4.27, $L_1 = L_2 = L_\phi$ and $L_3 = L_4 = L_\theta$. In this chapter, we assume $L_\phi = L_\theta \approx 135$ μm, considered as an example based on [130]. Leveraging (4.30), we can capture the impact of non-uniform variations in the waveguide width, SOI thickness, and slab thickness in any MZI design in coherent SPNNs. Note that FPVs will deviate the splitting ratio, which ideally should be 50:50, in the input and output directional couplers (DC1 and DC2 in Fig. 4.27), hence affecting the network accuracy [29]. However, this chapter focuses on the optical phase noise due to FPVs in MZIs and considers variation-free DCs in SPNNs. Also, note that the analyses proposed in this section are independent of the example MZI considered in Fig. 4.27.

4.5.2.2 Network Level: OIU Performance under FPVs

Here, we model the impact of FPVs on the performance of an OIU, shown in Fig. 4.26(b). Note that in this chapter we do not consider the impact of FPVs in the OAUs and NAUs (see Fig. 4.26(a)) as they are often implemented either electronically or opto-electronically [131, 132]. Recall from Section 4.5.1.2 that, given a weight matrix $W = U\Sigma V^H$, the matrices U, Σ, and V can be decomposed to θ and ϕ phase values on each MZI in an OIU using Clements decomposition [125]. Under FPVs, the transfer matrix of each MZI in the OIU will deviate in a manner that can be calculated using (4.30). To analyze the impact of such variations at the network level, we use relative-variation distance (RVD). RVD determines the deviation between an intended matrix and a deviated matrix [29]. We found that there is a strong correlation between network-level RVD and system-level inferencing accuracy in SPNNs (we will discuss this in Section V). RVD can be defined as:

$$RVD(W, \overline{W}) = \frac{\Sigma_m \Sigma_n |W^{m,n} - \overline{W}^{m,n}|}{\Sigma_m \Sigma_n |W^{m,n}|}, \tag{4.31}$$

where $|.|$ denotes the absolute value of a complex number. W is the nominal weight matrix and $\overline{W}$ is the deviated weight matrix under FPVs related to a fully connected layer. $W^{m,n}$ denotes the element at the m^{th} row and n^{th} column of W. Each MZI in the OIU has a unique impact on each element of the weight matrix. Accordingly, variations related to each MZI in the network have a unique effect on the overall RVD. Higher RVD means the actual weight matrix is more deviated

from the intended one, which can be interpreted as observing a lower inferencing accuracy at the system level.

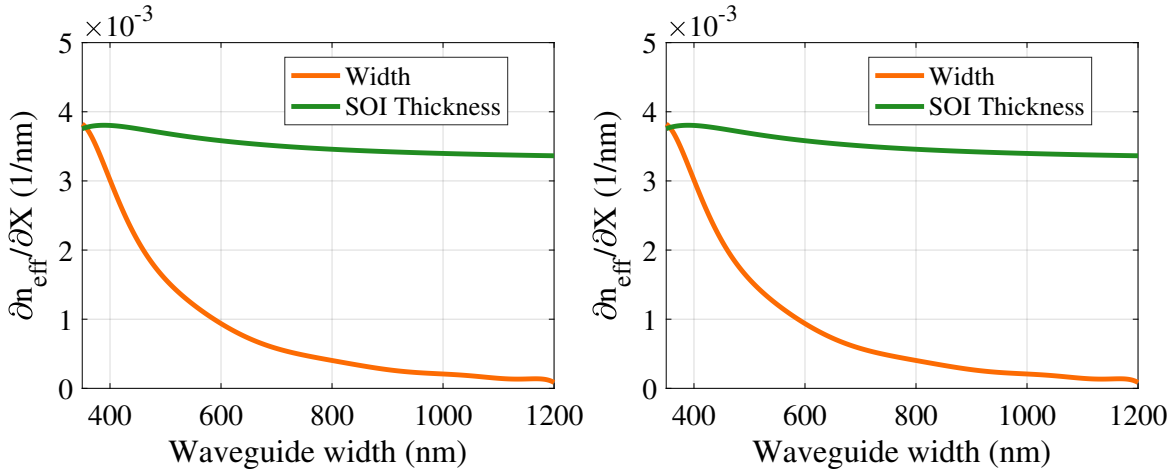
To compute $\overline{W}$ in (4.31), we first need to analyze the impact of non-uniform FPVs in the waveguide width, SOI thickness, and slab thickness on each individual MZI in an OIU. As the first step, we calculate the total dimension of an OIU based on the length of an individual MZI (l_{MZI}) and the distance between its input and output ports (g_{MZI}). In this work and as an example, we consider $l_{MZI} = 340$ μm and $g_{MZI} = 30$ μm, and based on the phase shifter length in [130] (i.e., $\approx$135 μm). Accordingly and using the Clements design for the OIU [125], the total dimension of the OIU can be calculated. As the second step, we use our in-house FPV wafer map simulator (see Section 4.5.2.1) to generate a die map that matches the size of OIU. We then place the OIU on the die map and extract FPV information for each individual MZI in the OIU. Note that each MZI itself experiences different variations (non-uniform variations across a single device; a.k.a. intra-device variations), and the FPVs between two different MZIs are also different. All such non-uniformities are considered in our device-level (Section 4.5.2.1) and network-level analyses. By capturing FPVs for each individual MZI in an OIU, we can calculate the deviated $\overline{W}$ in (4.31) based on:

$$\overline{W}_m = \overline{U}_m \times \overline{\Sigma}_m \times \overline{V_m^H}. \tag{4.32}$$

Here, $\overline{U}_m$ and $\overline{V_m^H}$ are the deviated unitary transfer matrices and $\overline{\Sigma}_m$ is the deviated diagonal matrix under FPVs. They can be calculated based on multiplying MZI transfer matrices under FPVs (the model in (4.30)), and in a specific order determined by the Clements design [125]. For example, for $\overline{U}_m$ we have:

$$\overline{U}_m = D\left(\prod_{(m,n)\in S} T'_{MZI,m,n}\right), \tag{4.33}$$

where m and n should be calculated based on the mapping method (e.g., Clements [125]) used to map the weight matrices to cascaded MZI arrays. Also, S is the order of multiplication which again should be determined by the mapping method. Moreover, D is a diagonal matrix with unity



(a) Strip waveguide **(b)** Shallow-etched ridge waveguide

Figure 4.28: Rate of changes in waveguide effective index (see the strip and ridge waveguides in Fig. 4.27) under FPVs $\frac{\partial n_{\text{eff}}}{\partial X}$, where X shows the design parameter under FPVs, in (a) a strip and (b) a shallow-etched ridge waveguide, when the waveguide width (w) increases from 350 to 1200 nm. Results are for $t = 220$ nm and $h = 150$ nm (for the ridge waveguide in (b)).

magnitude and is not related to the physical placement of MZIs. Similarly, we can calculate $\overline{V_m^H}$ and $\overline{\Sigma}_m$. Although we considered the Clements method for mapping the weights to phase settings of a cascaded MZI array in OIUs, our network-level models in this section can work with any mapping method.

4.5.3 SPNN Design Optimization Under FPVs

In this section, we explore the design space of MZI devices under different FPVs to optimize their performance in SPNNs. In particular, we focus on minimizing the impact of FPVs on MZI arms that imposes undesired optical phase noises in the device, leading to faulty matrix-vector multiplication. As discussed in Section 4.5.2, FPVs also deviate the splitting ratio in DCs in an MZI. Nevertheless, the design optimization solution in this section assumes ideal DCs. Note that FPV-tolerant DCs can be designed based on the method proposed in [133].

Considering (4.28) and (4.29), one can alleviate the impact of FPV-induced optical phase noise in an MZI by implying $|\Delta\beta_1 L_1 - \Delta\beta_2 L_2| \rightarrow 0$ and $|\Delta\beta_3 L_3 - \Delta\beta_4 L_4| \rightarrow 0$. Accordingly, to

obtain a phase-noise-free MZI, we should have:

$$\frac{L_1}{L_2} = \frac{\Delta\beta_2}{\Delta\beta_1} \text{ and } \frac{L_3}{L_4} = \frac{\Delta\beta_4}{\Delta\beta_3}. \tag{4.34}$$

This indicates that the length ratio between any two opposite arms in the MZI should be inversely proportional to the ratio of the changes in their waveguide propagation constants ($\Delta\beta$), under non-uniform FPVs. In this section and for brevity, we assume $L_1 = L_2 = L_3 = L_4$ and without variations. As a result and based on (4.34), we should minimize $|\Delta\beta_1 - \Delta\beta_2|$ and $|\Delta\beta_3 - \Delta\beta_4|$ under different FPVs. In other words, we should make sure that the propagation constant changes on the two opposite arms in an MZI are as small as possible (i.e., $\Delta\beta_1 \to 0$, $\Delta\beta_2 \to 0$, $\Delta\beta_3 \to 0$, and $\Delta\beta_4 \to 0$), or the propagation constant changes on the two opposite arms are as close as possible (i.e., $\Delta\beta_1 \to \Delta\beta_2$ and $\Delta\beta_3 \to \Delta\beta_4$).

Considering (4.27), $\Delta\beta$ is proportional to the rate of changes in the waveguide's effective index under FPVs (i.e., $\frac{\partial n_{\text{eff}}}{\partial X}$, where X denotes the design parameter under FPVs: i.e., $X = w$ for width, $X = t$ for SOI thickness, and $X = h$ for slab thickness variations. In our prior work [3], we found that as the waveguide width increases, $\frac{\partial n_{\text{eff}}}{\partial X}$ decreases, especially under waveguide width variations (i.e., when $X = w$). This is because as the waveguide width increases, a bigger portion of the optical mode is confined in the waveguide core (and the confinement is also stronger), and hence the variations in the waveguide width will create less distortion in the optical mode in the waveguide. Also, note that increasing the waveguide width helps reduce the propagation loss in strip and ridge waveguides [3]. Fig. 4.28(a) shows the rate of changes in the effective index of a strip waveguide with $t = 220$ nm (see Fig. 4.27) and when the waveguide width (w) changes from 350 nm to 1200 nm, both considered as an example. As it can be seen, as the waveguide width increases, $\frac{\partial n_{\text{eff}}}{\partial w}$ decreases sharply but $\frac{\partial n_{\text{eff}}}{\partial t}$ decreases sightly and stays higher than $\frac{\partial n_{\text{eff}}}{\partial w}$ under different waveguide widths. While waveguide width can be changed during the design time, the SOI thickness cannot be changed; this parameter is determined by the host SOI wafer.

As it can be seen from Fig. 4.28(a), both $\frac{\partial n_{\text{eff}}}{\partial t}$ and $\frac{\partial n_{\text{eff}}}{\partial w}$ are still high in strip waveguides. To address this, we also explore the design of MZIs using a shallow-etched ridge waveguide with a

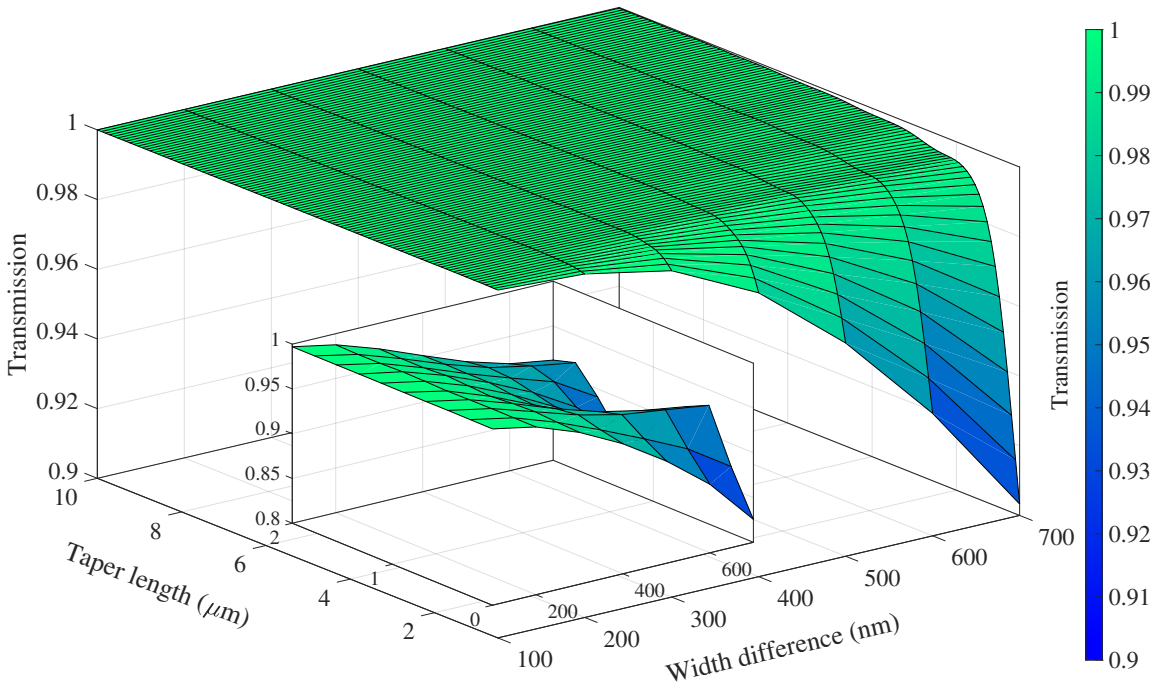


Figure 4.29: Minimum taper length required to keep the optical transmission between two waveguides of different widths consistent (i.e., at 1 in the figure) and to avoid mode distortion. The inset zooms in the results for the taper length of 0–2 μm.

slab thickness (h) of 150 nm, as shown in Fig. 4.27. We simulated different slab thicknesses from 60 nm to 180 nm (results are not shown in the paper), and $h = 150$ nm returned the best results. Such a ridge waveguide is common in the design of grating structures [134]. By adding the slab to a strip waveguide (i.e., making it a ridge waveguide), the optical mode is pulled mostly towards the slab region, hence SOI thickness variations should have less impact on the optical mode. Similar to Fig. 4.28(a), Fig. 4.28(b) shows the rate of changes in the effective index of the shallow-etched ridge waveguide with $t = 220$ nm and $h = 150$ nm, and when the waveguide width (w) changes from 350 nm to 1200 nm. Observe that, compared to the strip waveguide, both $\frac{\partial n_{\text{eff}}}{\partial w}$ and $\frac{\partial n_{\text{eff}}}{\partial t}$ are smaller in the ridge waveguide. The shallow-etched ridge waveguide also suffers from variations in its slab thickness. Nevertheless, as it is shown by Fig. 4.28(b), $\frac{\partial n_{\text{eff}}}{\partial h}$ is much smaller than $\frac{\partial n_{\text{eff}}}{\partial t}$ in the shallow-etched ridge waveguide, and it decreases sharply as the waveguide width increases.

Considering the results in Fig. 4.28, designing MZIs with wider strip and shallow-etched ridge waveguides should help minimize the changes in the propagating constant on each MZI arm (see

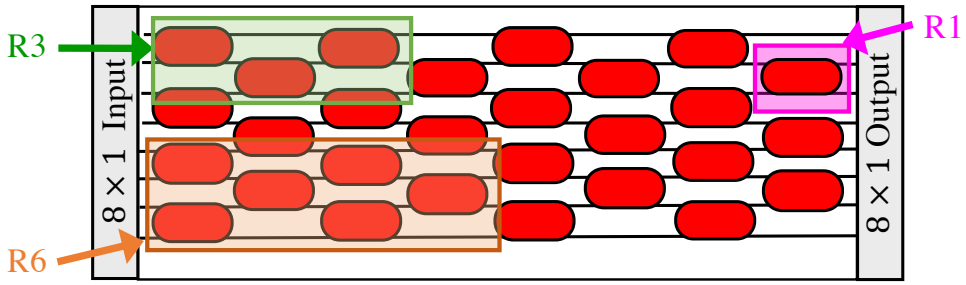


Figure 4.30: Different region sizes (R1, R3, and R6) and related MZIs in a single 8×8 OIU unit. R12 (not shown) can be obtained similarly. Each MZI block size is 30×340 μm^2.

(4.27) and (4.34)). Moreover, to increase the waveguide width, an important design consideration is to include waveguide tapers on the MZI arms as shown in Fig. 4.27. Waveguide tapers are essential to avoid optical mode distortion and higher order mode excitation when moving from the nominal waveguide width (i.e., 470 nm in this chapter—see Section V) to a wider waveguide and vice versa [135]. In particular, the taper length should be long enough to avoid any optical transmission and mode distortion. Using Lumerical MODE [65], we simulated the fundamental mode transmission between two waveguides of different widths in Fig. 4.29. As it can be seen, a waveguide taper length of ≈1 μm will be sufficient for every 100 nm width difference between two waveguides of different widths (see Fig. 4.27). This helps us calculate the area overhead when we optimize MZIs with wider waveguide widths.

Considering different FPVs in the waveguide width, SOI thickness, and slab thickness, modeled based on FPV wafer map models in [3], we consider two scenarios based on which the design of MZIs in an SPNN can be optimized. First, in the *region-based-tolerant MZI design*, we assume a designer may have some *a priori* knowledge of the FPVs. This is a valid assumption as silicon photonics foundries can provide some FPV maps, with different variation data resolutions, to the designers using their fabrication-processes. Second, we assume a *worst-case-tolerant MZI design* scenario, where a designer may have very little to no *a priori* knowledge of the FPVs, and hence the MZIs in an SPNN are designed considering the worst-case FPV scenarios (e.g., corner analysis). In such a scenario, we design the worst-case-tolerant MZIs with the largest possible waveguide

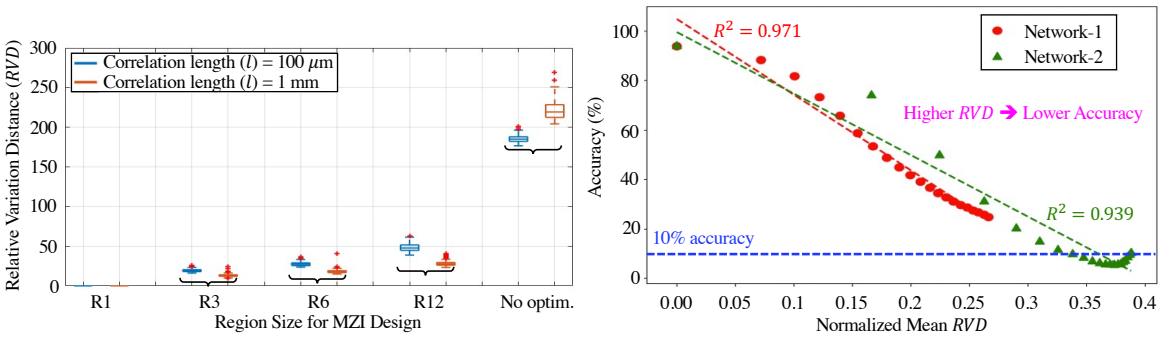
widths, and equal widths on all the arms, while considering the area overhead in the MZIs. This is discussed further in Section 4.5.4.

For the region-based-tolerant MZI design, we can define different regions of different sizes (i.e., number of MZIs) in an SPNN, an example of which is shown in Fig. 4.30. Such regions group the MZIs that are spatially close on the die with one another. We assume that the designer has some *a priori* knowledge of the FPVs for all (and not the individual) MZIs grouped in the same region. As a result, the smaller the region size (e.g., R1 with a single MZI in Fig. 4.30), the more detailed FPV information is available to the designer and vice versa (e.g., R6 with six MZIs in Fig. 4.30). Accordingly, we consider the average observed non-uniform variations in a region to design region-based-tolerant MZIs by performing an exhaustive search for the MZI waveguide widths (using results in Fig. 4.28) while considering the area overhead in the MZIs imposed by adding the tapers. Here, we might have different waveguide widths (between the search range of 350 nm to 1200 nm—see Fig. 4.28) on each MZI arm.

As we will show in Section V, our region-based-tolerant and worst-case-tolerant MZI designs minimize optical phase noises in MZIs under different FPVs, hence they improve the network accuracy in SPNNs. Nevertheless, it is important to note that our device-level design optimization solutions proposed in this section do not aim at completely eliminating, if at all feasible, the impact of FPVs in SPNNs. That being said, our optimization will reduce the impact of FPVs in SPNNs sufficiently so that the overhead and complexity of dynamic calibration techniques (e.g., [129]) to eliminate the impact of such variations in SPNNs will be significantly reduced.

4.5.4 Simulation Results and Discussions

FPVs lead to undesired optical phase noises, which, in turn, lead to faulty matrix-vector multiplication in the fully connected layers of an SPNN. The sensitivity of a standalone MZI to FPVs depends on its physical design parameters (see Fig. 4.28), among which only the MZI waveguide width can be freely altered during the design time. Leveraging realistic and correlated FPV maps developed based on [3] (see Fig. 4.27-top) and the proposed MZI design optimization in



(a) RVD for 100 random unitary matrices

(b) Accuracy and RVD under FPVs (different σ_w's and σ_t's)

Figure 4.31: (a) RVD for different region sizes and under FPVs with different correlation lengths. (b) High R^2 values denote a strong linear correlation between RVD and accuracy. FPVs are based on the parameters in Table 4.8. We consider the linear correlation as an example to calculate R^2.

Section 4.5.3, we explore and optimize the nominal waveguide widths in the MZIs in SPNN case studies considered in this section, to improve their tolerance under different FPVs.

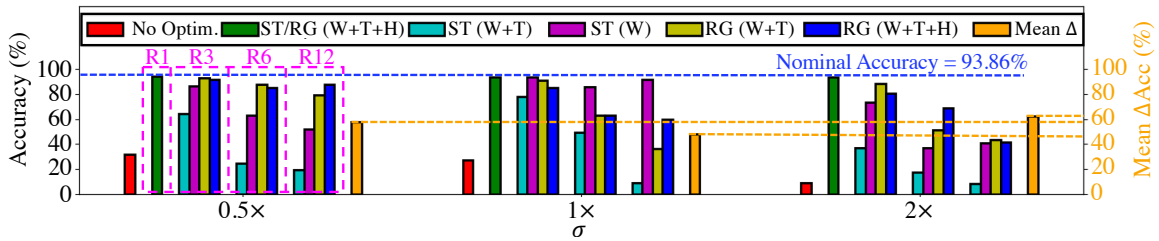
Prior to evaluating the impact of the proposed MZI optimization on SPNN accuracy under FPVs, let us explore whether such an optimization is independent of the model and the dataset. To examine this, Fig. 4.31(a) considers 100 randomly generated 16×16 unitary matrices—each of which belongs to a different weight matrix—and presents a box plot of the distribution of RVD's between the nominal and the deviated unitary matrices under FPVs, and for different region sizes. Recall from Section IV that to design the region-based-tolerant MZIs, we take the mean variations affecting a region on the FPV map with 1, 3, 6, or 12 MZIs into consideration (see Fig. 4.30). All the MZIs in a particular region are replaced with the optimized region-based-tolerant MZI, designed using shallow-etched ridge waveguides (see Figs. 4.27 and 4.28(b)). As it is shown in Fig. 4.31(a), in all cases (i.e., R1–R12), the mean RVD is significantly reduced when optimized MZIs are used. In particular, the interquartile ranges (IQRs) in the box plots are consistent among all the unitary matrices in a region: this shows that the proposed optimization is effective independent of the considered unitary matrix. Also, the mean RVD decreases with decreasing the region size (i.e., when a designer has access to more detailed FPV data from the foundry), and it is the highest when MZIs are not optimized.

Table 4.9: Architectures of the SPNNs considered. FC(x,y): Fully connected layer with x inputs and y outputs, SP: Softplus activation, LSM: LogSoftMax activation, PhS: phase shifters.

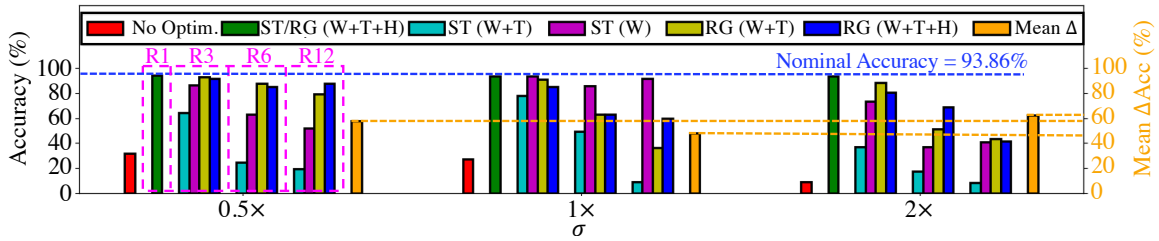
Model	Architecture	# PhS
Network-1	FC(16,16)-SP-FC(16,16)-SP-FC(16,10)-LSM	1380
Network-2	FC(64,64)-SP-FC(64,64)-SP-FC(64,10)-LSM	20,580

To explore the impact of our proposed MZI design optimization on SPNN accuracy, we consider a case study of two fully connected SPNNs with different footprint (see Table 4.9) trained on the MNIST dataset. To compress the $28\times28 = 784$ dimensional feature vector in the MNIST dataset, we take the shifted fast Fourier transform of each image. The compressed 16-dimensional feature vector for Network-1 is then obtained by considering the values within the 4×4 region at the center of the frequency spectrum. Similarly, for the larger Network-2, we use a 64-dimensional feature vector by considering the 8×8 region at the center of the frequency spectrum. In Fig. 4.31(b), we show the linear correlation between the accuracy and the mean RVD, averaged over the 6 unitary matrices (two in each of the three OIUs) and normalized over the number of phase shifters in each network. Given such a strong linear correlation, our method should improve the accuracy of all SPNNs under FPVs, irrespective of the nominal phase angles.

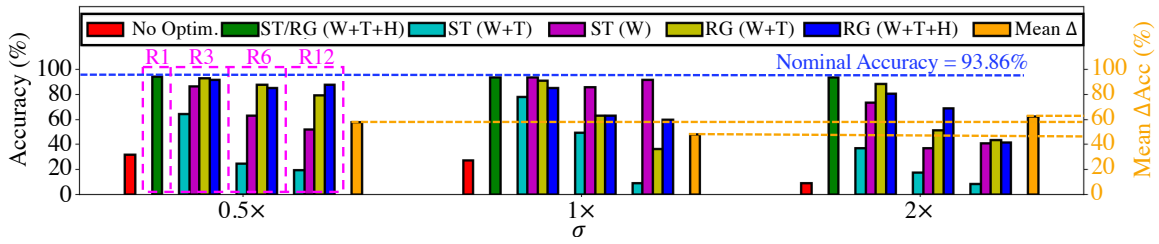
By applying realistic and correlated FPV maps—generated using our prior work in [3] and considering parameters in Table 4.8—to Network-1 and Network-2, Fig. 4.32 shows the inferencing accuracy in each network with conventional MZIs (No Optim.) and optimized region-based-tolerant MZIs, which can have different waveguide widths on each arm. For the conventional MZI, we designed an MZI using strip waveguides with $t = 220$ nm and $w = 470$ nm and variation-free $\approx$10-μm-long DCs with a 200 nm gap, to obtain 50:50 splitting ratio in the MZI. In each plot in Fig. 4.32, we consider three different standard deviations for the waveguide width, SOI thickness, and slab thickness variations: $0.5\times$, $1\times$, and $2\times$ the expected standard deviations (σ_w, σ_t, σ_h) in Table 4.8. Moreover, the variation maps are generated for two correlation lengths of 100 μm and 1 mm. Considering Figs. 4.32(a)–(d), with no MZI optimization (i.e., No Optim.; red bars), the SPNN accuracy is always the least (e.g., 7.73% with $1\times\sigma$ in Fig. 4.32(d)). Considering optimized region-based-tolerant MZI design with strip or shallow-etched ridge waveguides under all the vari-



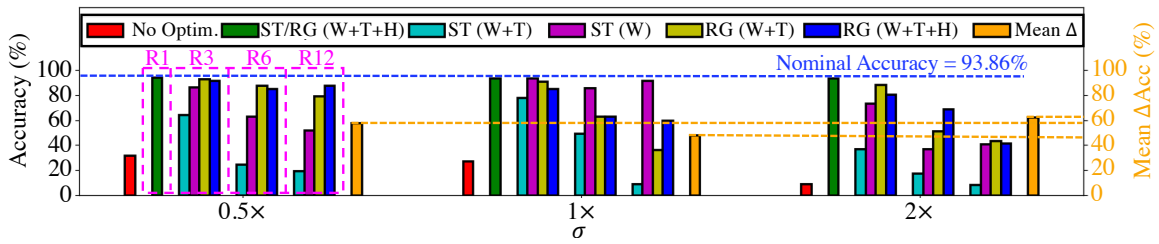
(a) Network-1, correlation length = 100 μm



(b) Network-1, correlation length = 1 mm



(c) Network-2, correlation length = 100 μm



(d) Network-2, correlation length = 1 mm

Figure 4.32: Accuracy of two SPNNs in Table 4.9 under correlated FPVs in width, SOI thickness, and slab thickness before and after the optimization. Here, ST and RG denote strip and shallow-etched ridge waveguide, respectively. The parameters listed inside (.) show the variations considered, with W, T, and H denoting waveguide width, SOI thickness, and slab thickness variations, respectively. Note that results for ST (W+T) and RG (W+T+H) are the same for R1. The second y-axis shows the average difference between No Optim. accuracy (i.e., using the conventional MZI) and the accuracy obtained using RG (W+T+H) for R1, R3, R6, and R12.

ations and R1 (i.e., when regions include a single MZI—see Fig. 4.30), the network accuracy in Figs. 4.32(a)–(d) is almost the same as the nominal accuracy. This is because each individual MZI has been optimized considering its exact FPV profile. This is for an ideal case when a designer has full access to variation data affecting each MZI in the network, so it may be impractical in most cases (R1 results are shown to indicate the efficiency of our optimization in such rare cases).

Considering more realistic and practical region sizes (R3, R6, and R12) and a region-based-tolerant MZI design using strip waveguides, Figs. 4.32(a)–(d) show the SPNN accuracy when considering (i) both waveguide width and SOI thickness variations (W+T; light green bars), and (ii) width variations only (W; magenta bars)—when SOI thickness variations are negligible, e.g., through SOI thickness uniformity improvement [136]. Observe that with (i), the optimized MZI can help retrieve some accuracy, which also decreases as the region size increases. But overall, the gain in accuracy in this case is small. This is due to the fact that optimized MZIs designed using strip waveguides are not sufficiently tolerant to thickness variations (see Fig. 4.28(a)). Considering (ii), the optimized MZI designed using strip waveguides achieves better accuracy improvements (and higher than (i)) in both the networks.

Figs. 4.32(a)–(d) also show the network accuracies with the optimized region-based-tolerant MZIs designed using shallow-etched ridge waveguides, under the presence of (i) waveguide width and SOI thickness variations (W+T; yellow bars), and (ii) all the variations (W+T+H; dark blue bars). Observe that with both (i) and (ii), the optimized MZIs using shallow-etched ridge waveguides perform much better compared to those using strip waveguides (i.e., ST (W+T)). Comparing (i) and (ii)—when considering slab thickness variations—the network accuracy is (slightly) lower in some cases. Nevertheless, considering the average difference between No Optim. accuracy and the accuracy obtained with (ii) for R1, R3, R6, and R12 (orange bars; second y-axis), using optimized MZIs designed with shallow-etched ridge waveguides under all the variations can significantly improve the network accuracy (e.g., by up to 72% in Network-2 with $0.5\times\sigma$). This shows that the proposed optimization can improve the resilience of SPNNs to FPVs.

Another observation is how the network accuracy in Figs. 4.32(a)–(d) seem to change across different region sizes and FPV correlation lengths. In fact, when the correlation length in variations is shorter, variations tend to "change more" within a region with a given size, and as the region size increases, they change even more across the region. This imposes a higher error for the region-based-tolerant MZIs designed per region. Therefore, the accuracy results are generally a bit lower in Figs. 4.32(a) and 4.32(c) compared to those in Figs. 4.32(b) and 4.32(d).

To assess SPNN accuracy using the optimized worst-case-tolerant MZI design (see Section IV), we consider, as an example, Network-2 with FPVs of different correlation lengths and standard deviations in Table 4.8. In this experiment, we assume no *a priori* FPV knowledge and using strip waveguides in the design of the optimized worst-case-tolerant MZIs while considering waveguide width variations only. The worst-case-tolerant MZI is optimized by widening all the MZI arms together—i.e., MZI arms all have the same width after optimization—while considering the resulting area overhead due to the required waveguide tapers (see Figs. 4.27, 4.28(a), and 4.29). Results for this experiment (before and after the optimization) are shown in Table 4.10 for different area overhead. We observe that even with 1% area overhead, the accuracy improves. However, the improvements are significant only when the area overhead is greater than 8% for both the correlation lengths.

4.5.5 Conclusion

In this section, we have analyzed the impact of FPVs in the waveguide width, SOI thickness, and slab thickness on coherent SPNNs. In particular, we have modeled undesired optical phase noises due to such variations at the MZI device level, and how such phase noises contribute to the performance degradation, for which we have considered relative variation distance (RVD) at the network level in optical unitary multipliers built using MZIs. Furthermore, we have proposed physical-level design optimization solutions to enhance MZI device tolerance under correlated FPVs in SPNNs. Our simulation results for two SPNN case studies of different sizes and considering realistic FPV maps show that the proposed physical-level optimization can help significantly

Table 4.10: Network-2 accuracy with worst-case-tolerant MZIs designed using strip waveguides under width variations.

Area Overhead	Correlation Length	Width (nm)	Arm Length (μm)	Pre-Opt Accuracy	Post-Opt Accuracy
1%	1 mm 100 μm	533	135.63	11.65% 54.27%	20.47% 77.08%
2%	1 mm 100 μm	589	136.19	11.65% 54.27%	45.79% 86.48%
4%	1 mm 100 μm	688	137.18	11.65% 54.27%	83.97% 92.17%
8%	1 mm 100 μm	853	138.83	11.65% 54.27%	92.36% 93.77%
16%	1 mm 100 μm	1111	141.41	11.65% 54.27%	93.97% 94.28%
32%	1 mm 100 μm	1200	142.3	11.65% 54.27%	94.07% 94.29%

improve the SPNN inferencing accuracy. In addition, the results in this chapter indicate the importance of considering variations during the design-phase of SPNNs to facilitate the application of online and dynamic calibration mechanisms in these networks, which are often complex and power- and area-hungry.

Chapter 5

ProVAT: An Automated Design and Analysis Framework for Process-Variation-Resilient Design of Silicon Photonic Microring Resonators

5.1 Introduction

MRRs are essential building blocks in silicon photonic integrated circuits (PICs). MRR is a looped waveguide that selectively transmits or drops specific wavelengths of light, known as resonant wavelengths (λ_r), based on its dimensions and optical path length. This resonance behavior makes MRRs versatile building blocks for filtering and switching applications in PICs [137, 138]. The wavelength selectivity in MRRs, determined by waveguide dimensions, silicon-on-insulator (SOI) layer thickness, and ring radius, is critical for applications like dense wavelength-division multiplexing (DWDM) and computational photonics [74]. However, the resonant wavelength in MRRs is highly sensitive to fabrication-process variations (FPVs), which can significantly impact device performance.

FPVs arise from optical lithography imperfections, and can induce significant shifts in the resonant wavelength of MRRs. These variations often manifest as changes in SOI thickness, waveguide width, waveguide edge roughness, and sidewall slope [25]. Previous studies have demonstrated the substantial impact of FPVs on MRR performance, with even single nanometer variation in waveguide thickness leading to notable shifts in resonant wavelength [27]. To address this challenge, researchers have explored various strategies for mitigating the impact of FPVs. One promising approach is the use of adiabatic MRRs, which employ curved, tapered waveguides with gradually increasing widths towards the ring's center [53, 139, 140]. Adiabatic designs have been shown to enhance resilience to FPVs while also offering advantages such as rapid thermal tuning and reduced power consumption [140, 141]. Additionally, other techniques like post-fabrication trim-

ming, statistical design optimization, and the use of feedback control mechanisms have been investigated to further improve MRR robustness in the face of FPVs [142].While these techniques are well-established, a unified framework for designing variation-resilient MRRs while comprehensively understanding and accounting for FPVs effects remains elusive.

In this chapter, we introduce ProVAT (Process Variation Analysis Tool), a comprehensive framework for designing MRRs that are resilient to FPVs. ProVAT empowers designers to assess the impact of FPVs on critical MRR performance metrics like resonant wavelength and free spectral range (FSR). The tool not only analyzes MRR performance but also allows users to understand the impact of FPVs on individual waveguide properties like effective index and group index. Furthermore, ProVAT provides the flexibility to create custom FPV maps tailored to specific fabrication scenarios. Finally, through layout and design optimization under FPV, ProVAT enables designers to generate optimized, FPV-resilient MRRs that meet unique performance metrics such as Q-factor, extinction ratio (ER), 3-dB bandwidth, and the total resonant wavelength shift that can be afforded given tuning power budget, by exploring various MRR design parameters, including waveguide width, MRR radius, and gap, as outlined in Fig. 5.1.

The rest of this chapter is structured as follows. Section 5.2 details the ProVAT framework, encompassing FPVs map generation, analysis of their impact on waveguides and MRRs, and optimization strategies for FPV-resilient MRR design. Section 5.3 presents the fabrication and validation of an optimized MRR, demonstrating its improved tolerance under FPVs. Finally, Section 5.4 concludes the paper.

5.2 Proposed Design Framework

This section details our open-source ProVAT design framework (see Fig. 5.1) for designing robust adiabatic MRRs that are resilient to FPVs. We begin by examining how FPVs in waveguide width, SOI layer thickness, and other critical parameters affect fundamental optical properties like effective index (n_{eff}) and group index (n_{g}), which in turn influence key MRR performance metrics such as resonant wavelength, extinction ratio (ER), and FSR. Leveraging this understanding, we

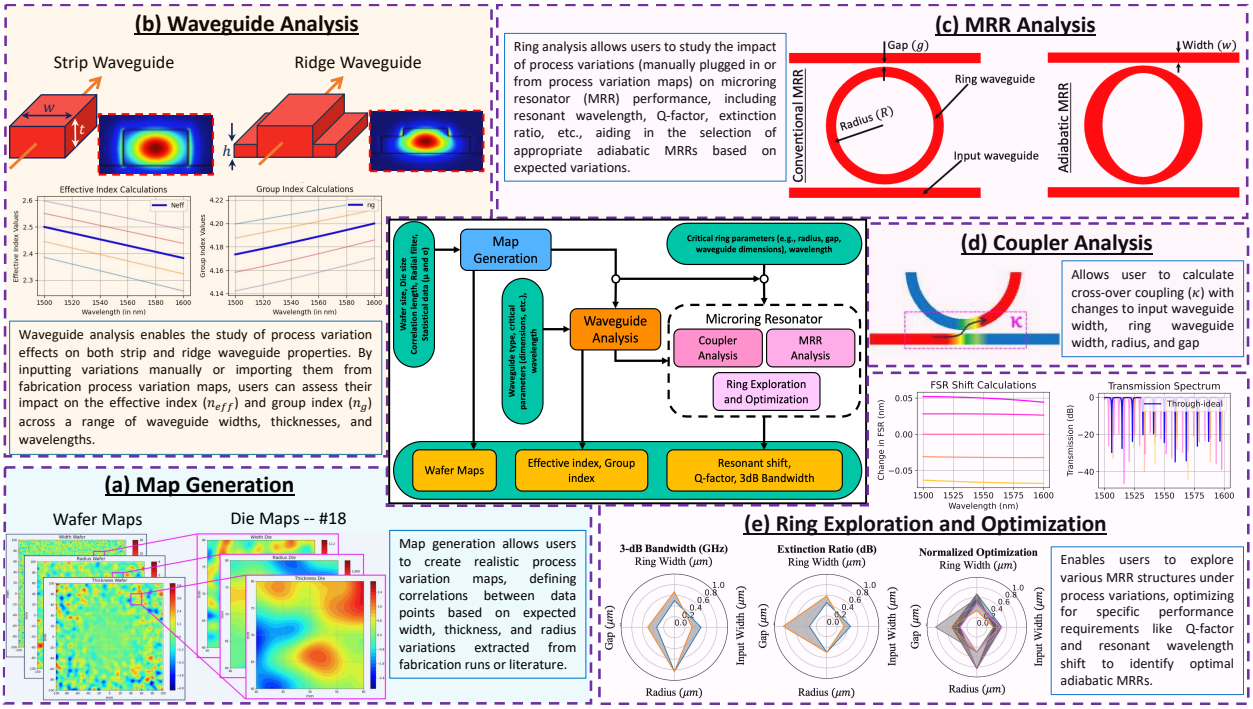


Figure 5.1: ProVAT workflow for designing FPV-resilient microring MRRs. The workflow encompasses: (a) generating realistic FPV maps across a wafer and individual dies, (b) analyzing the impact of FPVs on strip and ridge waveguides; (c) exploring adiabatic MRRs, considering coupling effects and variations on ring parameters; (d) calculating cross-over coupling coefficients for various waveguide dimensions; and (e) exploring and optimizing MRR designs to meet user-specified performance requirements. Top graphs in (e) show variations in free spectral range (FSR) and extinction ratio with ring radius, while bottom graphs demonstrate optimized values for specific design requirements.

systematically explore the design space in MRRs, exploring various design parameters to optimize performance and achieve an adiabatic MRR that is both tolerant to FPVs and tailored to specific user requirements.

5.2.1 FPV Maps

To design MRRs that are resilient to FPVs, it is crucial to understand how these variations occur not only within a single die but across the entire wafer. Building upon our previous work [3], we generate realistic FPV wafer maps using a procedure similar to that outlined in [33]. First, we create uncorrelated random variation maps for waveguide width, thickness, and radius. These maps have specified mean and standard deviation values, whose estimate can be extracted from literature [33] or obtained from fabrication foundries. Next, we convolve a Gaussian filter with a defined correlation length to these random distribution maps, resulting in correlated FPV maps as shown in Fig. 5.1(a). These wafer maps post convolution show spatial dependencies of different variations across the wafer. We further enhance these maps by incorporating radial variation effects, recognizing that variations tend to increase as one moves from the wafer's center to its edges [53, 55]. To model this, we ensure the least variation at the wafer center and progressively increase it towards the edges which is a close replica of variations in a fabricated wafer.

Our approach allows for the generation of diverse variation maps, with the option to include or exclude radial effects. Notably, users can customize these maps using mathematical models to simulate specific test cases. These maps are provided as *.csv* files, where the data represents the magnitude of the variation at specific locations on the wafer. For illustration, Fig. 5.1(a) shows a 200-mm wafer map with a randomly selected die (18) expanded to reveal these variations. This flexibility, along with the accessible data format, empowers designers to tailor FPV maps to their unique requirements.

5.2.2 Waveguide Analysis

In this section, we investigate how FPVs impact the properties of optical waveguides—structures that confine and guide light in MRRs—and the resulting effects on key performance metrics. We

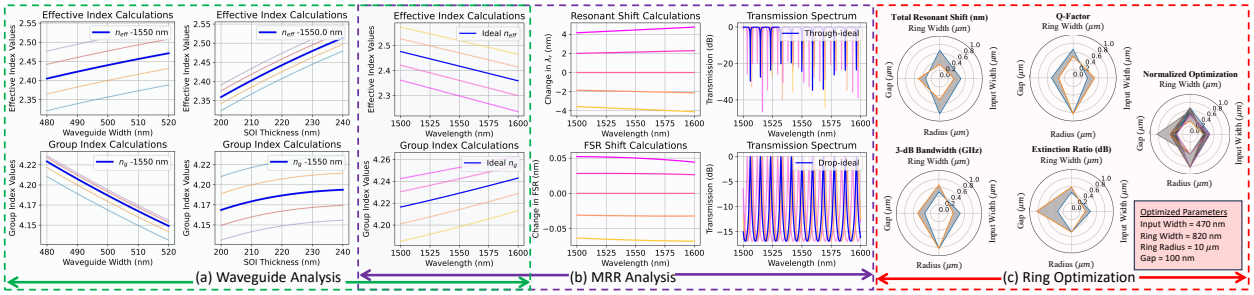


Figure 5.2: ProVAT workflow showcasing: (a) the effect of width and thickness variations on n_{eff} and n_{g} for strip/ridge waveguides, (b) the impact of FPVs on MRR performance metrics (resonant wavelength shift ($\Delta\lambda_r$), FSR, transmission spectrum), and (c) selection of optimal adiabatic MRR design parameters (see figure inset) aimed to reduce impact of $\Delta\lambda_r$ under FPVs.

focus on two common waveguide types: strip waveguides (rectangular cross-sections) and ridge waveguides (strip waveguides with a partial etch). For both types, we analyze how variations in waveguide width (w), SOI thickness (t), and slab thickness (h) influence the effective index (n_{eff}) and group index (n_{g}) (n_{eff} is the ratio of the speed of light in a vacuum to the phase velocity of light propagating in a specific mode within the waveguide and n_{g} is measure of how the group velocity of light, changes with wavelength within the waveguide). n_{eff} and n_{g} are crucial parameters in determining the behavior of light within the waveguide and, ultimately, the performance of MRRs. To conduct this analysis, we employ a modified Marcatili method [143], an approximate technique for calculating the propagation modes of waveguides, as detailed in [46]. Notably, the effective index in a waveguide depends not only on its dimensions but also on the wavelength of the propagating light [12]. This relationship can be expressed as:

$$n_{\text{eff}}(\lambda, w, t, h) = \left(\frac{\lambda}{2\pi}\right) \beta(\lambda, w, t, h), \tag{5.1}$$

where β is the propagation constant and λ is the optical wavelength. By considering these relationships, we can better assess the impact of FPVs on the performance of silicon photonic devices.

By systematically varying waveguide dimensions within the ranges defined by our FPV maps (or manually specified ranges), we investigate the resulting changes in n_{eff} and n_g, as illustrated in Fig. 5.2(a). These analyses provide crucial insights into how variations in w, t, and λ impact

the propagation of light within the waveguide. Our findings reveal distinct trends in the behavior of n_{eff} and n_g with respect to these parameters, aligning closely with those reported in existing literature [26]. This validation underscores the accuracy of our models and reinforces their utility for predicting waveguide behavior under realistic fabrication scenarios. By offering users the flexibility to input a wide range of parameter values, our approach enables a comprehensive understanding of the trends and changes in waveguide parameters, facilitating informed design decisions for robust MRR design.

5.2.3 MRR Analysis

To assess the impact of FPVs on MRR performance, we focus on analyzing critical characteristics such as resonant wavelength (λ_r), cross-over coupling (κ), FSR, 3-dB bandwidth, and ER. λ_r is a fundamental parameter defining the specific wavelength of light at which the MRR resonates. This resonance phenomenon occurs when the optical path length of the ring waveguide matches an integer multiple of the incident light's wavelength, leading to constructive interference within the ring. λ_r depends on various design parameters, including the ring radius, and the waveguide dimensions (width and SOI thickness). We utilize established mathematical models [12] to calculate λ_r using the n_{eff} obtained from waveguide analysis:

$$\lambda_r(w, t, h, R) = \left(\frac{2\pi R}{m}\right) n_{eff}(\lambda_r, w, t, h). \tag{5.2}$$

Here, R is the radius of the MRR and m is the integer that denotes the order of the resonance mode. κ is the fraction of light power transferred between the input waveguide and the ring. It is a crucial parameter that determines the magnitude of resonance and affects the shape of the transmission spectrum. The FSR, on the other hand, is the spacing between adjacent resonant wavelengths and is influenced by the ring radius and n_g. The transmission spectrum, which describes how the intensity of transmitted light varies with wavelength, reveals crucial information about the MRR's performance, including its 3-dB bandwidth (range of wavelengths around λ_r where the transmitted power is half of the peak power), Q-factor (parameter that quantifies the sharpness of the resonance,

indicating how well the MRR can store energy at a specific wavelength), and ER (The ratio of the maximum to minimum transmitted power in the MRR's transmission spectrum). Any deviation in λ_r can be referred to as resonant wavelength shifts ($\Delta\lambda_r$), which quantifies the total deviation of λ_r from its original design point and can be written as:

$$\Delta\lambda_R(w,t,h,R) = \left(\frac{2\pi R}{m}\right)\Delta n_{eff}(\lambda_R,w,t,h). \tag{5.3}$$

Based on (5.2) and (5.3), our goal is to minimize $\Delta\lambda_r$ caused by FPVs while maintaining user identified performance metrics, such as FSR, Q-factor, and ER. FPVs not only impact λ_r but also introduce additional losses as optical signals propagate through the MRR and couple in and out of it. These losses can significantly degrade the efficiency of MRR-based PICs. Therefore, it is crucial to monitor the cross-over coupling coefficient (κ), which quantifies the coupling strength between the input/drop waveguide and the ring resonator. κ is a key parameter for understanding MRR behavior and is typically determined using precise numerical methods like Finite-Difference Time-Domain (FDTD) simulations across a range of ring radii.

Building upon our previous work in [3], where we examined the impact of ring parameters like waveguide width on Q-factor, 3-dB bandwidth, and $\Delta\lambda_r$, ProVAT further extends to FSR and ER analysis by also exploring radius and gap along with other ring parameters for a comprehensive overview of impact on ring performance. We go beyond analyzing these metrics in ideal scenarios by incorporating the variation analysis from previous sections. This allows us to understand how FPVs affect not only the nominal values of these metrics but also their deviations under realistic FPV. Fig. 5.2(b) illustrates the changes in n_{eff} and n_{g} caused by FPVs and their corresponding impact on $\Delta\lambda_r$, FSR, and ER. By quantifying these effects, we can take a first step towards designing MRRs robust to FPVs, ensuring reliable performance and accurate reflection of expected metrics even in the presence of FPVs.

One strategy to mitigate the impact of FPVs on MRR performance is to increase the waveguide width [3]. Wider waveguides offer better mode confinement, reducing propagation loss (the loss of light as it travels along the waveguide) within the MRR. However, our analysis reveals that

increasing the waveguide width can also decrease κ, which represents the fraction of light power transferred between the input waveguide and the ring resonator. A lower κ value can lead to weaker resonances and less efficient power transfer. Using our framework ProVAT, we can address this trade-off much more efficiently. Not only designing appropriate adiabatic MRRs [140] to meet user performance metric but also aiming for a peripheral view of the impact of FPVs on such design and performance metrics boosting validation of the proposed design. In an adiabatic MRR, the waveguide width is gradually increased, starting with a narrow width at the coupling region to maximize κ, and then smoothly increasing towards the edges of the ring to enhance mode confinement. This design allows for both high coupling efficiency and reduced propagation loss, thereby improving the overall performance and tolerance of the MRR to FPVs. To showcase the unique advantages of our ProVAT, we present results in Section 5.3 where, FPVs tolerance in MRRs is enhanced without sacrificing other essential performance parameters. This demonstrates a significant advancement in MRR design.

5.3 Results and Discussions

To validate the versatility and effectiveness of ProVAT, we present a comprehensive case study of an MRR with optimized parameters shown in Fig. 5.2(c). We demonstrate the framework's ability to optimize MRR designs for improved FPVs tolerance while maintaining exceptional ER and Q-factor performance. This validation is further strengthened by fabricating and comparing the performance of optimized adiabatic MRRs against conventional MRR design. For this case study, we aim to achieve minimum $\Delta\lambda_r$ of 0–1 nm, maintaining Q-factor between 5000–20000, and an ER of 15–40 dB. These specifications align with photonic computing applications that require high FSR values for improved wavelength division multiplexing (WDM) properties [74]. In addition, we target a ring radius of 5–15 μm, considered as an example.

The design process for FPV-tolerant MRRs starts with creating FPV maps that consider expected variations in ring parameters, based on the findings in [33]. Since waveguide parameter variations directly affect n_{eff} and overall MRR performance, we utilize the total resonant wave-

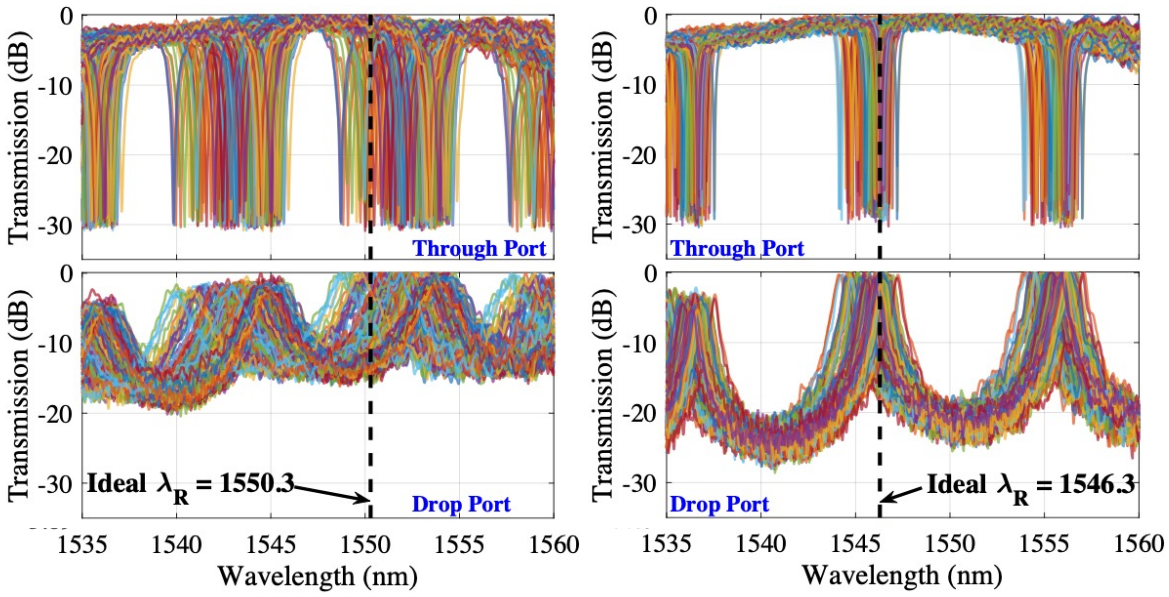
length shift $T_{\Delta\lambda_r}$ as a key metric to evaluate the combined impact of these variations on the MRR's resonant wavelength:

$$T_{\Delta\lambda_r}(\lambda_r, w', t', R') = \frac{\partial\lambda_r}{\partial w}(\sigma_w) + \frac{\partial\lambda_r}{\partial t}(\sigma_t) + \frac{\partial\lambda_r}{\partial R}(\sigma_R). \tag{5.4}$$

Here, $\partial\lambda_r/\partial w, t, R$, capture the sensitivity of λ_r to changes in waveguide width (w), SOI thickness (t), and radius (R), respectively. The standard deviations σ_w, σ_t, and σ_R are derived from the FPV maps. As noted in [3], a low $T_{\Delta\lambda_r}$, achievable by increasing the waveguide width, leads to a minimal λ_r shift.

Next, we automatically explore the MRR design space by systematically varying ring parameters which is set by the user at the start of the analysis. The parameters of interest are, the width of the input waveguide (the waveguide that couples light into the ring) and the ring waveguide (the waveguide forming the ring itself), as well as the gap between them and the ring radius. For each configuration, we calculate $T_{\Delta\lambda_r}$, Q-factor, ER, and 3-dB bandwidth. Figure 5.2(c) illustrates this exploration, highlighting feasible designs (gray) that meet designer requirements. Normalized results aid readability. The overlapping region reveals optimal designs, leading us to select a ring with an input waveguide width of 470 nm, ring waveguide width of 800 nm, gap of 100 nm, and a radius of 10 μm. This ring is further designed adiabatically to improve coupling while maintaining low tolerance to FPV. To validate our findings, we fabricate an MRR with this optimized design.

Conventional and adiabatic MRRs were designed with the same overall size and near-identical nominal resonant wavelengths (1550 nm conventional, 1546 nm adiabatic) to ensure similar FPVs exposure. This wavelength difference does not affect our analysis. We strategically placed 268 identical conventional MRRs and 289 identical adiabatic MRRs across a 10×10 mm chip, fabricated using a standard E-beam multi-project wafer process at Applied Nanotools Inc. [73]. Each conventional-adiabatic MRR pair was placed close to each other to ensure similar impact on both MRRs due to FPVs. In-house testing was conducted at a constant chip stage temperature of 300 K to minimize thermal variations. With an input power of 7.5 dBm, we observed a total output loss



(a) Conventional MRR **(b)** Adiabatic MRR

Figure 5.3: Through-port and drop-port response in (a) conventional and (b) adiabatic MRRs. Black-dotted lines show the ideal (nominal) resonant wavelengths.

of 25.2 dB. Grating couplers contributed 17.6 dB (8.8 dB each), and per-device loss was estimated at ≈1 dB, leaving ≈6.6 dB unaccounted for, likely due to test and alignment losses.

Fig. 5.3 reveals that adiabatic MRRs demonstrate a through- and drop-port response up to 70% closer to the ideal response at 1546.3 nm compared to conventional MRRs (ideal response at 1550.3 nm). Table 5.1 further quantifies this performance advantage across multiple metrics, confirming the effectiveness of our proposed tool for creating robust MRRs. This successful validation through fabrication and testing demonstrates the practical utility of our framework in achieving high-performance, FPV-tolerant MRR designs.

5.4 Conclusion

This chapter introduced ProVAT, a comprehensive and user-friendly tool that helps design process-variation-resilient adiabatic MRRs. ProVAT's systematic exploration of critical design parameters and performance metrics enables users to develop robust and efficient MRR designs that meet specific fabrication constraints and performance targets. By thoroughly analyzing the impact of FPVs on key metrics like resonant wavelength, Q-factor, and ER, ProVAT provides

Table 5.1: Characterized device performance (Avg.: Average, SD: Standard Deviation)

	MRR1 (Conventional)		**MRR2 (Adiabatic)**	
	Through	**Drop**	**Through**	**Drop**
Avg. λ_r	1552.8 nm		1546.1 nm	
SD λ_r ($\propto \delta\lambda_r$)	1.3 nm		0.5 nm	
Avg. Q-factor	3567	590	10067	790
Avg. ER	27.7 dB	12.8 dB	25 dB	21.8 dB
$\sigma\lambda_r/FSR$	0.15		0.06	

invaluable insights for achieving reliable post-fabrication performance. This comprehensive approach, integrating waveguide and ring parameters along with detailed FPVs modeling, represents a significant advancement in the design of high-performance, FPV-resilient MRRs, all within a single, streamlined tool.

Chapter 6

Conclusion and Future Work

In this thesis, we have traversed the complete cycle of photonic microring resonator development, from design and optimization to fabrication, testing, characterization, and analysis under the influence of fabrication process variations (FPVs). We began by establishing a framework for realistically modeling FPV maps that replicate process variations across both wafers and individual dies. This crucial step allowed us to understand how such variations can significantly affect the overall device performance, yield, and reliability of any photonic device. We also explored the diverse types of geometric variations that can occur within a wafer and between wafers, establishing a foundation for our subsequent investigations.

Building on this foundation, we systematically analyzed the impact of FPVs on a wide range of photonic devices, from single waveguides to both passive and active microring resonators (MRRs). By conducting a comprehensive design space exploration, we not only identified the key metrics that influence device performance but also developed opportunities to optimize designs in the presence of variations. Our computationally efficient analytical models, specifically tailored to capture the effects of physical-level variations on MRR device-level performance, enabled an exhaustive exploration of the MRR design space. This led to the development of optimal MRR designs that exhibit high tolerance to FPVs while maintaining the desired quality factor and 3dB bandwidth performance – all of which can be determined during the design phase. The results presented in this thesis provide valuable insight for silicon photonic designers, enabling them to proactively consider FPVs during design and ultimately reduce tuning power consumption and improve circuit yield after fabrication.

To validate our theoretical findings, we performed an experimental analysis comparing conventional and adiabatic MRR designs. Our results consistently demonstrated the superior performance of adiabatic MRRs across multiple metrics. Notably, adiabatic MRRs achieved a remarkable 70% alignment with their ideal resonant wavelengths, showcasing their robustness to fabrication varia-

tions. Additionally, they outperformed conventional MRRs in terms of uniformity in frequency response, Q-factor, extinction ratio (ER), and free spectral range (FSR). This exceptional uniformity makes adiabatic MRRs particularly promising for photonic integrated circuit (PIC) applications in both datacom and computation, where high inter-device matching is essential. By simplifying device placement, routing, and tuning processes, adiabatic MRRs have the potential to revolutionize PIC design and contribute to more efficient and reliable photonic systems.

Upon building confidence in our optimized design methods, we successfully applied our MRR design optimization to various applications. In one instance, we developed a two-channel wavelength-selective MRR-based demultiplexer that demonstrated exceptional channel-spacing accuracy within 0.5 nm, even when the MRRs were positioned 500 μm apart on a chip. This achievement underscores the practical utility and robustness of our optimization techniques.

Furthermore, we extended our optimized rings to the realm of photonic accelerators, culminating in the development of CrossLight, a novel cross-layer optimized photonic neural network accelerator. By integrating device-level fabrication-driven optimizations seamlessly with circuit-level and architecture-level optimizations, CrossLight achieved remarkable energy efficiency and performance. Notably, it demonstrated a 9.5$\times$ reduction in energy-per-bit and a 15.9$\times$ improvement in performance-per-watt compared to state-of-the-art photonic DNN accelerators. CrossLight even surpassed several CPU, GPU, and custom electronic accelerator platforms, showcasing the potential of cross-layer optimization strategies to address critical challenges such as crosstalk, fabrication process variations, high laser power, and excessive tuning power. The results presented in this thesis highlight the transformative potential of photonic DNN accelerators to meet the growing demand for energy-efficient and high-performance deep neural network acceleration.

In another application, we proposed ROBIN, an optical-domain binarized neural network (BNN) accelerator. Through meticulous optimizations at the device, circuit, and architecture levels, ROBIN achieved significant energy and area savings while improving overall throughput. In particular, we developed two variants of ROBIN: ROBIN_EO, optimized for energy and area efficiency, and ROBIN_PO, prioritizing higher frames-per-second (FPS) performance. Our simulation analysis

revealed that ROBIN outperforms various state-of-the-art optical neural network accelerators in terms of energy-per-bit (EPB). While electronic BNN accelerators exhibit lower power consumption, both ROBIN variants demonstrated superior throughput, underscoring ROBIN's potential for accelerating BNN model execution on resource-constrained platforms.

For the final application, we explored the impact of fabrication process variations (FPVs) on coherent silicon photonic neural networks (SPNNs). By analyzing the influence of FPVs in waveguide width, SOI thickness, and slab thickness, we modeled the resulting optical phase noises at the Mach-Zehnder interferometer (MZI) device level. We then investigated how these phase noises propagate through optical unitary multipliers built using MZIs and contribute to performance degradation at the network level, as measured by the relative variation distance (RVD). To mitigate these effects, we proposed physical-level design optimization solutions that significantly enhance MZI device tolerance under correlated FPVs in SPNNs. Our simulation results, incorporating realistic FPV maps for two SPNN case studies of varying sizes, demonstrated the effectiveness of our proposed optimizations in improving SPNN inferencing accuracy. These findings emphasize the importance of considering variations during the design phase of SPNNs, facilitating the implementation of online and dynamic calibration mechanisms that are often complex and resource-intensive.

Finally, to ensure that our photonic design methodologies are accessible to a wide range of photonic designers, we introduced ProVAT, a comprehensive and user-friendly tool that empowers users to design process-variation-resilient adiabatic MRRs. ProVAT's systematic exploration of critical design parameters and performance metrics allows users to develop robust and efficient MRR designs that meet specific fabrication constraints and performance targets. By meticulously analyzing the impact of FPVs on key metrics such as resonant wavelength, Q-factor, and extinction ratio (ER), ProVAT provides invaluable insights for achieving reliable post-fabrication performance. This comprehensive approach, integrating waveguide and ring parameters with detailed FPV modeling, represents a significant advancement in the design of high-performance, FPV-resilient MRRs, all consolidated within a single, streamlined tool.

In conclusion, this thesis has presented a comprehensive exploration of photonic microring resonator design, optimization, and application under the influence of fabrication process variations. Our work has not only deepened the understanding of FPVs and their impact on device performance but has also yielded innovative design methodologies, optimization techniques, and practical tools that empower designers to create robust, high-performance photonic systems. The advancements presented in this thesis pave the way for the widespread adoption of photonic technologies in diverse fields, from telecommunications and data centers to artificial intelligence and scientific computing. By addressing the challenges posed by fabrication variations, we have contributed to the development of more reliable, efficient, and scalable photonic solutions that hold the promise of transforming the technological landscape.

While this thesis focused on silicon photonics, the field of photonic integrated circuits (PICs) encompasses a wider range of material platforms, each with unique advantages and challenges. Future research could explore alternative platforms such as silicon nitride (SiN), which has gained significant attention for its applications in non-linear optics, telecommunications, and low-loss optical gyroscopes.

A comprehensive design space exploration for SiN-based MRRs would be invaluable. This would involve systematically investigating the impact of various design parameters (e.g., waveguide dimensions, ring radius, coupling gap) and fabrication process variations on key performance metrics (e.g., Q-factor, extinction ratio, free spectral range). By developing computationally efficient models that accurately capture the behavior of SiN devices, researchers could identify optimal designs that meet specific application requirements while exhibiting robustness to fabrication variability. Integrating these findings into the ProVAT tool would significantly enhance its versatility. Users could then leverage ProVAT's capabilities to not only optimize MRR designs for a given material platform but also to compare and select the most suitable platform for their specific application, considering trade-offs between performance, cost, and fabrication complexity.

The current work was limited by the 220 nm thickness constraint imposed by the fabrication foundry. However, high-aspect-ratio waveguides, with thicknesses beyond this limit, hold promise

for further optimization and novel applications. For instance, increasing the waveguide thickness can enhance the confinement of transverse magnetic (TM) modes, potentially leading to improved performance in TM-based devices.

Future research could systematically explore the design space of high-aspect-ratio waveguides, investigating the impact of thickness variations on various device performance metrics. This would involve developing accurate models that capture the complex behavior of light propagation in such waveguides, as well as considering fabrication challenges associated with high-aspect-ratio structures. The insights gained from this exploration could open up new avenues for designing and optimizing PICs for a wider range of applications, including sensing, biosensing, and non-linear optics.

While our work focused primarily on passive MRR designs, active devices, such as those incorporating thermo-optic or electro-optic tuners, are crucial for many practical applications. Future research could extend our design space exploration and optimization techniques to active devices. This would involve analyzing the impact of fabrication variations on the interaction between MRRs and tuners, considering factors such as the distance between the MRR and tuner, the materials used, and the specific tuning mechanism. By optimizing these parameters, researchers could develop active MRR designs that are not only high-performance but also robust to fabrication variability.

Additionally, the experimental results in this thesis revealed higher-than-expected reflections in the fabricated ring resonators. Future work could focus on mitigating these reflections through further design optimization. This could involve modifying the ring geometry, adjusting the coupling gap, or incorporating anti-reflection coatings. Reducing reflections would improve the overall efficiency and performance of the MRR devices, leading to lower power consumption and improved signal integrity.

By addressing these diverse research directions, future work has the potential to significantly advance the field of photonic integrated circuits. The development of comprehensive design tools, optimized material platforms, and robust device designs would accelerate the adoption of PICs

in a wide range of applications, ultimately contributing to more powerful, efficient, and reliable photonic systems.

Bibliography

[1] Lukas Chrostowski and Michael Hochberg. *Silicon Photonics Design: From Devices to Systems*. Cambridge University Press, 2015.

[2] 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.

[3] Asif Mirza, Febin Sunny, Peter Walsh, Karim Hassan, Sudeep Pasricha, and Mahdi Nikdast. Silicon photonic microring resonators: A comprehensive design-space exploration and optimization under fabrication-process variations. *IEEE TCAD*, 2021.

[4] Attila Mekis, Chris Poulton, Peter T Rakich, Gautam Bahl, Erich Buckley, Aharon Biberman, Lin Chang, Xingchen Jiang, Hao Liu, Tyson Mansfield, et al. A guide to the design, fabrication and characterization of photonic integrated circuits. *Optical Express*, 25(12):13678–13704, 2017.

[5] David Thomson, Andrew Zilkie, John E Bowers, Tin Komljenovic, Graham T Reed, Laurent Vivien, Delphine Marris-Morini, Eric Cassan, Leon Wehner, Sébastien Tanzilli, et al. Roadmap on silicon photonics. *Journal of Optics*, 18(7):073003, 2016.

[6] Michael Hochberg and Tom Baehr-Jones. Towards fabless silicon photonics. *Nature Photonics*, 4(8):492–494, 2010.

[7] Bahram Jalali and Sasan Fathpour. Silicon photonics. *Journal of Lightwave Technology*, 24(12):4600–4615, 2006.

[8] Graham T Reed and Andrew P Knights. *Silicon Photonics: An Introduction*. John Wiley Sons, 2004.

[9] Hadi Esmaeilzadeh, Emily Blem, Renée St. Amant, Karthikeyan Sankaralingam, and Doug Burger. Dark silicon and the end of multicore scaling. In *2011 38th Annual International Symposium on Computer Architecture (ISCA)*, pages 365–376. IEEE, 2011.

[10] David AB Miller. Device scaling challenges for optical interconnects. *Proceedings of the IEEE*, 97(7):1166–1185, 2010.

[11] David AB Miller. Device requirements for optical interconnects to silicon chips. *Proceedings of the IEEE*, 97(7):1166–1185, 2009.

[12] 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 and Photonics Reviews*, 6(1):47–73, 2012.

[13] M Lipson. *Optical waveguide theory and practice*. Optical Waveguide Theory and Technology, 2005.

[14] Theodor Tamir, editor. *Integrated optics*, volume 7. Springer Science & Business Media, 1973.

[15] Laurent Vivien and Lorenzo Pavesi. Ge on si integrated photonics. In *Handbook of Silicon Photonics*, pages 1013–1062. CRC Press, 2016.

[16] Janibul Bashir, Eldhose Peter, and Smruti R Sarangi. A survey of on-chip optical interconnects. *ACM Computing Surveys (CSUR)*, 51(6):1–34, 2019.

[17] Mahdi Nikdast, Sudeep Pasricha, Gabriela Nicolescu, Ashkan Seyedi, and Di Liang. *Silicon photonics for high-performance computing and beyond*. CRC Press, 2021.

[18] Kyle Shiflett, Dylan Wright, Avinash Karanth, and Ahmed Louri. Pixel: Photonic neural network accelerator. In *2020 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pages 474–487. IEEE, 2020.

[19] Charles Roques-Carmes, Yichen Shen, Cristian Zanoci, Mihika Prabhu, Fadi Atieh, Li Jing, Tena Dubček, Chenkai Mao, Miles R Johnson, Vladimir Čeperić, et al. Heuristic recurrent algorithms for photonic ising machines. *Nature communications*, 11(1):249, 2020.

[20] William A Zortman, Douglas C Trotter, and Michael R Watts. Silicon photonics manufacturing. *Optics express*, 18(23):23598–23607, 2010.

[21] Laurent Vivien and Lorenzo Pavesi, editors. *Handbook of silicon photonics*. CRC Press, 2013.

[22] Yufei Xing, Jiaxing Dong, Umar Khan, and Wim Bogaerts. Capturing the effects of spatial process variations in silicon photonic circuits. *ACS Photonics*, 10(4):928–944, 2022.

[23] Duane S Boning, Sally I El-Henawy, and Zhengxing Zhang. Variation-aware methods and models for silicon photonic design-for-manufacturability. *Journal of Lightwave Technology*, 40(6):1776–1783, 2022.

[24] Sudip Shekhar, Wim Bogaerts, Lukas Chrostowski, John E Bowers, Michael Hochberg, Richard Soref, and Bhavin J Shastri. Roadmapping the next generation of silicon photonics. *Nature Communications*, 15(1):751, 2024.

[25] Erik Rosseel, Luis Fernandez, Martin Tabat, Wim Bogaerts, John Hautala, and Philippe Absil. SOI thickness uniformity improvement using wafer-scale corrective etching for silicon nano-photonic device. In *IEEE Photonics Society*, pages 289–292, 2011.

[26] Mahdi Nikdast, Gabriela Nicolescu, Jelena Trajkovic, and Odile Liboiron-Ladouceur. Modeling fabrication non-uniformity in chip-scale silicon photonic interconnects. In *IEEE/ACM Design, Automation and Test in Europe Conference and Exhibition*, pages 115–120, 2016.

[27] Wim Bogaerts, Yufei Xing, and Umar Khan. Layout-aware variability analysis, yield prediction, and optimization in photonic integrated circuits. *IEEE Journal of Selected Topics in Quantum Electronics*, 25(5):1–13, 2019.

[28] Nur Musyiirah Haji Masri, M Rakib Uddin, and Law Foo Kui. WDM system based on radius variation of photonic microring resonators. In *IEEE Student Conference on Research and Development*, pages 243–246, 2017.

[29] Sanmitra Banerjee, Mahdi Nikdast, and Krishnendu Chakrabarty. Modeling silicon-photonic neural networks under uncertainties. *IEEE Journal of Selected Topics in Quantum Electronics*, 28(1):1–12, 2022.

[30] Sally I El-Henawy, Ryan Miller, and Duane S Boning. Effects of a random process variation on the transfer characteristics of a fundamental photonic integrated circuit component. In *Optical Modeling and Performance Predictions X*, volume 10743, pages 162–171. SPIE, 2018.

[31] Yufei Xing, Jiaxing Dong, Umar Khan, and Wim Bogaerts. Hierarchical model for spatial variations of integrated photonics. In *IEEE International Conference on Group IV Photonics (GFP)*, pages 1–2, 2018.

[32] Sherief Reda and Sani R Nassif. Analyzing the impact of process variations on parametric measurements: Novel models and applications. In *2009 Design, Automation & Test in Europe Conference & Exhibition*, pages 375–380. IEEE, 2009.

[33] Zeqin Lu, Jaspreet Jhoja, Jackson Klein, Xu Wang, Amy Liu, Jonas Flueckiger, James Pond, and Lukas Chrostowski. Performance prediction for silicon photonics integrated circuits with layout-dependent correlated manufacturing variability. *Optics express*, 25(9):9712–9733, 2017.

[34] Robert L Wagner, Jiming Song, and Weng C Chew. Monte carlo simulation of electromagnetic scattering from two-dimensional random rough surfaces. *IEEE Transactions on Antennas and Propagation*, 45(2):235–245, 1997.

[35] Yiwen Shen, Xiang Meng, Qixiang Cheng, Sébastien Rumley, Nathan Abrams, Alexander Gazman, Evgeny Manzhosov, Madeleine Strom Glick, and Keren Bergman. Silicon photonics for extreme scale systems. *Journal of Lightwave Technology*, 37(2):245–259, 2019.

[36] Zhiping Zhou, Ruixuan Chen, Xinbai Li, and Tiantian Li. Development trends in silicon photonics for data centers. *Optical Fiber Technology*, 44:13–23, 2018.

[37] 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):441–446, 2017.

[38] Thomas Yong Long Ang, Soon Thor Lim, Shuh Ying Lee, Ching Eng Png, and Mee Koy Chin. How small can a microring resonator be and yet be polarization independent? *Applied optics*, 48(15):2821–2835, 2009.

[39] 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. *Optics Express*, 18(18):19055–19063, 2010.

[40] CL Manganelli, Paolo Pintus, Fabrizio Gambini, D Fowler, M Fournier, Stefano Faralli, C Kopp, and CJ Oton. Large-FSR thermally tunable double-ring filters for WDM applications in silicon photonics. *IEEE Photonics Journal*, 9(1):1–10, 2017.

[41] Ting Hu, Chen Qiu, Ping Yu, LongZhi Yang, WanJun Wang, XiaoQing Jiang, Mei Yang, Lei Zhang, and JianYi Yang. Silicon photonic network-on-chip and enabling components. *Science China Technological Sciences*, 56(3):543–553, 2013.

[42] Sudeep Pasricha and Mahdi Nikdast. A survey of silicon photonics for energy-efficient manycore computing. *IEEE Design and Test*, 37(4):60–81, 2020.

[43] Sai Vineel Reddy Chittamuru and Sudeep Pasricha. Improving crosstalk resilience with wavelength spacing in photonic crossbar-based network-on-chip architectures. In *IEEE International Midwest Symposium on Circuits and Systems*, pages 1–4, 2015.

[44] Meisam Bahadori, Mahdi Nikdast, Sébastien Rumley, Liang Yuan Dai, Natalie Janosik, Thomas Van Vaerenbergh, Alexander Gazman, Qixiang Cheng, Robert Polster, and Keren Bergman. Design space exploration of microring resonators in silicon photonic interconnects: Impact of the ring curvature. *IEEE Journal of Lightwave Technology*, 36(13):2767–2782, Jul 2018.

[45] Lukas Chrostowski, Xu Wang, Jonas Flueckiger, Yichen Wu, Yun Wang, and S Talebi Fard. Impact of fabrication non-uniformity on chip-scale silicon photonic integrated circuits. In *IEEE/OSA Optical Fiber Communication Conference*, pages Th2A–37, 2014.

[46] Mahdi Nikdast, Gabriela Nicolescu, Jelena Trajkovic, and Odile Liboiron-Ladouceur. Chip-scale silicon photonic interconnects: A formal study on fabrication non-uniformity. *Journal of Lightwave Technology*, 34(16):3682–3695, 2016.

[47] Meisam Bahadori, Alexander Gazman, Natalie Janosik, Sébastien Rumley, Ziyi Zhu, Robert Polster, Qixiang Cheng, and Keren Bergman. Thermal rectification of integrated microheaters for microring resonators in silicon photonics platform. *IEEE Journal of Lightwave Technology*, 36(3):773–788, 2017.

[48] Stefan Abel, Thilo Stöferle, Chiara Marchiori, Daniele Caimi, Lukas Czornomaz, Michael Stuckelberger, Marilyne Sousa, Bert J Offrein, and Jean Fompeyrine. A hybrid barium titanate–silicon photonics platform for ultraefficient electro-optic tuning. *IEEE Journal of Lightwave Technology*, 34(8):1688–1693, 2016.

[49] Fuwan Gan, Tymon Barwicz, MA Popovic, MS Dahlem, CW Holzwarth, PT Rakich, HI Smith, EP Ippen, and FX Kartner. Maximizing the thermo-optic tuning range of silicon photonic structures. In *IEEE Photonics in Switching*, pages 67–68, 2007.

[50] Hasitha Jayatilleka, Harel Frish, Ranjeet Kumar, John Heck, Chaoxuan Ma, Meer N Sakib, Duanni Huang, and Haisheng Rong. Post-fabrication trimming of silicon photonic ring resonators at wafer-scale. *Journal of Lightwave Technology*, 39(15):5083–5088, 2021.

[51] Jared C Mikkelsen, Wesley D Sacher, and Joyce KS Poon. Adiabatically widened silicon microrings for improved variation tolerance. *Opt. Express*, 22(8):9659–9666, 2014.

[52] Ying Luo, Xuezhe Zheng, Shiyun Lin, Jin Yao, Hiren Thacker, Ivan Shubin, John E Cunningham, Jin-Hyoung Lee, Stevan S Djordjevic, Jock Bovington, et al. A process-tolerant ring modulator based on multi-mode waveguides. *IEEE Photonics Technology Letters*, 28(13):1391–1394, 2016.

[53] Zhan Su, Ehsan S Hosseini, Erman Timurdogan, Jie Sun, Gerald Leake, Douglas D Coolbaugh, and Michael R Watts. Reduced wafer-scale frequency variation in adiabatic microring resonators. In *IEEE/OSA Optical Fiber Communication Conference*, pages 1–3, 2014.

[54] Raymond G Beausoleil, Andrei Faraon, David Fattal, Marco Fiorentino, Zhen Peng, and Charles Santori. Devices and architectures for large-scale integrated silicon photonics circuits. In *Optoelectronic Integrated Circuits XIII*, volume 7942, page 794204, 2011.

[55] Ashok V. Krishnamoorthy, Xuezhe Zheng, Guoliang Li, Jin Yao, Thierry Pinguet, Attila Mekis, Hiren Thacker, Ivan Shubin, Ying Luo, Kannan Raj, and John E. Cunningham. Exploiting CMOS manufacturing to reduce tuning requirements for resonant optical devices. *IEEE Photonics Journal*, 3(3):567–579, 2011.

[56] Xi Chen, Moustafa Mohamed, Zheng Li, Li Shang, and Alan R. Mickelson. Process variation in silicon photonic devices. *Applied optics*, 52(31):7638–7647, 2013.

[57] Yuyang Wang, Jared Hulme, Peng Sun, Mudit Jain, M Ashkan Seyedi, Marco Fiorentino, Raymond G Beausoleil, and Kwang-Ting Cheng. Characterization and applications of spatial variation models for silicon microring-based optical transceivers. In *IEEE ACM/IEEE Design Automation Conference*, pages 1–6, 2020.

[58] Yanir London, Thomas Van Vaerenbergh, Marco Piorentino, Ashkan Seyedi, Peng Sun, and Keren Bergman. Behavioral model of silicon photonics microring with unequal ring and bus widths. In *IEEE Optical Interconnects Conference*, pages 1–2, 2019.

[59] Danielius Kramnik et al. Fast-tuning adiabatic microrings for CROW filters and athermal WDM receivers in a 45 nm SOI CMOS process. In *IEEE/Optica CLEO*, pages SF4M–2, 2022.

[60] Akhilesh SP Khope. Ultralow loss adiabatic microring resonator with thermal tuning. 2021.

[61] GB Hocker and William K Burns. Mode dispersion in diffused channel waveguides by the effective index method. *Applied optics*, 16(1):113–118, 1977.

[62] Mahdi Nikdast, Gabriela Nicolescu, Jelena Trajkovic, and Odile Liboiron-Ladouceur. Photonic integrated circuits: A study on process variations. In *IEEE/OSA Optical Fiber Communication Conference*, pages W2A–22, 2016.

[63] Meisam Bahadori, Sébastien Rumley, Dessislava Nikolova, and Keren Bergman. Comprehensive design space exploration of silicon photonic interconnects. *IEEE Journal of Lightwave Technology*, 34(12):2975–2987, 2016.

[64] Meisam Bahadori, Sébastien Rumley, Hasitha Jayatilleka, Kyle Murray, Nicolas A. F. Jaeger, Lukas Chrostowski, Sudip Shekhar, and Keren Bergman. Crosstalk penalty in microring-based silicon photonic interconnect systems. *IEEE Journal of Lightwave Technology*, 34(17):4043–4052, 2016.

[65] MODE, Ansys Lumerical.

[66] Rainer Hainberger. Structural optimization of silicon-on-insulator slot waveguides. *IEEE photonics technology letters*, 18(24):2557–2559, 2006.

[67] Daniele Melati, Francesco Morichetti, and Andrea Melloni. A unified approach for radiative losses and backscattering in optical waveguides. *Journal of Optics*, 16(5):055502, 2014.

[68] Minh A Tran, Duanni Huang, Tin Komljenovic, Jonathan Peters, Aditya Malik, and John E Bowers. Ultra-low-loss silicon waveguides for heterogeneously integrated silicon/III-V photonics. *Applied Sciences*, 8(7), 2018.

[69] Yuguang Zhang, Xiao Hu, Daigao Chen, Lei Wang, Miaofeng Li, Peng Feng, Xi Xiao, and Shaohua Yu. Design and demonstration of ultra-high-Q silicon microring resonator based on a multi-mode ridge waveguide. *Optics letters*, 43(7):1586–1589, 2018.

[70] Bryce A Dorin and N Ye Winnie. Two-mode division multiplexing in a silicon-on-insulator ring resonator. *Optics express*, 22(4):4547–4558, 2014.

[71] Paolo Pintus, Michael Hofbauer, Costanza L Manganelli, Maryse Fournier, Sarat Gundavarapu, Olivier Lemonnier, Fabrizio Gambini, Laetitia Adelmini, Carl Meinhart, Christophe Kopp, et al. PWM-driven thermally tunable silicon microring resonators: design, fabrication, and characterization. *Laser & Photonics Reviews*, 13(9):1800275, 2019.

[72] FDTD, Ansys Lumerical.

[73] Applied Nanotools Inc.

[74] Febin Sunny, Asif Mirza, Mahdi Nikdast, and Sudeep Pasricha. Crosslight: A cross-layer optimized silicon photonic neural network accelerator. In *2021 58th ACM/IEEE Design Automation Conference (DAC)*, pages 1069–1074, 2021.

[75] Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, Rick Boyle, Pierre-luc Cantin, Clifford Chao, Chris Clark, Jeremy Coriell, Mike Daley, Matt Dau, Jeffrey Dean, Ben Gelb, Tara Vazir Ghaemmaghami, Rajendra Gottipati, William Gulland, Robert Hagmann, C. Richard Ho, Doug Hogberg, John Hu, Robert Hundt, Dan Hurt, Julian Ibarz, Aaron Jaffey, Alek Jaworski, Alexander Kaplan, Harshit Khaitan, Daniel Killebrew, Andy Koch, Naveen Kumar, Steve Lacy, James Laudon, James Law, Diemthu Le, Chris Leary,

Zhuyuan Liu, Kyle Lucke, Alan Lundin, Gordon MacKean, Adriana Maggiore, Maire Mahony, Kieran Miller, Rahul Nagarajan, Ravi Narayanaswami, Ray Ni, Kathy Nix, Thomas Norrie, Mark Omernick, Narayana Penukonda, Andy Phelps, Jonathan Ross, Matt Ross, Amir Salek, Emad Samadiani, Chris Severn, Gregory Sizikov, Matthew Snelham, Jed Souter, Dan Steinberg, Andy Swing, Mercedes Tan, Gregory Thorson, Bo Tian, Horia Toma, Erick Tuttle, Vijay Vasudevan, Richard Walter, Walter Wang, Eric Wilcox, and Doe Hyun Yoon. In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the 44th annual international symposium on computer architecture*, pages 1–12, 2017.

[76] Intel Movidius VPU, 2020.

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

[78] Sudeep Pasricha and Nikil Dutt. *On-chip communication architectures: system on chip interconnect*. Morgan Kaufmann, 2010.

[79] Amir Kavyan Kavyan Ziabari, Jose L Abellán, Rafael Ubal, Chao Chen, Ajay Joshi, and David Kaeli. Leveraging Silicon-Photonic NoC for Designing Scalable GPUs. In *Proceedings of the 29th ACM on International Conference on Supercomputing*, pages 273–282, 2015.

[80] David AB Miller. Meshing optics with applications. *Nature Photonics*, 11(7):403–404, 2017.

[81] 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):7430, 2017.

[82] Viraj Bangari, Bicky A Marquez, Heidi Miller, Alexander N Tait, Mitchell A Nahmias, Thomas Ferreira De Lima, Hsuan-Tung Peng, Paul R Prucnal, and Bhavin J Shastri. Dig-

ital electronics and analog photonics for convolutional neural networks (deap-cnns). *IEEE Journal of Selected Topics in Quantum Electronics*, 26(1):1–13, 2019.

[83] Weichen Liu, Wenyang Liu, Yichen Ye, Qian Lou, Yiyuan Xie, and Lei Jiang. Holylight: A nanophotonic accelerator for deep learning in data centers. In *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 1483–1488. IEEE, 2019.

[84] Zheng Zhao, Derong Liu, Meng Li, Zhoufeng Ying, Lu Zhang, Biying Xu, Bei Yu, Ray T Chen, and David Z Pan. Hardware-software co-design of slimmed optical neural networks. In *Proceedings of the 24th Asia and South Pacific Design Automation Conference*, pages 705–710, 2019.

[85] 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, 2020.

[86] Colin Pask. Generalized parameters for tunneling ray attenuation in optical fibers. *JOSA*, 68(1):110–116, 1978.

[87] Liangjun Lu, Xiaorui Li, Wei Gao, Xin Li, Linjie Zhou, and Jianping Chen. Silicon non-blocking $4\times$ 4 optical switch chip integrated with both thermal and electro-optic tuners. *IEEE Photonics Journal*, 11(6):1–9, 2019.

[88] Maziyar Milanizadeh, Douglas Aguiar, Andrea Melloni, and Francesco Morichetti. Canceling thermal cross-talk effects in photonic integrated circuits. *Journal of Lightwave Technology*, 37(4):1325–1332, 2019.

[89] HEAT, Ansys Lumerical.

[90] Souvaraj De, Ranjan Das, Ravi K Varshney, and Thomas Schneider. Design and simulation of thermo-optic phase shifters with low thermal crosstalk for dense photonic integration. *IEEE Access*, 8:141632–141640, 2020.

[91] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[92] https://github.com/google/qkeras.

[93] Ziliang Ruan, Yuntao Zhu, Pengxin Chen, Yaocheng Shi, Sailing He, Xinlun Cai, and Liu Liu. Efficient hybrid integration of long-wavelength vcsels on silicon photonic circuits. *Journal of Lightwave Technology*, 38(18):5100–5106, 2020.

[94] Ali Doguş Güngördü, Günhan Dündar, and Mustafa Berke Yelten. A high performance tia design in 40 nm CMOS. In *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE, 2020.

[95] Binhao Wang, Zhihong Huang, Wayne V Sorin, Xiaoge Zeng, Di Liang, Marco Fiorentino, and Raymond G Beausoleil. A low-voltage si-ge avalanche photodiode for high-speed and energy efficient silicon photonic links. *Journal of Lightwave Technology*, 38(12):3156–3163, 2019.

[96] L Hagedorn Frandsen, P Ingo Borel, YX Zhuang, Anders Harpøth, Morten Thorhauge, Martin Kristensen, Wim Bogaerts, Pieter Dumon, Roel Baets, Vincent Wiaux, et al. Ultralow-loss 3-db photonic crystal waveguide splitter. *Optics letters*, 29(14):1623–1625, 2004.

[97] Yi-Chou Tu, Po-Han Fu, and Ding-Wei Huang. High-efficiency ultra-broadband multi-tip edge couplers for integration of distributed feedback laser with silicon-on-insulator waveguide. *IEEE Photonics Journal*, 11(4):1–13, 2019.

[98] Sudeep Pasricha and Shirish Bahirat. Opal: A multi-layer hybrid photonic noc for 3d ics. In *16th Asia and South Pacific Design Automation Conference (ASP-DAC 2011)*, pages 345–350. IEEE, 2011.

[99] Hasitha Jayatilleka, Michael Caverley, Nicolas AF Jaeger, Sudip Shekhar, and Lukas Chrostowski. Crosstalk limitations of microring-resonator based wdm demultiplexers on soi. In *2015 IEEE Optical Interconnects Conference (OI)*, pages 48–49. IEEE, 2015.

[100] Erman Timurdogan, Cheryl M Sorace-Agaskar, Ehsan S Hosseini, Gerald Leake, Douglas D Coolbaugh, and Michael R Watts. Vertical junction silicon microdisk modulator with integrated thermal tuner. In *CLEO: 2013*, pages 1–2. IEEE, 2013.

[101] Matteo Pisati, Fernando De Bernardinis, Paolo Pascale, Claudio Nani, Marco Sosio, Enrico Pozzati, Nicola Ghittori, Federico Magni, Marco Garampazzi, Giacomino Bollati, et al. 6.3 a sub-250mw 1-to-56gb/s continuous-range pam-4 42.5 db il ADC/dac-based transceiver in 7nm finfet. In *2019 IEEE International Solid-State Circuits Conference-(ISSCC)*, pages 116–118. IEEE, 2019.

[102] Luan HK Duong, Mahdi Nikdast, Sebastien Le Beux, Jiang Xu, Xiaowen Wu, Zhehui Wang, and Peng Yang. A case study of signal-to-noise ratio in ring-based optical networks-on-chip. *IEEE Design & Test*, 31(5):55–65, 2014.

[103] Maurizio Capra, Beatrice Bussolino, Alberto Marchisio, Muhammad Shafique, Guido Masera, and Maurizio Martina. An updated survey of efficient hardware architectures for accelerating deep convolutional neural networks. *Future Internet*, 12(7):113, 2020.

[104] Febin P Sunny, Ebadollah Taheri, Mahdi Nikdast, and Sudeep Pasricha. A survey on silicon photonics for deep learning. *ACM Journal of Emerging Technologies in Computing System*, 17(4):1–57, 2021.

[105] Jeff Anderson, Shuai Sun, Yousra Alkabani, Volker Sorger, and Tarek El-Ghazawi. Photonic processor for fully discretized neural networks. In *2019 IEEE 30th International Conference on Application-specific Systems, Architectures and Processors (ASAP)*, volume 2160, pages 25–32. IEEE, 2019.

[106] Farzaneh Zokaee, Qian Lou, Nathan Youngblood, Weichen Liu, Yiyuan Xie, and Lei Jiang. Lightbulb: A photonic-nonvolatile-memory-based accelerator for binarized convolutional neural networks. In *2020 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 1438–1443. IEEE, 2020.

[107] Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. *arXiv preprint arXiv:1602.02830*, 2016.

[108] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

[109] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. *Advances in neural information processing systems*, 29, 2016.

[110] Yingjie Liu, Wenzhao Sun, Hucheng Xie, Nan Zhang, Ke Xu, Yong Yao, Shumin Xiao, and Qinghai Song. Adiabatic and ultracompact waveguide tapers based on digital metamaterials. *IEEE Journal of Selected Topics in Quantum Electronics*, 25(3):1–6, 2018.

[111] Qianfan Xu, David Fattal, and Raymond G Beausoleil. Silicon microring resonators with 1.5-μm radius. *Optics express*, 16(6):4309–4315, 2008.

[112] Brent E Little, Sai T Chu, Hermann A Haus, JAFJ Foresi, and J-P Laine. Microring resonator channel dropping filters. *Journal of lightwave technology*, 15(6):998–1005, 1997.

[113] Jinan Xia, Andrea Bianco, Edoardo Bonetto, and Roberto Gaudino. On the design of microring resonator devices for switching applications in flexible-grid networks. In *2014 IEEE International Conference on Communications (ICC)*, pages 3371–3376. IEEE, 2014.

[114] Tianshi Chen, Zidong Du, Ninghui Sun, Jia Wang, Chengyong Wu, Yunji Chen, and Olivier Temam. Diannao: A small-footprint high-throughput accelerator for ubiquitous machine-learning. *ACM SIGARCH Computer Architecture News*, 42(1):269–284, 2014.

[115] Junhua Shen, Akira Shikata, Lalinda D Fernando, Ned Guthrie, Baozhen Chen, Mark Maddox, Nikhil Mascarenhas, Ron Kapusta, and Michael CW Coln. A 16-bit 16-ms/s SAR ADC with on-chip calibration in 55-nm CMOS. *IEEE Journal of Solid-State Circuits*, 53(4):1149–1160, 2018.

[116] Bo Wu, Shuang Zhu, Benwei Xu, and Yun Chiu. A 24.7 mW 65 nm CMOS SAR-assisted ct $\Delta\Sigma$ modulator with second-order noise coupling achieving 45 MHz bandwidth and 75.3 db SNDR. *IEEE Journal of Solid-State Circuits*, 51(12):2893–2905, 2016.

[117] Eric Qin, Ananda Samajdar, Hyoukjun Kwon, Vineet Nadella, Sudarshan Srinivasan, Dipankar Das, Bharat Kaul, and Tushar Krishna. Sigma: A sparse and irregular gemm accelerator with flexible interconnects for dnn training. In *2020 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pages 58–70. IEEE, 2020.

[118] Stephen Cass. Taking ai to the edge: Google's tpu now comes in a maker-friendly package. *IEEE Spectrum*, 56(5):16–17, 2019.

[119] Tao Luo, Shaoli Liu, Ling Li, Yuqing Wang, Shijin Zhang, Tianshi Chen, Zhiwei Xu, Olivier Temam, and Yunji Chen. Dadiannao: A neural network supercomputer. *IEEE Transactions on Computers*, 66(1):73–88, 2016.

[120] Alessandro Aimar, Hesham Mostafa, Enrico Calabrese, Antonio Rios-Navarro, Ricardo Tapiador-Morales, Iulia-Alexandra Lungu, Moritz B Milde, Federico Corradi, Alejandro Linares-Barranco, Shih-Chii Liu, et al. Nullhop: A flexible convolutional neural network accelerator based on sparse representations of feature maps. *IEEE transactions on neural networks and learning systems*, 30(3):644–656, 2018.

[121] Peng Guo, Hong Ma, Ruizhi Chen, Pin Li, Shaolin Xie, and Donglin Wang. Fbna: A fully binarized neural network accelerator. In *2018 28th International conference on field programmable logic and applications (FPL)*, pages 51–513. IEEE, 2018.

[122] Yaman Umuroglu, Nicholas J Fraser, Giulio Gambardella, Michaela Blott, Philip Leong, Magnus Jahre, and Kees Vissers. Finn: A framework for fast, scalable binarized neural network inference. In *Proceedings of the 2017 ACM/SIGDA international symposium on field-programmable gate arrays*, pages 65–74, 2017.

[123] Michael Y-S Fang, Sasikanth Manipatruni, Casimir Wierzynski, Amir Khosrowshahi, and Michael R DeWeese. Design of optical neural networks with component imprecisions. *Opt. Express*, 27(10):14009–14029, 2019.

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

[125] 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, 2016.

[126] Farhad Shokraneh, Simon Geoffroy-Gagnon, and Odile Liboiron-Ladouceur. The diamond mesh, a phase-error- and loss-tolerant field-programmable MZI-based optical processor for optical neural networks. *Opt. Express*, 28:23495–23508, 2020.

[127] Ying Zhu, Grace Li Zhang, Bing Li, Xunzhao Yin, Cheng Zhuo, Huaxi Gu, Tsung-Yi Ho, and Ulf Schlichtmann. Countering variations and thermal effects for accurate optical neural networks. In *IEEE/ACM Int. Conf. Comput.*, pages 1–7. IEEE, 2020.

[128] Sanmitra Banerjee, Mahdi Nikdast, and Krishnendu Chakrabarty. Optimizing coherent integrated photonic neural networks under random uncertainties. In *IEEE/OSA OFC*, 2021.

[129] Saumil Bandyopadhyay et al. Hardware error correction for programmable photonics. *Optica*, 8(10):1247–1255, 2021.

[130] Farhad Shokraneh, Mohammadreza Sanadgol Nezami, and Odile Liboiron-Ladouceur. Theoretical and experimental analysis of a $4\times$ 4 reconfigurable mzi-based linear optical processor. *J. Light. Technol.*, 38(6):1258–1267, 2020.

[131] 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. *Opt. Express*, 28(8):12138–12148, 2020.

[132] Qixiang Cheng, Jihye Kwon, Madeleine Glick, Meisam Bahadori, Luca P Carloni, and Keren Bergman. Silicon photonics codesign for deep learning. *Proc. IEEE*, 108(8):1261–1282, 2020.

[133] Zeqin Lu, Han Yun, Yun Wang, Zhitian Chen, Fan Zhang, Nicolas AF Jaeger, and Lukas Chrostowski. Broadband silicon photonic directional coupler using asymmetric-waveguide based phase control. *Opt. Express*, 23(3):3795–3808, 2015.

[134] Guomin Jiang et al. Slab-modulated sidewall bragg gratings in silicon-on-insulator ridge waveguides. *IEEE Photon. Technol. Lett.*, 23(1):6–8, 2010.

[135] Yunfei Fu, Tong Ye, Weijie Tang, and Tao Chu. Efficient adiabatic silicon-on-insulator waveguide taper. *Photonics Res.*, 2(3):A41–A44, 2014.

[136] Shankar Kumar Selvaraja, Erik Rosseel, Luis Fernandez, Martin Tabat, Wim Bogaerts, John Hautala, and Philippe Absil. SOI thickness uniformity improvement using corrective etching for silicon nano-photonic device. In *IEEE Int. Conf. Group IV Photonics*, pages 71–73, 2011.

[137] Deepak Shekhawat et al. Design of ultra-compact and highly-sensitive graphene assisted silicon micro-ring resonator modulator for switching applications. *Silicon*, 14(8):4383–4390, 2022.

[138] Andrey A Nikitin et al. Optical bistable soi micro-ring resonators for memory applications. *Optics Communications*, 511, 2022.

[139] Michael R Watts et al. Adiabatic resonant microrings (arms) with directly integrated thermal microphotonics. In *Conference on Lasers and Electro-Optics*. Optica Publishing Group, 2009.

[140] Asif Mirza et al. Experimental analysis of adiabatic silicon photonic microring resonators under process variations. *IEEE PTL*, 36(6):417–420, 2024.

[141] Xiaoxi Wang et al. Wide-range and fast thermally-tunable silicon photonic microring resonators using the junction field effect. *Optics express*, 24(20):23081–23093, 2016.

[142] Hasitha Jayatilleka et al. Post-fabrication trimming of silicon photonic ring resonators at wafer-scale. *JLT*, 39(15):5083–5088, 2021.

[143] Wouter J. Westerveld et al. Extension of marcatili's analytical approach for rectangular silicon optical waveguides. *JLT*, 30(14):2388–2401, 2012.

Appendix A

Cross-Over Coupling in Unconventional MRRs

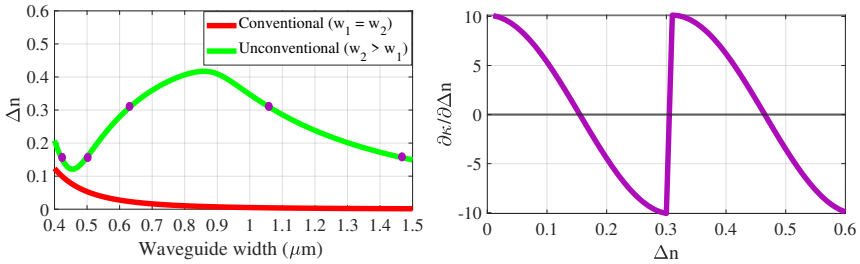
In Section 3.4, we show that the cross-over coupling (κ) in an MRR with wider waveguides and $w_i \neq w_r$ can be improved when $w_r \approx \rho w_i$, where ρ is an integer (see Figs. 3.4 and 3.6). Here, we analytically investigate such an improvement in κ by studying the cross-over coupling in a directional coupler (DC), which can represent the coupling region in an MRR. Moreover, we employ supermode theory [1] to analyze the cross-over coupling in the DC. Supermode analysis studies waveguides by considering the interfaces of the modes of the total structure. According to the supermode analysis, the effective indices of the first two eigenmodes of the coupled waveguides, which are known as symmetric and antisymmetric modes, determine the cross-over coupling in a DC. Considering supermode analysis and a DC with a gap of g and waveguide widths of w_1 and w_2, the cross-over coupling can be defined as:

$$\kappa = \left| sin\left(\frac{\pi \cdot \Delta n}{\lambda} \cdot L\right)\right|, \tag{A.0.1}$$

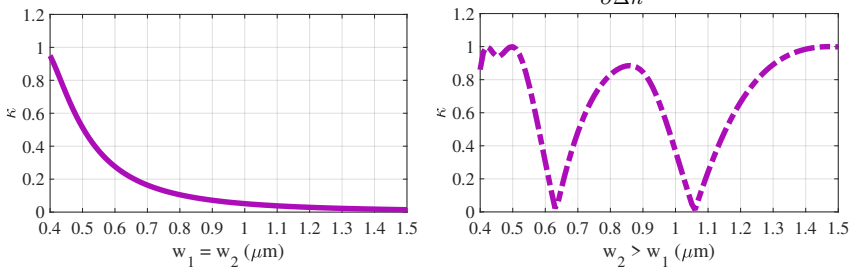
where Δn is the difference between the effective indices of the symmetric (n_e) and antisymmetric (n_o) supermodes (i.e., $n_e - n_o$), λ is the wavelength, and L is the coupler length. Variations in the DC (e.g., increasing/decreasing the waveguide width) change Δn in (A.0.1). Considering the first derivative of the cross-over coupling (κ) with respect to Δn, we have:

$$\frac{\partial \kappa}{\partial \Delta n} = \left(\frac{\pi L}{2\lambda}\right) \frac{sin\left(\frac{2\pi L}{\lambda} \cdot \Delta n\right)}{\left|sin\left(\frac{\pi L}{\lambda} \cdot \Delta n\right)\right|}. \tag{A.0.2}$$

As a case study, we quantitatively simulate κ and $\frac{\partial \kappa}{\partial \Delta n}$ in a DC with $L = 5$ μm and $g = 100$ nm, both considered as an example. We assume two scenarios: 1) a conventional DC in which $w_1 = w_2$, and 2) an unconventional DC in which $w_2 > w_1$. Using Lumerical MODE [65], Fig. A.1(a) shows



(a) $\Delta n = n_e - n_o$ **(b)** $\frac{\partial \kappa}{\partial \Delta n}$ in (A.0.2)



(c) κ in conventional DC **(d)** κ in unconventional DC

Figure A.1: (a) Difference between the effective indices of the symmetric (n_e) and antisymmetric (n_o) supermodes in a conventional and unconventional DC. (b) Rate of changes in the DC cross-over coupling (κ) w.r.t. the changes in Δn. Cross-over coupling in (c) a conventional and (d) an unconventional DC (x-axis shows w_2). In these simulations, $L = 5$ μm and $g = 100$ nm. Also, $w_1 = 400$ nm in the unconventional DC.

Δn where $w_1 = w_2$ increases from 400 to 1500 nm in the conventional DC, and w_2 increases over the same range in the unconventional DC with $w_1 = 400$ nm (considered as an example). Note that the gap is maintained at 100 nm. While increasing the waveguide width, Δn decreases when $w_1 = w_2$. However, when $w_2 > w_1$, Δn decreases at first and then it peaks when $w_2 \approx 2w_1$ in the unconventional DC in Fig. A.1(a).

Employing (A.0.2), Fig. A.1(b) shows $\frac{\partial \kappa}{\partial \Delta n}$ for Δn within the range [0, 0.6]—see the y-axis in Fig. A.1(a)—and with $L = 5$ μm. When $\frac{\partial \kappa}{\partial \Delta n} = 0$, we have a local maximum or minimum in κ. Figs. A.1(c) and A.1(d) show the cross-over coupling in the conventional and unconventional DC, respectively. As can be seen, when the waveguide width increases, κ decreases in the conventional DC, but it increases at some specific waveguide widths in the unconventional DC. In Fig. A.1(a) and for the unconventional DC (when $w_2 > w_1$), we specified (see the circles in the figure) the waveguide widths corresponding to the Δn values at which $\frac{\partial \kappa}{\partial \Delta n} = 0$ in Fig. A.1(b). Considering these waveguide widths (w_2), we can observe multiple maxima and minima in Fig. A.1(d). In

particular, $\frac{\partial^2 \kappa}{\partial \Delta n^2} = 0$ at $\Delta n = 0.3$ in Fig. A.1(b) (i.e., the inflection point), corresponding to $w_2 = 630$ and 1050 nm (see Fig. A.1(a) when $w_2 > w_1$) within which $w_2 \approx 2w_1$: there is a second maximum in κ in Fig. A.1(d) when 630 nm$\leq w_2 \leq$ 1050 nm. Note that one can change L to show a similar trend and extend this to when $w_2 \approx \rho w_1$, where ρ is an integer. This explains the trend observed in the cross-over coupling in MRRs with $w_r > w_i$ (see Fig. 3.6). As discussed in Section 3.5, such an increase in κ helps design MRRs which are not only tolerant to FPVs but also can achieve high Q-factor and 3dB bandwidth.

Appendix B

Impact of Radius Variations on Resonant Wavelength Shift

Assuming that variations in SOI thickness, waveguide width, and slab are independent. Changes in MRR radius can impact both, MRR radius and waveguide width. In order to accurately determine changes in resonant wavelength shift ($\Delta\lambda_{MRR}$) caused due to changes in radius, we divide radius into two different parts, inner (R_i) and outer (R_o) radii as shown in Fig. B.1(a). The relationship between R_i, R_o, and radius (R) of MRR can be defined as distance from center of MRR to $(R_i + R_o)/2$ and we individually study the impact of factors R_i and R_o on $\Delta\lambda_{MRR}$.

Variations in radius can affect individually on R_i and R_o causing changes in MRR radius and width which can be seen in Fig. B.1(c) where changes in either R_i or R_o cause changes in MRR radius as well as width. The amount of width variations the MRR has undergone in such a scenario can be defined by the factor α which accounts for the percent changes in MRR width. The following equation defines $\Delta\lambda_{MRR}$ due to changes in MRR radius and width which can be modeled as:

$$\Delta\lambda_{Radius(\alpha)} = \frac{\lambda_{MR}}{n_{\text{g}}}(\alpha(\delta n_{\text{eff}}) + n_{\text{eff}} \cdot \frac{\delta R}{R}), \tag{B.0.1}$$

Where, $\Delta\lambda_{Radius(\alpha)}$ is changes in resonant wavelength due to change in radius. Fig. B.1(c) clearly shows changes in either outer (R_o') or inner (R_i') radius, due to which both the radius and width of the MRR have been impacted.

Further in our studies we model $\Delta\lambda_{MRR}$ due to change in MRR radius. In such a situation changes in both R_i and R_o lead to changes in MRR radius. The width of MRR remains the same as seen in Fig. B.1(b). $(R_i + R_o)/2$ defines the width of MRR, and since there is no change in MRR width, both R_i and R_o remain the same in other words, $\alpha = 0$. One can observe changes in MRR radius while there are no changes due to MRR width. Such a model can help us analyze the individual effect of changes in resonant wavelength due to changes in individual factors such

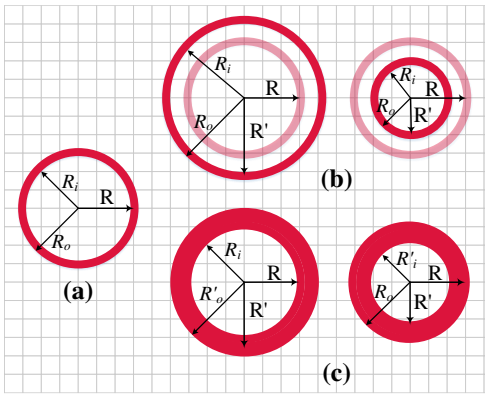


Figure B.1: (a) MRR showing inner and outer radius R_i and R_o respectively. The average radius (R) without any radius variations, (b) variations in MRR radius (R') due to changes in MRR radius, and (c) variations in MRR radius (R')

as radius which in turn helps us accurately model $\Delta\lambda_{MRR}$ and factors related to $\Delta\lambda_{MRR}$ such as total resonant wavelength shift.

Appendix C

Resonant-Wavelength Shift in MRRs

In Section 3.3, we present a model to capture the impact of different variations in the critical dimensions of an MRR on its resonant wavelength (i.e., the resonant-wavelength shift model in (5.3)). Here, we show how this model can be derived. The resonant wavelength (λ_R) in an MRR with a radius R can be defined as:

$$\lambda_R \cdot m = 2\pi R \cdot n_{\text{eff}}, \tag{C.0.1}$$

where n_{eff} is the effective index of the MRR and m is an integer denoting the resonance order. As the n_{eff} changes, the resonant wavelength shifts, which in turn will change the n_{eff} because of the non-zero slope of $\frac{\partial n_{\text{eff}}}{\lambda_R}$. Applying differentiation to both sides of (C.0.1) and observing the same resonance mode (i.e., $\delta m = 0$), we have:

$$m \cdot \delta\lambda_R = 2\pi \left(\delta R \cdot n_{\text{eff}} + R \cdot \delta n_{\text{eff}}\right). \tag{C.0.2}$$

Rearranging (C.0.2) and using $\lambda_R = 2\pi R n_{\text{eff}}$ (A.1), we have:

$$\frac{\delta\lambda_R}{\lambda_R} = \frac{\delta R}{R} + \frac{\delta n_{\text{eff}}}{n_{\text{eff}}}. \tag{C.0.3}$$

Note that n_{eff} in an MRR depends on the MRR's resonant wavelength (λ_R), waveguide width (w), SOI thickness (t), slab thickness (h, for active MRRs), refractive index of the waveguide core (n_c), and refractive index of cladding/substrate (n_s). Therefore, δn_{eff} in (C.0.3) can be defined as:

$$\delta n_{\text{eff}} = \delta\lambda_R \cdot \frac{\partial n_{\text{eff}}}{\partial \lambda_R} + \delta w \cdot \frac{\partial n_{\text{eff}}}{\partial w} + \delta t \cdot \frac{\partial n_{\text{eff}}}{\partial t} \tag{C.0.4}$$

$$+ \delta h \cdot \frac{\partial n_{\text{eff}}}{\partial h} + \delta n_c \cdot \frac{\partial n_{\text{eff}}}{\partial n_c} + \delta n_s \cdot \frac{\partial n_{\text{eff}}}{\partial n_s}. \tag{C.0.5}$$

Considering material dispersion, n_c and n_s also depend on λ_R, hence δn_c and δn_s in (C.0.5) are:

$$\delta n_c = \delta\lambda_R \cdot \frac{\partial n_c}{\partial \lambda_R}, \tag{C.0.5}$$

$$\delta n_s = \delta\lambda_R \cdot \frac{\partial n_s}{\partial \lambda_R}. \tag{C.0.6}$$

Assuming SOI waveguides, n_c and n_s are the refractive indices of silicon and silicon dioxide (SiO_2), respectively. The wavelength dependency of SiO_2 within the wavelength range 1500 to 1600 nm is negligible, hence $\frac{\partial n_s}{\partial \lambda_R} \approx 0$ in (C.0.6). Leveraging group index (n_{g}) definition in (3.2), $\frac{\partial n_{\mathrm{eff}}}{\partial \lambda_R}$ in (C.0.5) and $\frac{\partial n_c}{\partial \lambda_R}$ in (C.0.5) can be defined as:

$$\frac{\partial n_{\mathrm{eff}}}{\partial \lambda_R} = \frac{n_{\mathrm{eff}} - n_{\mathrm{g}}}{\lambda_R}, \tag{C.0.7}$$

$$\frac{\partial n_c}{\partial \lambda_R} = \frac{n_c - n_{\mathrm{g}}^c}{\lambda_R}, \tag{C.0.8}$$

where n_{g}^c is the group index of silicon (waveguide core). Applying (C.0.7) and (C.0.8) to (C.0.5), we have:

$$\delta n_{\mathrm{eff}} = \delta\lambda_R \left(\frac{n_{\mathrm{eff}} - n_{\mathrm{g}}}{\lambda_R} \right) + \delta w \cdot \frac{\partial n_{\mathrm{eff}}}{\partial w} + \delta t \cdot \frac{\partial n_{\mathrm{eff}}}{\partial t} \tag{C.0.9}$$

$$+ \delta h \cdot \frac{\partial n_{\mathrm{eff}}}{\partial h} + \delta\lambda_R \cdot \frac{\partial n_{\mathrm{eff}}}{\partial n_c} \left(\frac{n_c - n_{\mathrm{g}}^c}{\lambda_R} \right). \tag{C.0.10}$$

Applying (C.0.10) to (C.0.3) and after simplification, we have:

$$\frac{\delta\lambda_R}{\lambda_R} = \frac{n_{\mathrm{eff}} \cdot \frac{\delta R}{R} + \delta w \cdot \frac{\partial n_{\mathrm{eff}}}{\partial w} + \cdot \delta t \cdot \frac{\partial n_{\mathrm{eff}}}{\partial t} + \delta h \cdot \frac{\partial n_{\mathrm{eff}}}{\partial h}}{n_{\mathrm{g}} - \left(n_{\mathrm{g}}^c - n_c\right) \cdot \frac{\partial n_{\mathrm{eff}}}{\partial n_c}}, \tag{C.0.11}$$

where in the denominator $n_{\mathrm{g}} >> \left(n_{\mathrm{g}}^c - n_c\right) \cdot \frac{\partial n_{\mathrm{eff}}}{\partial n_c}$, and hence $\left(n_{\mathrm{g}}^c - n_c\right) \cdot \frac{\partial n_{\mathrm{eff}}}{\partial n_c}$ can be ignored. Leveraging (C.0.11), the resonant-wavelength shift model in (3.4) can be derived.