UNIVERSITY OF CALIFORNIA
Santa Barbara

Modeling, Simulation, and Optimization of Variation-Aware Runtime-Reconfigurable Optical Interconnects

A dissertation submitted in partial satisfaction of the
requirements for the degree of

Doctor of Philosophy
in
Electrical and Computer Engineering

by

Yuyang Wang

Committee in charge:

Professor Kwang-Ting Cheng, Co-Chair
Professor John E. Bowers, Co-Chair
Professor Yuan Xie
Professor Clint L. Schow
Professor Zheng Zhang

September 2021

The Dissertation of Yuyang Wang is approved.

Professor Yuan Xie

Professor Clint L. Schow

Professor Zheng Zhang

Professor John E. Bowers, Committee Co-Chair

Professor Kwang-Ting Cheng, Committee Co-Chair

July 2021

Modeling, Simulation, and Optimization of Variation-Aware Runtime-Reconfigurable Optical Interconnects

by

Yuyang Wang

To my wife, Crystal

Acknowledgements

My sincerest thanks go to my advisor, Professor Kwang-Ting Cheng, for being my role model of academic professionalism and a true mentor over the past six years. I am eternally grateful for his intellectual and personal advice and his support in all forms, all of which have encouraged me to become an independent researcher, a dependable collaborator, and above all, a better person.

I am deeply grateful to the rest of my dissertation committee, Professor John E. Bowers, Professor Yuan Xie, Professor Clint L. Schow, and Professor Zheng Zhang, for their insightful guidance and valuable feedback on my research, as well as their generous help in facilitating collaboration and internship opportunities over the years. Besides, I sincerely thank Professor Nadir Dagli for his vital input to my research, and Professor Larry Coldren, Professor Margaret Marek-Sadowska, Professor Li-C. Wang, Professor Forrest Brewer, Professor Behrooz Parhami, Professor Jason Marden, and Dr. Robby Nadler for their informative lectures.

I gratefully acknowledge the financial support from Semiconductor Research Corporation (SRC) and American Institute for Manufacturing Integrated Photonics (AIM Photonics) during my Ph.D. study.

I extend my heartfelt gratitude to my industry collaborators at Hewlett Packard Labs, Dr. M. Ashkan Seyedi, Dr. Peng Sun, Dr. Jared Hulme, Dr. Marco Fiorentino, Dr. Raymond G. Beausoleil, and Mudit Jain, for sharing firsthand data of fabricated devices and providing thoughtful input to our collaborative papers. I also thank Dr. Tsung-Ching Huang for offering his professional advice on various occasions.

I greatly appreciate the invaluable help from Professor Jiang Xu at HKUST, who offered me opportunities to promote my research in various workshops and departmental seminars and shared his extensive experience and broad vision of the field. I am also grateful for

the mentorship I received from Dr. Gilles Lamant and Dr. Ahmadreza Farsaei during my internship at Cadence Design Systems, Inc.

I sincerely thank my fellow collaborators at UCSB, including Dr. Rui Wu, Dr. Zeyu Zhang, and Dr. Chong Zhang, for stimulating fruitful discussions and contributing to several of our research projects. I also thank Dr. Fan Lan, Dr. Di Liang, and Dr. Ping-Lin Yang for patiently answering my technical questions and offering helpful suggestions during my research.

I gratefully acknowledge the dedicated service of the staff of the ECE Department, including Valerie De Veyra, Shannon Gann, Robin Jenneve, Adriana Aguirre, Olivia La Pierre, Tori Smith, and Paul Gritt, for providing all sorts of help and relieving me from the burden of paperwork. I also thank the staff of the OISS and the Housing Services of UCSB for their indispensable support.

I extend my special thanks to my labmates unmentioned, Dr. Miguel A. Lastras Montaño, Dr. Amirali Ghofrani, Dr. Nicole Fern, Dr. Chun-Kai Hsu, Dr. Fan Lin, Dr. Chong Huang, and Dr. Leilai Shao, for our time together in the lab and our discussions on various topics. I also appreciate the friendship offered by fellow students from other research groups at UCSB, including Dr. Peng Gu, Dr. Maohua Zhu, and Fengqiao Sang. Besides, I thank my friends at HKUST, including Dr. Xuanqi Chen, Dr. Zengqiang Yan, Shichao Li, Guang Chen, Jeffry Wicaksana, and Weihang Dai, for their hospitality during my visit.

Finally, I would like to express my eternal gratefulness to my family: my parents, Dr. Zhihua Wang and Xuefei Yang, for their selfless devotion to my growth and education over the past 28 years; my late grandma, for her continued love and encouragement; my sister, Jacqueline, for simply being there for me; and my wife and life partner, Crystal, for her unconditional trust, support, and companionship, without which I would not have been able to make this journey.

Curriculum Vitæ

Yuyang Wang

Education

2021	Ph.D., Electrical and Computer Engineering (Expected), University of California, Santa Barbara, California, USA.
2018	M.S., Electrical and Computer Engineering, University of California, Santa Barbara, California, USA.
2015	B.Eng., Electronic Engineering, Tsinghua University, Beijing, China.

Professional Experience

Fall 2019	*Visiting Intern*, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong SAR, China.
Winter 2019	*Teaching Assistant*, Department of Electrical and Computer Engineering, University of California, Santa Barbara, California, USA.
Summer 2018	*Design Engineering Intern*, Cadence Design Systems, Inc., San Jose, California, USA.
Summer 2014	*Student Intern*, Rice University, Houston, Texas, USA.

Publications

Yuyang Wang and Kwang-Ting Cheng, "Traffic-Adaptive Power Reconfiguration for Energy-Efficient and Energy-Proportional Optical Interconnects," in *IEEE/ACM International Conference on Computer-Aided Design (ICCAD),* Nov. 2021.

Yuyang Wang, Peng Sun, Jared Hulme, M. Ashkan Seyedi, Marco Fiorentino, Raymond G. Beausoleil, and Kwang-Ting Cheng, "Energy Efficiency and Yield Optimization for Optical Interconnects via Transceiver Grouping," *IEEE/OSA Journal of Lightwave Technology,* vol. 39, no. 6, 2021.

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 *ACM/IEEE Design Automation Conference (DAC),* Jun. 2020.

Yuyang Wang and Kwang-Ting Cheng, "Task Mapping-Assisted Laser Power Scaling for Optical Network-on-Chips," in *IEEE/ACM International Conference on Computer-Aided Design (ICCAD),* Nov. 2019.

Yuyang Wang, Leilai Shao, Miguel A. Lastras-Montaño, and Kwang-Ting Cheng, "Taming Emerging Devices' Variation and Reliability Challenges with Architectural and System Solutions [Invited]," in *IEEE International Conference on Microelectronic Test Structures (ICMTS),* Mar. 2019.

Yuyang Wang, M. Ashkan Seyedi, Jared Hulme, Marco Fiorentino, Raymond G. Beausoleil, and Kwang-Ting Cheng, "Bidirectional Tuning of Microring-Based Silicon Photonic Transceivers for Optimal Energy Efficiency," in *ACM/IEEE Asia and South Pacific Design Automation Conference (ASP-DAC),* Jan. 2019.

Yuyang Wang, M. Ashkan Seyedi, Rui Wu, Jared Hulme, Marco Fiorentino, Raymond G. Beausoleil, and Kwang-Ting Cheng, "Energy-Efficient Channel Alignment of DWDM Silicon Photonic Transceivers," in *IEEE Design, Automation & Test in Europe Conference & Exhibition (DATE),* Mar. 2018.

Rui Wu, M. Ashkan Seyedi, **Yuyang Wang**, Jared Hulme, Marco Fiorentino, Raymond G. Beausoleil, and Kwang-Ting Cheng, "Pairing of Microring-Based Silicon Photonic Transceivers for Tuning Power Optimization," in *IEEE/ACM Asia and South Pacific Design Automation Conference (ASP-DAC),* Jan. 2018.

Zeyu Zhang, Rui Wu, **Yuyang Wang**, Chong Zhang, Eric J. Stanton, Clint L. Schow, Kwang-Ting Cheng, and John E. Bowers, "Compact Modeling for Silicon Photonic Heterogeneously Integrated Circuits," *IEEE/OSA Journal of Lightwave Technology,* vol. 35, no. 14, 2017.

Rui Wu, **Yuyang Wang**, Zeyu Zhang, Chong Zhang, Clint L. Schow, John E. Bowers, and Kwang-Ting Cheng, "Compact Modeling and Circuit-Level Simulation of Silicon Nanophotonic Interconnects," in *IEEE Design, Automation & Test in Europe Conference & Exhibition (DATE),* Mar. 2017.

Amirali Ghofrani, Miguel A. Lastras-Montaño, **Yuyang Wang**, and Kwang-Ting Cheng, "Inplace Repair for Resistive Memories Utilizing Complementary Resistive Switches," in *ACM International Symposium on Low Power Electronics and Design (ISLPED),* Aug. 2016.

Chao Xu, Felix X. Lin, **Yuyang Wang**, and Lin Zhong, "Automated OS-level Device Runtime Power Management," in *ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS),* Mar. 2015.

Abstract

Modeling, Simulation, and Optimization of Variation-Aware Runtime-Reconfigurable Optical Interconnects

by

Yuyang Wang

The explosive growth of data volume brought by the pervasiveness of artificial intelligence (AI) applications is calling for interconnect technologies that enable higher bandwidth capacity at a lower cost. In particular, optical interconnects based on silicon photonics are considered a promising substitute for electrical ones, given their cost-effectiveness and scalability enabled by a CMOS-compatible fabrication process. However, as the optical interconnects further penetrate the shorter-reach regime, several issues arise from the growing complexity of system integration and pose challenges to their design quality, including 1) inadequate support from design automation methodologies for the modeling and simulation of the optical interconnects, 2) oversimplified characterization of process variations, resulting in variation alleviation techniques with limited effectiveness, and 3) the lack of runtime reconfiguration strategies for the optical interconnects under traffic dynamics, leading to unoptimized energy efficiency. This dissertation is devoted to addressing the above issues by solutions proposed at the device, link, and system levels, paving the way to the quality design of variation-aware runtime-reconfigurable optical interconnects with optimized energy efficiency.

The first part of this dissertation focuses on device-level methodologies for electronic-photonic design automation (EPDA), including compact models developed for lasers and modulators and a novel hierarchical spatial variation model characterized for silicon microring resonators. Extensively validated by measurement data, the library of device-level

models enables accurate circuit-level simulation of optical links and variation-aware estimation of the link power budget, serving as the fundamentals of the optimization techniques proposed at the link and system levels.

The second part of this dissertation proposes three link-level techniques to improve the energy efficiency of the optical interconnects under wafer-scale process variations. The three techniques exploit, respectively, 1) sub-channel redundancy of carrier wavelengths, 2) a combination of electrical and thermal tuning mechanisms, and 3) optimal mixing and matching of a batch of fabricated transceivers, achieving significant reductions in the energy required for transmitting a single bit of data.

The third and final part of this dissertation proposes two strategies at the system level that further improve the energy efficiency of the optical interconnects under traffic dynamics by reconfiguring the link power at application runtime. The two strategies incorporate assistance from 1) traffic adjustment enabled by task mapping exploration and 2) traffic adaptability enabled by traffic prediction, respectively, and achieve substantial improvements in energy consumption with minimal overhead to application execution time, notably outperforming existing strategies.

Contents

List of Figures

List of Tables

List of Abbreviations

A

AC alternating current.
AI artificial intelligence.
AIM Photonics American Institute for Manufacturing Integrated Photonics.
AWG arrayed waveguide grating.

B

BER bit error rate.
BPM beam-propagation method.

C

CMOS complementary metal-oxide-semiconductor.
CMT coupled-mode theory.
CW continuous-wave.

D

DAG directed acyclic graph.
DBR distributed Bragg reflector.
DC direct current.
DCN data center network.
DFB distributed feedback.
DGTD discontinuous Galerkin time-domain.
DLPS Dynamic Laser Power Scaling.
DML directly-modulated laser.
DRC design rule checking.
DSE design space exploration.
DVFS dynamic voltage and frequency scaling.
DWDM dense wavelength-division multiplexing.

E

EAM electroabsorption modulator.
EDA electronic design automation.
EDP energy-delay product.
EEE effective energy efficiency.
EME eigenmode expansion.
EO electro-optical.
EPDA electronic-photonic design automation.
ER extinction ratio.

F

FDTD finite-difference time-domain.
FEM finite-element method.
FFT fast Fourier transform.
FSR free spectrum range.
FWHM full width at half maximum.

G

GA genetic algorithm.

H

HDL hardware description language.
HPC high-performance computing.

I

ISA instruction set architecture.
IV current-voltage.

L

LCA lightwave component analyzer.
LI light-current.
LVS layout versus schematic.

M

MPD monitoring photodetector.
MPSoC multi-processor system-on-chip.
MRR microring resonator.
MWMR multiple-reader-multiple-writer.
MZM Mach-Zehnder modulator.

N

NEE nominal energy efficiency.
NoC network-on-chip.
NP nondeterministic polynomial-time.

O

OMA optical modulation amplitude.
ONoC optical network-on-chip.
OOK on-off keying.

P

p.p. percentage point.
P2P point-to-point.
PARAFAC parallel factor.
PCE polynomial chaos expansion.
PCell parameterized cell.
PD photodetector.
PDA photonic design automation.
PDK process design kit.
PIC photonic integrated circuit.
PRBS pseudorandom binary sequence.
PSA Pareto simulated annealing.
PSD power spectral density.

Q

Q quality factor.
QCSE quantum-confined stark effect.
QD quantum-dot.
QQ-plot quantile-quantile plot.

R

RCWA rigorous coupled-wave analysis.
RIN relative intensity noise.
Rx receiver.

S

SA simulated annealing.
SDL schematic-driven layout.
SerDes serializer/deserializer.
SIE surface integral equation.
SNR signal-to-noise ratio.
SOA semiconductor optical amplifier.
SOI silicon-on-insulator.
SPICE Simulation Program with Integrated Circuit Emphasis.
SSM split-step method.

T

TDTW time-domain traveling-wave.
TIA transimpedance amplifier.
TO thermal-optical.
TRx transceiver.
Tx transmitter.

V

VCSEL vertical-cavity surface-emitting laser.
VIE volume integral equation.
VP Virtual Probe.

W

WDM wavelength-division multiplexing.
WPE wall-plug efficiency.

Part I

Introduction and Literature Review

Chapter 1

Introduction

Optical interconnects enabled by silicon photonics are promising to solve the communication bottleneck in future data center networks (DCNs) and high-performance computing (HPC) systems, a problem with growing prominence due to the rapid expansion of data volume in the era of artificial intelligence (AI). However, the practical application and broad adoption of optical interconnects in short-reach datacom solutions rely on methodologies and techniques for addressing the issues that arise from the increasing complexity of system integration, notably including 1) inadequate support from photonic design automation methodologies, 2) limited effectiveness of techniques for alleviating process variations, and 3) deteriorated energy efficiency under runtime traffic dynamics. This dissertation aims to tackle these issues by solutions proposed at the device, link, and system levels, paving the way to the quality design of variation-aware runtime-reconfigurable optical interconnects with optimized energy efficiency.

By briefly introducing the background of optical interconnects and identifying the challenges posed to their design automation and optimization, this chapter describes the motivation and the scope of this dissertation.

1.1 Background of Optical Interconnects

1.1.1 Evolution Toward Shorter Reach

Optical interconnects offer potential benefits of higher bandwidth capacity, lower propagation delay, and greater tolerance of electromagnetic interference compared to electrical ones [1]. As the optical communication technologies continuously evolve over the past few decades to become more energy-efficient, the substitution of traditional electrical interconnects by optical ones gradually unrolls from the long-reach telecom regime to the short-reach datacom regime [2]. As of today, optical interconnects have dominated the data communication solutions above the rack-to-rack level in data centers and HPC systems, as illustrated in Fig. 1.1.

In recent years, the explosive growth of data-intensive AI applications has stimulated advances in computational capability through hardware parallelism and specialization, which further shifts the performance bottleneck of parallel computing infrastructures from computation to communication [3]. For avoiding data-starved computation nodes, the peak

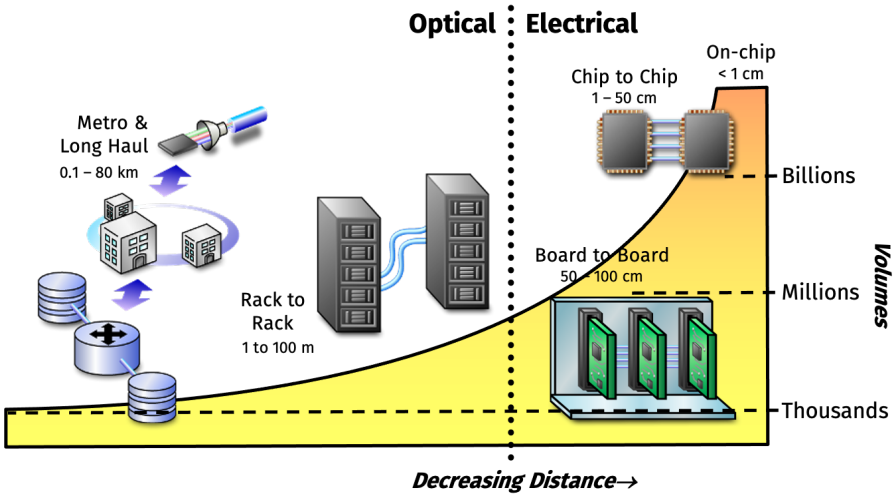


Figure 1.1: Evolution of optical interconnects toward shorter reach (courtesy of Intel Corporation).

bandwidth requirement provisioned for DCNs and HPC interconnects has exceeded hundreds of Gb/s, a data rate at which traditional electrical interconnects become uneconomical even for short distances within a rack [4]. On the other hand, short-reach optical links leveraging dense wavelength-division multiplexing (DWDM) have demonstrated an aggregated data rate of 400 Gb/s [5], and technologies toward the Tb/s class are under active investigation [6, 7]. In the foreseeable future, optical interconnects are expected to continue their penetration into the intra-rack regime to provide high-throughput and cost-effective communication solutions at the board-to-board, chip-to-chip, and ultimately core-to-core levels [8].

The evolution of optical interconnects toward shorter reach is inevitably accompanied by the need for addressing greater integration complexity. According to data from [9], the number of devices integrated on a photonic integrated circuit (PIC) grows exponentially over the years, as summarized in Fig. 1.2, often referred to as the Moore's Law for photonics [10]. To this end, silicon photonics has drawn particular attention, among various enabling technologies for optical interconnects, for its cost-effectiveness and scalability achieved by a CMOS-compatible fabrication process [11, 12, 13]. The plasma dispersion effect of the silicon material enables the implementation of electro-optical (EO) modulators in silicon without the need to modify the CMOS process [14]. Although it is hard to implement light emitting or amplifying devices in silicon due to its indirect bandgap, several solutions have emerged using wafer-bonded or directly grown III-V materials on silicon [15, 16].

1.1.2 Overview of Optical Devices and Interconnect Architectures

An optical network is a collection of optical links that provide data communication between various processing nodes. The basic building block of an optical network is the optical transceiver (TRx), which, in turn, comprises several essential components, including

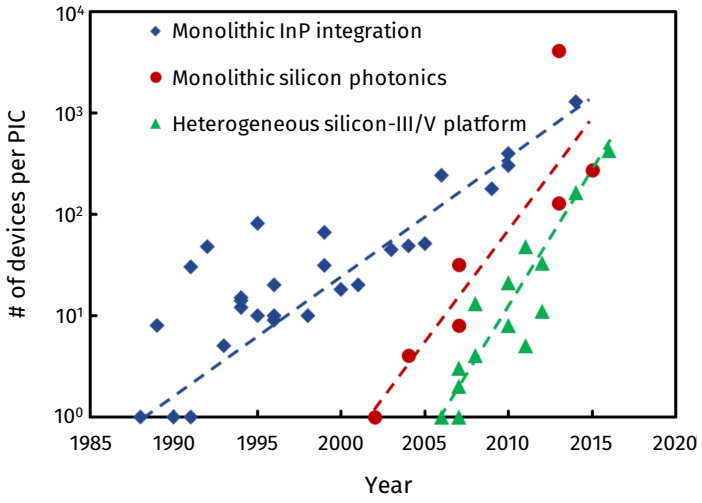


Figure 1.2: Number of photonic devices integrated on a PIC over the past decades for three technology platforms of integrated photonics (data from [9]).

lasers, modulators, photodetectors (PDs), and the medium for light propagation, namely waveguides and optical fibers. The adoption of wavelength-division multiplexing (WDM) has become a standard practice for better bandwidth capacity, which transmit multiple wavelengths in the same medium concurrently [17].

From the perspective of laser integration, optical interconnects can be categorized by using either off-chip or on-chip lasers, the latter gaining in popularity for less coupling loss, better energy proportionality, and greater layout flexibility [18]. In terms of the multiplexing architecture, optical interconnects can be distinguished by using either an array of single-wavelength lasers multiplexed by an external multiplexer [19] or a quantum-dot (QD) comb laser that generates a group of evenly-spaced comb lines from a single device output [20,21]. As for the modulation scheme, optical interconnects may opt for either directly-modulated lasers (DMLs) or continuous-wave (CW) lasers modulated by external modulators [22]. Some of the popular modulators implemented in silicon photonics include the electroabsorption modulator (EAM) [23], the Mach-Zehnder modulator (MZM) [24], and the silicon microring modulator [25].

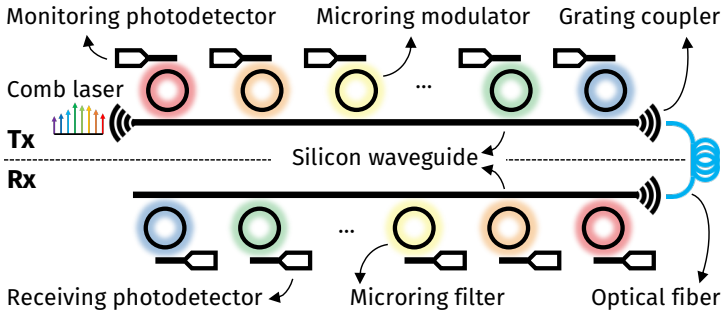


Figure 1.3: Architectural illustration of a silicon microring–based optical transceiver.

In recent years, optical transceivers based on QD comb lasers and silicon microring resonators have drawn increasing attention for its (de)multiplexer-free architecture that achieves DWDM within compact footprints [26,27,28,29]. Fig. 1.3 illustrates the architecture of a microring-based optical transceiver, showing cascaded microrings deployed alongside a shared waveguide. At the transmitter (Tx) side, each microring modulator modulates a specific wavelength at its resonance. At the receiver (Rx) side, a corresponding microring filter couples the signal out for detection. Optical interconnect architectures based on other types of lasers and modulators have also been reported [30].

1.2 Scope of Challenges to be Addressed

Despite the legacy of the well-developed CMOS process that is partially reusable for reducing the production cost of silicon photonics, new challenges still emerge with the increasing number of devices that require close integration, as well as the growing complexity of traffic patterns in modern computing infrastructures, manifesting themselves in the design difficulty and the energy efficiency issues of the optical interconnects.

1.2.1 Compact Modeling and Simulation

The first and fundamental challenge is the insufficient support for the design automation of PICs, especially the compact modeling and simulation methodologies that are indispensable for the co-design of photonic devices and their electronic driving circuitry. According to the 2020 Integrated Photonic Systems Roadmap on electronic-photonic design automation (EPDA) [31], the modeling and simulation of photonic devices and circuits are at the foundation of the EPDA framework, among other essential methodologies like layout and verification of curvilinear structures, collaboratively envisioned by task forces from both academia and industry, as illustrated in Fig. 1.4. Therefore, it is desirable for the modeling and simulation methodologies for optical interconnects to build upon the successful experience of the electronic design automation (EDA) community by incorporating support for photonic features, such as optical power and phase, into existing EDA toolsets. In addition, the development of link- and system-level optimization techniques for optical interconnects also relies on credible models and simulation methodologies validated by actual measurement data of fabricated devices.

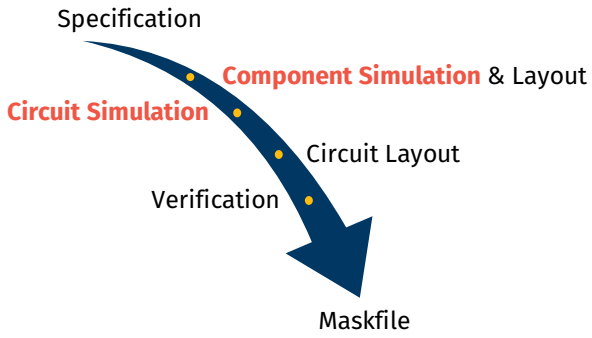


Figure 1.4: Design automation methodologies required for designing a PIC, as envisioned by [31], where the modeling and simulation methodologies for photonic devices and circuits are at the foundation.

1.2.2 Variation Characterization and Alleviation

Another major challenge posed to optical interconnects stems from their particular vulnerability to process variations. Due to the relatively large ratio between the device size and the operating wavelength, photonic devices are inherently more sensitive to geometry uncertainties than electronic ones, even leveraging the same fabrication process. For wavelength-critical devices, such as microring resonators (MRRs) [25], post-fabrication tuning mechanisms are mandatory for rectifying the deviation of the resonance wavelengths of the MRRs caused by the process variations. In a microring-based optical transceiver as illustrated in Fig. 1.3, each microring is typically fabricated with a resistive thermal tuner that can be electrically heated up to shift its resonance wavelength. As the power consumption for variation rectification can take up over half of that of a point-to-point (P2P) link (as shown in Fig. 1.5) [32], the energy efficiency of optical interconnects strongly depends on the effectiveness of link-level variation alleviation techniques, which, in turn, calls for accurate characterization of the variation patterns of fabricated devices.

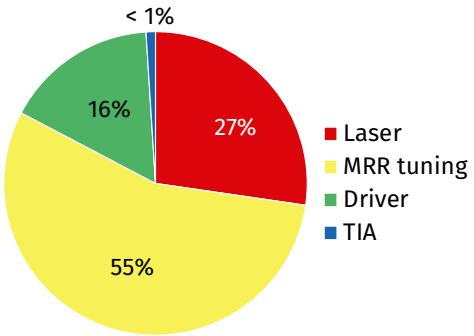


Figure 1.5: Energy decomposition of a microring-based optical link reported in [32] showing the shares of the laser, MRR tuning (for variation rectification), electronic drivers, and transimpedance amplifiers (TIAs).

1.2.3 Runtime Power Reconfiguration

A third critical challenge arises for optical interconnects in terms of deteriorated energy efficiency when operating in real-world systems under runtime traffic dynamics. Despite recent advances in device technologies [33, 34, 35] that are pushing the best-case energy efficiency of an individual link toward ~1 pJ/b, the effective energy efficiency of an interconnected network is often worse by orders of magnitude due to traffic dynamics [36], if without proper power reconfiguration for photonic devices at application runtime. Moreover, failure to properly manage the link power during idle or low-utilization periods is devastating to the energy proportionality of the network [37]. As the traffic patterns in modern DCNs and HPC systems are both temporally and spatially non-uniform [38,39], much different from the relatively steady patterns found in long-haul communication [40], the optical links in short-reach application scenarios frequently switch between idle and various utilization levels, intensifying the demand for traffic-adaptive power management. As a result, the need for system-level runtime reconfiguration strategies has become unprecedentedly imminent for developing energy-optimized optical interconnects.

1.3 Clarification on Energy Efficiency

As described in Sections 1.2.2 and 1.2.3, both the link-level variation alleviation techniques and the system-level runtime reconfiguration strategies aim to tackle the energy efficiency issues of the optical interconnects, nevertheless, from two different dimensions. This section thus clarifies the terminologies for describing the energy efficiency of optical interconnects.

The energy efficiency of optical interconnects is usually measured in pJ/b. However, in most literature, this metric is computed as mW/ (Gb/s), a unit equivalent to pJ/b as 1 W = 1 J/s, reflecting the power required to attain a target data rate. In this dissertation, we refer

to this power-oriented metric as the nominal energy efficiency (NEE) to distinguish it from the effective energy efficiency (EEE), the latter measuring the actual energy consumption associated with data movement.

> **Definition 1.1** (nominal energy efficiency, NEE)**.** The nominal energy efficiency describes the power consumption of an individual optical link required to communicate at a target data rate while maintaining the bit error rate (BER) below a given threshold (e.g., 10^{-12}).

Thus, reducing the power consumption of various link components, such as the circuitry for variation rectification, is beneficial to the NEE of the optical link.

> **Definition 1.2** (effective energy efficiency, EEE)**.** The effective energy efficiency, on the other hand, measures the actual energy consumed by the optical interconnects, which may comprise multiple links, to transfer a total number of bits during the entire timespan of application execution.

In the presence of traffic dynamics, optical links without proper power management may still consume energy when they are idle, causing the EEE of the optical network to be orders of magnitude worse than the NEE of individual links. Therefore, traffic-adaptive runtime reconfiguration strategies are imperative for improving the effective energy efficiency of the optical interconnects.

1.4 Dissertation Outline

For developing energy-optimized optical interconnects with the increasing complexity of system integration, this dissertation is devoted to resolving the three critical needs of

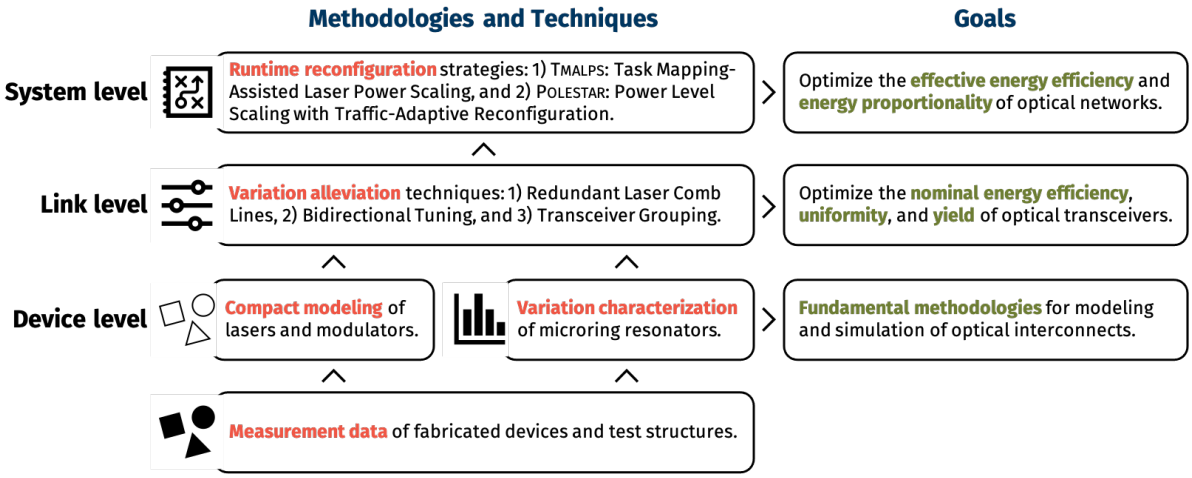


Figure 1.6: Overview of the solutions proposed in this dissertation.

optical interconnects identified in Section 1.2, namely 1) validated methodologies for photonic modeling and simulation, 2) effective techniques for variation characterization and alleviation, and 3) runtime power reconfiguration strategies for traffic adaptability. Fig. 1.6 summarizes the solutions proposed in this dissertation and categorizes them into the device, link, and system levels, detailed in Parts II, III, and IV, respectively.

Part II focuses on device-level EPDA methodologies, including compact models developed for lasers and modulators (Chapter 3) and a novel hierarchical spatial variation model characterized for silicon microring resonators (Chapter 4). Extensively validated by measurement data, the library of developed models enables accurate circuit-level simulation of optical links and variation-aware estimation of the link power budget, serving as the fundamentals of the optimization techniques proposed at the link and system levels.

Part III introduces three link-level techniques for improving the energy efficiency of the optical interconnects under wafer-scale process variations. The three techniques exploit, respectively, 1) sub-channel redundancy of carrier wavelengths (Chapter 5), 2) a combination of electrical and thermal tuning mechanisms (Chapter 6), and 3) optimal mixing and matching of a batch of fabricated transceivers (Chapter 7), achieving significant reductions

in the nominal energy efficiency of optical links.

Part IV further proposes two strategies at the system level that improve the effective energy efficiency of optical interconnects under traffic dynamics by reconfiguring the link power at application runtime. The two strategies incorporate assistance from 1) traffic adjustment enabled by task mapping exploration (Chapter 8) and 2) traffic adaptability enabled by traffic prediction (Chapter 9), respectively, and achieve substantial improvements in energy consumption with minimal overhead to application execution time, notably outperforming existing strategies.

1.5 Permissions and Attributions

This dissertation has the following permissions and attributions:

1. Chapter 3 contains material from "Compact Modeling for Silicon Photonic Heterogeneously Integrated Circuits," by Zeyu Zhang, Rui Wu, Yuyang Wang, Chong Zhang, Eric J. Stanton, Clint L. Schow, Kwang-Ting Cheng, and John. E. Bowers, which appears in *IEEE/OSA Journal of Lightwave Technology,* vol. 35, no. 14, 2017.

2. Chapter 4 contains material from "Characterization and Applications of Spatial Variation Models for Silicon Microring–Based Optical Transceivers," by Yuyang Wang, Jared Hulme, Peng Sun, Mudit Jain, M. Ashkan Seyedi, Marco Fiorentino, Raymond G. Beausoleil, and Kwang-Ting Cheng, which appears in *ACM/IEEE Design Automation Conference (DAC),* Jun. 2020.

3. Chapter 5 contains material from "Energy-Efficient Channel Alignment of DWDM Silicon Photonic Transceivers," by Yuyang Wang, M. Ashkan Seyedi, Rui Wu, Jared Hulme, Marco Fiorentino, Raymond G. Beausoleil, and Kwang-Ting Cheng, which appears in *IEEE Design, Automation & Test in Europe Conference & Exhibition (DATE),* Mar. 2018.

4. Chapter 6 contains material from "Bidirectional Tuning of Microring-Based Silicon Photonic Transceivers for Optimal Energy Efficiency," by Yuyang Wang, M. Ashkan Seyedi, Jared Hulme, Marco Fiorentino, Raymond G. Beausoleil, and Kwang-Ting Cheng, which appears in *ACM/IEEE Asia and South Pacific Design Automation Conference (ASP-DAC),* Jan. 2019.

5. Chapter 7 contains material from "Energy Efficiency and Yield Optimization for Optical Interconnects via Transceiver Grouping," by Yuyang Wang, Peng Sun, Jared Hulme, M. Ashkan Seyedi, Marco Fiorentino, Raymond G. Beausoleil, and Kwang-Ting Cheng, which appears in *IEEE/OSA Journal of Lightwave Technology,* vol. 39, no. 6, 2021.

6. Chapter 8 contains material from "Task Mapping–Assisted Laser Power Scaling for Optical Network-on-Chips," by Yuyang Wang and Kwang-Ting Cheng, which appears in *IEEE/ACM International Conference on Computer-Aided Design (ICCAD),* Nov. 2019.

7. Chapter 9 contains material from "Traffic-Adaptive Power Reconfiguration for Energy-Efficient and Energy-Proportional Optical Interconnects," by Yuyang Wang and Kwang-Ting Cheng, which appears in *IEEE/ACM International Conference on Computer-Aided Design (ICCAD),* Nov. 2021.

Chapter 2

Prior Work

This chapter summarizes some prior studies that address the challenges in the design automation and optimization of optical interconnects by device-, link-, and system-level tools, methodologies, and techniques.

2.1 Device-Level Design Automation and Optimization

The design of photonic devices for photonic integrated circuits (PICs) involves iterations between physical simulation of device specifications and optimization of design parameters. With the ever-growing number of devices to be integrated, both steps require support from design automation methodologies for improved design efficiency. Meanwhile, high-level characterization methodologies, including device compact modeling and process variation characterization, are also needed for efficient and accurate prediction of system performance when many devices are integrated.

2.1.1 Simulation Tools and Methodologies for Device Physics

Simulators for electromagnetic propagation are heavily used by photonic device designers and have been well developed [41, 42] based on methodologies such as the finite-difference time-domain (FDTD) method, finite-element method (FEM), eigenmode expan-

sion (EME) method, and beam-propagation method (BPM). A comprehensive review of these numerical methods is conducted in [43]. Additional methods for numerical simulation of photonic devices have also been extensively investigated, including the time-domain traveling-wave (TDTW) method [44, 45], split-step method (SSM) [46], and coupled-mode theory (CMT) [47], as well as the discontinuous Galerkin time-domain (DGTD) method, volume integral equation (VIE) methods, surface integral equation (SIE) methods, and rigorous coupled-wave analysis (RCWA) as summarized in [48]. A trade-off often explored by the simulation methodologies for device physics is the between efficiency and accuracy. Some novel methodologies, such as [49] and [50], achieve significant improvement in simulation efficiency with comparable accuracy to traditional methods through carefully designed approximations.

2.1.2 Physical-Level Device Optimization

The optimization of photonic devices draws particular research attention due to the need for high-performance devices with compact footprints. The optimization targets of photonic devices often vary with intended applications and require different optimization techniques.

Optimization of Active Devices

The design optimization of active devices requires knowledge of the unique properties of the active region and the electro-optical (EO) effects involved. For lasers, trade-offs are often explored between several optimization targets, including but not limited to a faster modulation speed, a higher wall-plug efficiency (WPE), a smaller threshold current, a narrower linewidth, a lower level of noise, and better temperature stability. Through the usage of a hybrid III-V active layer, the distributed feedback (DFB) laser reported in [51] demon-

strates a direct modulation rate of 25.8 Gb/s and an energy efficiency of 0.5 pJ/b. A hybrid quantum-dot (QD) laser for on-off keying (OOK) modulation has reported a record high direct modulation rate of 15 Gb/s with 1.2 pJ/b at up to 70 °C [52]. In [53], a detailed model is developed based on a novel measurement technique for the gain characteristics of QD laser material with different growth methods. The model has been demonstrated to provide useful guidance on the design of complex laser cavities, such as distributed Bragg reflector (DBR) lasers and mode-locked lasers, resulting in several device designs with record-breaking performances [54].

Physical models for EO modulators are also developed and continuously refined for guiding the optimization of modulator design in terms of energy efficiency, bandwidth, and insertion loss [55, 56, 57]. Common optimization trade-offs explored for the design of optical modulators are between the modulation speed, the driving voltage, and the device footprint. In [58], a modulator with semiconductor hybrid polymer integration suitable for operation at 40 Gb/s is demonstrated with a phase-shifting efficiency smaller than 0.5 V·mm and an energy efficiency of 0.4 pJ/b. Recently, a silicon microring modulator with integrated thermal-optical (TO) resonance tuning has been proposed in [59] with a data rate up to 128 Gb/s and energy efficiency of 18 fJ/b. In addition, similar trade-offs have been explored for the design optimization of other active devices, such as between the speed and the sensitivity requirement of photodetectors (PDs), a state-of-the-art result shown in [60].

Optimization of Passive Devices

The optimization of passive devices is also vital to the improvement of bandwidth and the reduction of crosstalk and losses in optical interconnects. However, the room for optimization is usually limited if starting from existing templates of device geometries. Recently, advances in computational inverse-design approaches, i.e., algorithmic techniques for exploring optical structures to satisfy the desired functionalities, have demonstrated

groundbreaking success in discovering non-standard geometry designs for passive optical devices to achieve exceptional performance enhancement/novel functionalities with tiny footprints [61, 62, 63]. Addressing the non-convex solution space, approaches to the inverse design of passive devices are often formulated as an optimization problem [64] and employ algorithmic solvers based on simulated annealing (SA) [65, 66], evolutionary [67, 68], objective-first [69], and adjoint-variables [61, 70]. As these algorithms are inherently computational-intensive and the numerical simulation of the electromagnetic-wave equations at each iteration renders the problem evaluation-expensive, the authors of [71, 72, 73, 74] propose machine learning–based algorithms to assist the evaluation of candidate geometry designs and accelerate the inverse design process.

2.1.3 Device Compact Modeling

For the accurate modeling and simulation of integrated photonics, two fundamental building blocks are indispensable, namely 1) the compact models for photonic devices and 2) the simulation tools for PICs comprising a collection of electronic and photonic devices. This section reviews some prior work on device compact modeling, while the circuit-level simulation methodologies are revisited in Section 2.2.1. In contrast to physical models that are primarily used for device design and optimization, compact models for photonic devices target circuit designers and aim to enable efficient and accurate circuit-level simulation of PICs.

The compact modeling of photonic devices can take different approaches, such as using scripting languages for numeric computing, e.g., MATLAB [75], using open-source modeling tools like the Python-based Caphe [47] and openEPDA [76], using platform-specific scripts for model description specified by photonic design automation (PDA) software vendors, e.g., Lumerical INTERCONNECT [77], Synopsys Rsoft Optsim Circuit [78], IPKISS by

Luceda Photonics [79], and VPIcomponentMaker™ Photonic Circuits by VPIphotonics [80], or using a standard analog hardware description language (HDL) like Verilog-A [81]. In particular, the Verilog-A method has drawn increasing attention for its SPICE-compatibility [82] that enables the co-simulation of photonic device models with their electronic driving circuitry by well-developed mixed-signal simulation engines, such as Cadence Spectre [83]. Examples of Verilog-A models for photonic devices include vertical-cavity surface-emitting lasers (VCSELs) [84, 85], traveling-wave Mach-Zehnder modulators (MZMs) [86], carrier-depletion/injection–based microring modulators [87, 88, 89], and multiple devices forming an optical link [90, 91, 92, 93]. Some photonic foundries and fabless design houses have also developed their Verilog-A models for photonic devices, such as imec [94], CEA-Leti [95], and Luxtera [96].

2.1.4 Process Variation Characterization

Due to the immature fabrication process of nanophotonics, photonic devices often suffer from significant process variations that lead to the degraded performance of integrated circuits and systems. Therefore, the variation patterns must be characterized for process optimization and variation alleviation.

Process variations of semiconductor manufacturing can exist in hierarchies. Spatial variations often manifest as a combination of wafer-level, intra-die, and inter-die patterns, while temporal variations may present at wafer-to-wafer and lot-to-lot levels [97]. Limited by the fabrication quantity, prior studies on variation characterization of photonic devices have mainly focused on sub-wafer levels. The methods proposed in [98, 99, 100] extract wafer-level variation patterns by modeling the variation magnitude as a function of the device location on the wafer. However, these methods only capture a smooth trend of the variations across the wafer and leave high-frequency components in the residuals that may

not be entirely random. As the measurement data reported in [101, 102, 103] indicate that the variation magnitude between two devices on the same wafer is roughly proportional to their physical distance, the authors of [104] and [105] accordingly propose to characterize the intra-die and inter-die variations separately. Nevertheless, the limited number of dies per wafer and devices per die in these studies prohibited further extraction of location dependencies of the variations. Instead, they assume independent Gaussian distributions for both the intra-die and inter-die variations, resulting in oversimplified variation models. In [106] and [107], variation characterization methods encompassing both the wafer-level and intra-die components are presented but still limited by the small number of fabricated devices.

On the other hand, there have been extensive studies on the variation characterization of electronic manufacturing by leveraging mass fabrication and rich measurement data. The extraction of spatial variation models can take approaches that roughly fall into two categories, using either decomposition-based methods at various levels [108, 109, 110, 111, 112, 113, 114] or estimation-based methods characterizing the entire wafer [115, 116, 117, 118, 119, 120]. However, the effectiveness of these methodologies requires a large number of devices per wafer. As the photonic devices tend to be bulkier than electronic ones and thus less can be fabricated on a single wafer, they are usually less effective if directly applied for characterizing the process variations of photonic wafers.

2.1.5 PDK Availability

Process design kits (PDKs) provide the fundamental building blocks for enabling efficient PIC design. These are libraries of optimized photonic devices developed by the foundries for integration with electronic-photonic design automation (EPDA) tools. Ultimately, a photonic PDK is expected to include accurate compact models for electronic-

photonic co-simulation, parameterized mask layouts for automated generation of device layouts with customizable geometry, detailed design rules that ensure manufacturability of user-submitted design, as well as statistics on process variations for quantification of performance uncertainty. Nevertheless, existing photonic PDKs are still immature and often only carry a limited amount of fixed device layouts and primitive design rules [121]. Examples of available photonic PDKs include those provided by AIM Photonics [122], imec [123], Tower Semiconductor [124], Advanced Micro Foundry [125], and CompoundTek [126] for silicon photonics, LIGENTEC [127] and LioniX International [128] for silicon nitride photonics, Fraunhofer HHI [129] and SMART Photonics [130] for InP/III-V photonics, as well as others listed in [131, 132, 133, 134].

2.1.6 Limitations Addressed in This Dissertation

At the device level, this dissertation addresses some limitations of existing methodologies for compact modeling and variation characterization of photonic devices. Specifically, existing compact models for photonic devices are usually developed based on a fixed device design and provided as black boxes for enabling circuit-level simulation. However, the lack of customizable parameters restricts the models from being used for design space exploration (DSE) and design optimization. Besides, many of the existing models are only validated by measurement data at the device level, and the results of circuit-level simulation may not be accurate by simply connecting these models. To this end, this dissertation includes compact models for crucial components of optical interconnects, validated by both device- and circuit-level measurement data, providing fundamentals for the optimization techniques proposed at the link and system levels. As for variation characterization, this dissertation addresses the limitations of prior work in terms of limited fabrication quantity and oversimplified variation models by developing a novel hierarchical model for the spa-

tial variations of photonic devices, characterized from the measurement data of a record number of fabricated microring devices.

2.2 Link-Level Design Automation and Optimization

The design and optimization of optical interconnects involve substantial effort put into the development of efficient methods for designing optimized optical links. Solutions at the link level include design automation methodologies for photonic integrated circuits, as well as techniques for analyzing and managing the impact of process variations on the performance and yield of fabricated optical links.

2.2.1 Electronic-Photonic Design Automation

Built upon the successful experience of CMOS electronic design, researchers and developers of the EPDA framework have focused on methodologies for circuit-level simulation, layout generation, and verification of PICs by adding support for photonic properties to existing electronic design environments. Other important features such as schematic-driven layout (SDL), automatic placement and routing (including phase-sensitive waveguides), layout versus schematic (LVS) checking, post-layout thermal effect simulation, and parasitic extraction (loss, coupling, reflection, etc) are also under active investigation and projected for future incorporation [31].

Electronic-Photonic Co-Simulation

Due to the extensive experience in electronic simulation possessed by electronic design automation (EDA) companies from the industry, the development of circuit-level EO co-simulation methodologies is mainly led by traditional EDA software vendors, e.g., Cadence [135], Synopsys [136], and Mentor Graphics (now part of Siemens [137]), by adding

interoperability with emerging PDA tools, including those from Lumerical (now part of Ansys [138]), Luceda Photonics [139], and VPIphotonics [140]. For frequency-domain simulation, the interactions between electronic and photonic devices are typically computed by S-parameters [141]. For time-domain simulation, however, the optimal time steps for simulating electronic and photonic responses can be drastically different. A workaround is to simulate the electronic and photonic components in dedicated engines (e.g., electronics in Cadence Spectre and photonics in Lumerical INTERCONNECT) with different time steps and periodically synchronize data between the two engines [142].

Curvilinear Layout Generation and Verification

The curvilinear shape of photonic components poses another challenge to electronic-photonic design automation in terms of both layout generation and verification, as the layouts of traditional electronic components are snapped to a fine Cartesian grid and comprised of Manhattan-style shapes. Algorithms are thus explored for efficiently representing curvilinear shapes within a Cartesian grid without affecting the desired optical properties. Options include approximating the curvilinear shapes with either many-vertex polygons [143] or rectangles [144]. To ease the burden of manual layout, parameterized cells (PCells) are made for photonic devices and circuit blocks to enable a hierarchical layout process [145]. PCells are usually implemented in a scripting language, such as SKILL in Cadence Virtuoso [146], AMPLE in Mentor Graphics Pyxis (now part of Siemens) [147], SPT in PhoeniX OptoDesigner (now part of Synopsys) [148], or a standard language (e.g., Python) as used in IPKISS by Luceda Photonics [79] and KLayout [149]. There are also standalone layout tools developed for improved efficiency, such as the Python-based PHIDL [150] and the C#-based VANDAL [151]. Recently, methodologies for schematic-driven layout are being actively investigated for further improving the productivity of PIC design [152, 153]. Originated from analog electronic design, SDL refers to the circuit schematic and populates the

layout with PCells and indicative connections between optical ports. Tools for automatic waveguide creation have also emerged, which can either calculate the waveguide shape between two ports without violating the design rules or generating the shape from the desired path drawn by the user [154].

Design rule checking (DRC) for photonic circuits also becomes challenging due to the curvilinear shapes. Standard DRC algorithms tailored to electronic design often result in false positives when directly applied to photonic layouts. To this end, recent developments with new DRC rules specially designed for curvilinear structures and all-angle polygons have been continuously improved for addressing this issue [155]. Another direction is to develop DRC algorithms with knowledge of the design intent [156]. For example, the linewidth of a discretized waveguide polygon is compared to the desired linewidth over the entire length of the waveguide, and excessive width variations are reported. Another level of verification is layout versus schematic checking, which verifies that the functional behavior of the layout is consistent with the schematic design. A preliminary step toward comprehensive photonic LVS has been initiated, which targets methodologies for automatically extracting the photonic netlist from the layout and verifying proper waveguide alignment [157]. More advanced LVS techniques, such as those for identifying unintentional waveguide crossings, are still under investigation. Furthermore, methodologies for parasitic extraction, such as coupling or reflection caused by the proximity of components, are also projected for future incorporation [158].

2.2.2 Process Variation Analysis and Management

Techniques at the link level for addressing the process variation challenges of optical interconnects are mainly proposed for 1) predicting the performance uncertainties and the yield of fabricated circuits and 2) mitigating the power consumption required for post-fab-

rication rectification of the process variations.

Uncertainty Quantification and Yield Prediction

Sensitivity analyses are required to predict the performance uncertainties and the yield of fabricated photonic circuits based on the process variation models characterized for individual devices. The Monte Carlo method has been considered a golden standard for yield prediction and uncertainty quantification, which requires a large number of circuit simulations with device variations randomly drawn from their modeled distributions [159]. However, the computational complexity of circuit simulations renders the Monte Carlo method evaluation-expensive, and the time required for simulating an adequate amount of samples may be prohibitive when there are many random variables. For electronic design, a simplified method called corner analyses can be used to reduce the number of required simulations, which only considers the best/worst case of individual devices. However, some photonic properties, such as the resonance wavelength, do not have an intrinsic good or bad value. To this end, novel techniques for uncertainty quantification and yield prediction of PICs have focused on reducing the required number of circuit simulations while keeping the prediction confidence comparable to the Monte Carlo method. The authors of [160] propose to use stochastic collocation, which builds a surrogate model of the performance metric of interest from a small set of circuit simulations, achieving accurate estimation of the performance uncertainties within a significantly less amount of computation time. Furthermore, the authors of [161, 162] propose to use polynomial chaos expansion (PCE) to model the performance metric of interest with a higher-order polynomial and determine the polynomial coefficients with a small number of circuit simulations. These methods have been successfully applied to photonic circuits for assessing the sensitivity of circuit performance to device-level variations [163,164,165,166]. In [167,168,169,170,171], the algorithm for coefficient computation is further accelerated with the adoption of techniques such as

tensor completion, which requires an even less number of circuit simulations.

Link-Level Variation Alleviation

Variation alleviation techniques aim to reduce the cost of post-fabrication rectification of process variations and are particularly important to the nominal energy efficiency (Definition 1.1 in Section 1.3) of microring-based optical links. Through device-level engineering, Liang *et al.* have demonstrated microring resonators that are hyper-sensitive to thermal changes and a theoretical reduction in thermal tuning power by orders of magnitude [172]. However, the feedback control of such sensitive devices becomes challenging in practice. There are more solutions proposed at the link level for variation alleviation. Techniques based on channel shuffling [173,174] and sub-channel redundancies [175,176] are proposed to reduce the expected power for thermally tuning the resonance wavelengths of the microrings. Another work by Wu *et al.* proposes an optimal pairing scheme for a batch of fabricated transceivers, which reduces the average tuning power required for each pair of transceivers formed from the batch [105]. Nevertheless, all of the above studies are based on oversimplified models for process variations characterized from outdated devices, and none have addressed the inherently low energy efficiency of resistive thermal tuners.

2.2.3 Limitations Addressed in This Dissertation

At the link level, this dissertation focuses on energy efficiency optimization for silicon microring–based optical interconnects by proposing multiple techniques for variation alleviation. These techniques are proposed to address the limitations of existing ones in terms of a limited design space, the inherently low efficiency of thermal tuners, and the inadequate exploitation of wafer-scale fabrication, all of which have restricted the room for energy optimization. Link-level simulation methodologies with variation awareness for ana-

lyzing the power budget of microring-based optical interconnects are also investigated for supporting the development of variation alleviation techniques.

2.3 System-Level Design Automation and Optimization

Challenges faced by optical interconnects at the system level come not only from the need for optimized devices and links but also from the runtime interaction and reconfiguration of system components as required by the running application. In other words, unlike that of device and circuit levels, system-level simulation and optimization solutions for optical interconnects are usually application-specific.

2.3.1 Simulation Tools and Methodologies for Optical Interconnects

The simulation of optical interconnects requires a higher level of model abstraction for photonic devices and switching fabrics, in pursuit of a balance between simulation speed and accuracy. Existing tools and methodologies for system-level simulation of optical interconnects roughly falls into two categories:

Configuration-based estimation The first category aims to provide estimations of various static properties of an optical network, such as the area, optical loss, crosstalk noise, link power budget, and link latency, based on a user-defined network configuration. For example, DSENT [177] is developed for fast area/power evaluations of up to hundreds of different network configurations, allowing for quick identification of optimal ones. Similarly, OEIL [178] is proposed for evaluating the energy efficiency, bandwidth density, crosstalk noise, and latency of an optical network based on its specific configuration. There are also system-level simulators dedicated to estimating the crosstalk noise [179] or thermal effects [180] of a given network configuration. In

addition, computational-efficient models for photonic devices are proposed in [181] for accelerating the evaluation of the propagation loss, interference, and phase shift in complex optical networks.

Event-driven simulation The second category further incorporates the capability of simulating traffic patterns or workload traces, thus providing access to some dynamic properties of the optical interconnects, such as the instantaneous power consumption and utilization of optical links at any time during the simulation. These tools are typically developed on top of existing event-driven simulators for (not necessarily optical) networks by adding power and latency models for optical components. Examples include PhoenixSim [182] based on the OMNeT++ framework [183], LioeSim [184] based on Orion [185], JADE [186] based on GEMS [187], and others reported in [188, 189, 190]. Besides, some architectural simulators, such as gem5 [191], Graphite [192], and Multi2Sim [193], as well as simulation frameworks for large-scale distributed data centers and computer networks, such as SimGrid [194] and CloudSim [195], can also be modified to simulate the dynamic properties of on-chip and off-chip optical interconnects driven by task execution.

2.3.2 Runtime Reconfigurable Optical Interconnects

Runtime reconfiguration strategies are indispensable for the effective energy efficiency (Definition 1.2 in Section 1.3) of optical interconnects under traffic dynamics. As the traffic patterns are significantly different for on-chip and off-chip communication scenarios, the reconfiguration strategies proposed so far are specially tailored to either optical network-on-chips (ONoCs) or off-chip optical networks for data center network (DCN) and high-performance computing (HPC) applications. The authors of [196] provide a comprehensive survey of reconfigurable optical networks in terms of the enabling technologies, algorithms,

and implementation considerations.

For ONoCs, several techniques are proposed to switch off the lasers for idle links at application runtime [197, 198]. However, due to the turn-on delay of the lasers (up to ~ 100 ns [18]), the energy wasted during the turn-on period may completely offset the energy saved by turning them off, and the application execution time may also be prolonged. Lan *et al.* address this issue by proposing a fine-grained strategy called Dynamic Laser Power Scaling (DLPS), which includes more power states for the lasers in addition to simple on and off [199]. Nevertheless, the effectiveness of DLPS is only observed for a small subset of traffic patterns, resulting in limited practicality.

As in ONoCs, the inter-arrival time between data transmission requests usually ranges from nanoseconds to hundreds of nanoseconds [200, 201], much smaller than the thermal time constants of microring tuning (~ 1 µs to ~ 1 ms [202, 203, 204]), the power reconfiguration for the microring tuning circuitry has been deemed unnecessary. On the contrary, in off-chip scenarios like DCNs and HPC interconnects, where the optical links can often stay idle for milliseconds to seconds [38], such reconfiguration capability becomes imperative for the effective energy efficiency of the network. However, strategies that include the microring tuning circuitry as a reconfiguration target are still lacking due to the long stabilization time required for microring tuning.

Another line of work focuses on reconfiguring the connectivity of the optical network at application runtime, which reduces the need for bandwidth overprovisioning by letting busy links borrow bandwidth from idle ones [205, 206]. These techniques also help to improve the network energy efficiency but are orthogonal to the runtime power reconfiguration strategies.

2.3.3 Limitations Addressed in This Dissertation

At the system level, this dissertation first develops simulation methodologies that support multiple power states of optical devices with variable reconfiguration time, a feature that is lacked by all existing simulators for optical interconnects. Then, this dissertation further proposes runtime power reconfiguration strategies for both on-chip and off-chip optical interconnects for further optimizing their effective energy efficiency.

Part II

Device-Level Modeling and Variation Characterization

Chapter 3

Compact Modeling for Lasers and Modulators

Photonic integrated circuits (PICs) fabricated on a heterogeneously-integrated silicon platform have demonstrated record levels of integration and bandwidth capacity. As PICs become more complex, the design and optimization of them demand modeling and simulation methodologies implemented in emerging electronic-photonic design automation (EPDA) platforms. In this chapter, the development of compact models for typical building blocks of optical network-on-chips (ONoCs) is introduced. These models are implemented in both SPICE-compatible electronic design automation (EDA) tools and dedicated photonic circuit simulators. Model validation is conducted at both device and link levels by the measurement data of a fabricated ONoC, allowing circuit designers to study the impact of individual device design on the overall link performance, paving the way to model-based design optimization of photonic integrated circuits.

3.1 Introduction

The ever-increasing demand for bandwidth capacity in telecommunication infrastructures has pushed the replacement of electronic transmission technologies by optical ones in links exceeding 10 m over the past decades [2]. In recent years, the explosive growth of

data-intensive computing applications and the tremendous advances in photonic integration technologies have driven the adoption of optical links based on photonic integrated circuits for short-reach datacom solutions [13]. Analogous to electronics, the integration density of PICs grows exponentially over time, often referred to as the Moore's Law for photonics [10]. The fast growing complexity of PICs calls for photonic design automation (PDA) tools [102]. A number of link- and system-level analyses for optical interconnects have been reported [174,207,208,209,210]. However, due to the analog nature of photonics, these studies depend on accurate modeling and simulation methodologies for integrated photonics at both device and circuit levels. Compact models for a variety of photonic devices have been reported, e.g., vertical-cavity surface-emitting lasers (VCSELs) [84], Mach-Zehnder modulators (MZMs) [86], and carrier-depletion/injection–based silicon microring resonators (MRRs) [87, 88, 89]. However, the lack of customizable parameters restricts these models from being used for design space exploration (DSE) and design optimization. Moreover, many of the existing models for photonic devices are only validated by measurement data at the device-level, and the results of circuit-level simulation may not be accurate by simply connecting these models.

In this chapter, we develop accurate compact models for key components of an ONoC, including lasers and modulators of multiple types. The models are implemented in formats that can be handled directly by traditional EDA tools (e.g., Cadence Virtuoso [211]) as well as dedicated photonic simulators (e.g. Lumerical INTERCONNECT [77] and Synopsys Rsoft Optsim Circuit [78]). We then validate the modeling methodologies by simulating each individual devices and an ONoC comprising multiple of them. The accuracy of the simulation results are extensively verified by measurement data at both device and circuit levels. Our proposed modeling and simulation approach enables the co-design of PICs and their electronic driving circuitry, which has to be designed separately in the past. Additionally, the enriched library of photonic devices paves the way to a process design kit (PDK) for our

heterogeneously-integrated silicon photonics platform. The methodologies introduced in this chapter allow an EDA-style design process for PICs and electro-optical (EO) systems.

The rest of the chapter is organized as follows. In Section 3.2, we present the compact modeling, parameter extraction, and model validation for individual devices. In Section 3.3, we introduce the architecture of a fabricated ONoC and further validate our developed models by circuit-level simulation of the ONoC. In Section 3.4, we demonstrate the application of our modeling and simulation methodologies for design space exploration and design optimization. And finally, in Section 3.5, we draw the conclusion of this chapter.

3.2 Device Characterization and Model Implementation

3.2.1 Compact Modeling for Lasers

An on-chip laser operating with a single wavelength and a high wall-plug efficiency (WPE) is crucial for a cost-efficient wavelength-division multiplexing (WDM) system [2]. The heterogeneously-integrated III-V on Si platform has been shown to offer several on-chip single-frequency laser solutions with progressively lower threshold and higher direct-modulation bandwidth [212, 213]. This section presents our compact models developed for lasers of multiple types, namely distributed feedback (DFB), distributed Bragg reflector (DBR), and microring lasers.

Distributed Feedback Laser

The DFB laser model is developed based on the one reported in [214]. The carrier and photon density of a diode laser, in both continuous-wave (CW) or directly-modulated oper-

ation modes, are governed by the coupled rate equations [22, 215]:

$$\frac{\partial N}{\partial t} = \frac{\eta_i I}{qV} - \left(BN^2 + CN^3\right) - g v_g N_p, \tag{3.1}$$

$$\frac{\partial N_p}{\partial t} = \Gamma g v_g N_p - \Gamma \beta_{sp} B N^2 - \frac{N_p}{\tau_p}, \tag{3.2}$$

where $g = g_{0N} \ln(N/N_{tr})$. The DC electrical characteristics follow the Shockley diode equation:

$$I = I_0 \exp \frac{q(V - IR)}{nkT}. \tag{3.3}$$

As the target application scenario of the developed model is for simulating on-chip optical links shorter than 1 cm, and the desired spacing of the multiplexed channels is much larger than the laser chirp, the current laser model does not include chirp or phase change caused by direct modulation. In addition, we further assume that the laser bandwidth is RC-limited and neglect the transit time effect. The model is based on the effective mirror mode, instead of a traveling-wave analysis, for simulation efficiency. Longitudinal variations of carrier (hole burning) is also not included. A list of the laser parameters included in the model is shown in Table 3.1. The numerical values are extracted from various sources including laser design parameters, material gain experiments, and laser light-current (LI) and current-voltage (IV) tests. Fig 3.1 shows the LI and IV curves of an on-chip DFB laser with a 400 µm–long cavity, a 26 textmu m mesa width, and a 1 µm Si waveguide.

The laser model as described by Eqs. (3.1)–(3.3) is implemented in Verilog-A [81], a standard hardware description language (HDL) language. The format is supported by traditional EDA tools for co-simulation of active photonic devices jointly with their electronic driving circuitry. The same set of laser parameters are also passed as inputs to the built-in laser models in specialized photonic circuit simulators, such as Lumerical INTERCONNECT and

Table 3.1: DFB laser parameter list.

Parameter	Symbol	Value	Unit
Current injection efficiency	η_i	0.5	
Electron charge	q	1.6×10^{-19}	C
Active region volume	V	3.36×10^{-11}	cm^3
Radiative recombination factor	B	4.29×10^{-10}	cm^3/s
Auger recombination factor	C	3.5×10^{-30}	cm^6/s
Gain coefficient	g_{0N}	1.96×10^{3}	cm^{-1}
Transparent carrier density	N_{tr}	1.08×10^{18}	cm^{-3}
Group velocity	v_g	7.5×10^{9}	cm/s
Active region confinement factor	Γ	5.6×10^{-2}	
Spontaneous emission factor	β_{sp}	1×10^{-4}	
Photon lifetime	τ_p	4.98×10^{-12}	s
Reverse saturation current	I_0	8.05×10^{-11}	A
Series resistance	R	5.77	Ω
Ideality factor	n	2	

Synopsys Rsoft Optsim Circuit. Having device models in the photonic circuit simulators helps accommodate the need for an accurate description of both active and passive components in an optical link. The simulated eye diagrams of a directly-modulated laser (DML) match the measured results in terms of modulation depth as well as overall eye shapes (Fig. 3.2). The agreement across different platforms supports the conclusion that the laser parameter extraction techniques are robust as the laser rate equation is implemented differently in each software platform. Equipped with these laser parameters, circuit designers

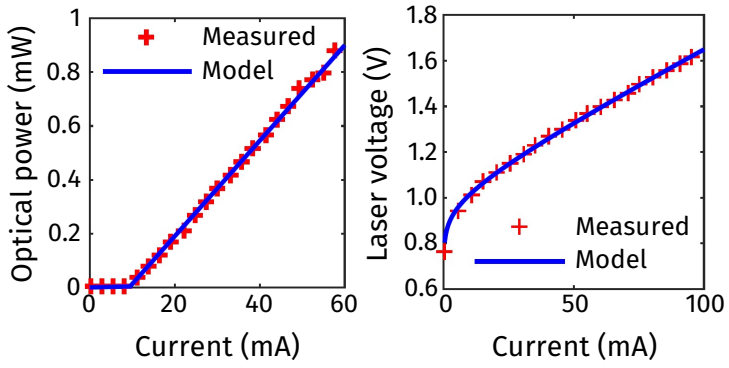


Figure 3.1: Measured and simulated LI and IV curves of the DFB laser (light output is from one facet).

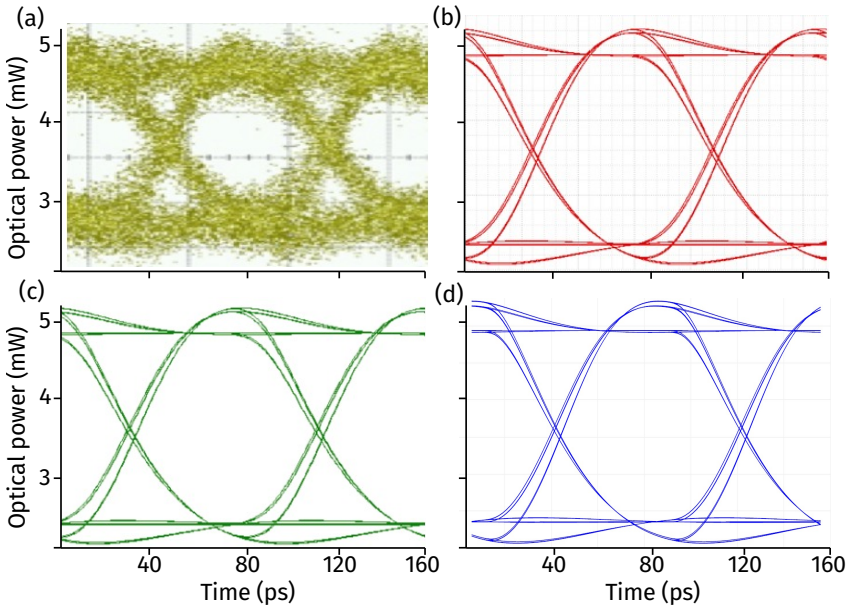


Figure 3.2: Eye diagrams of a 12.5 Gb/s directly-modulated DFB laser [214]: (a) measurement; (b) simulation in Cadence Virtuoso; (c) simulation in Synopsys Rsoft Optsim Circuit; and (d) simulation in Lumerical INTERCONNECT. The modulation depth is 3 dB in all four eye diagrams. In both the experiment and simulation setups, the laser is biased at 2.14 V and driven by a pseudorandom binary sequence (PRBS) signal with 0.75 Vpp. The eyes share the same axes.

can accurately simulate the static and dynamic behaviors of the DFB lasers.

Distributed Bragg Reflector and Microring Lasers

Compact models for DBR and microring lasers are also characterized and implemented based on the measurement data of the laser diodes reported in [216] and [217]. The models are also implemented in Verilog-A other platform-specific scripts for supporting simulations in Cadence Virtuoso, Synopsys Rsoft Optsim Circuit, and Lumerical INTERCONNECT. Fig. 3.3 shows the comparison of measured and simulated eye diagrams for lasers of multiple types in all three design automation platforms. The agreement of eye shapes between the measured and simulated results further confirms the robustness of our compact modeling methodology for hybrid III-V on Si lasers.

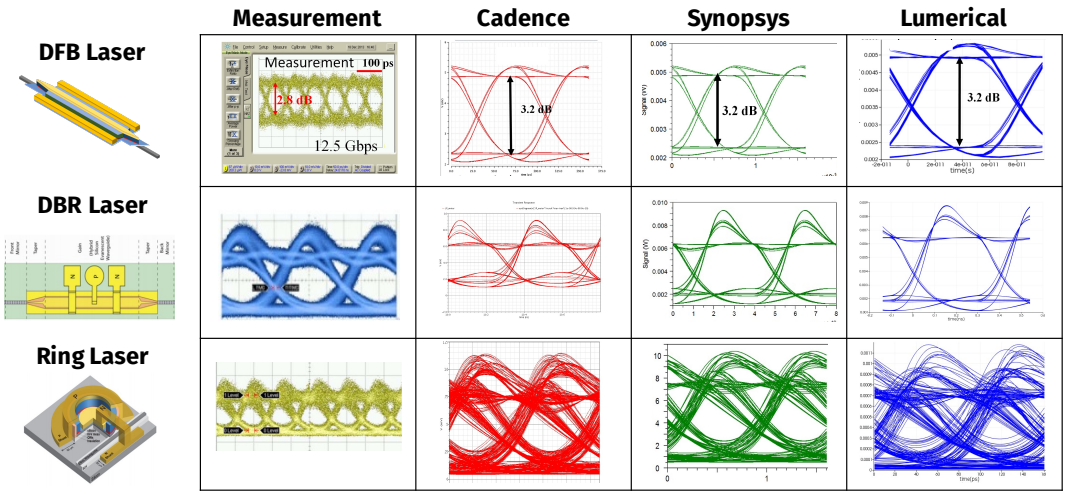


Figure 3.3: DFB, DBR, and microring laser models simulated in multiple design automation platforms.

Thermal Effects of Lasers

Our compact models for lasers account for the thermal effect and can simulate the degradations in laser performance caused by self-heating. First, the thermal coefficients for the threshold current and the slope efficiency [22] are extracted by fitting the following equations to the measured LI curve that shows a clear thermal roll-over (Fig. 3.4a):

$$I_{\mathrm{th}} = I_{\mathrm{th0}} \cdot e^{\Delta T / T_0}, \tag{3.4}$$

$$\eta_{\mathrm{d}} = \eta_{\mathrm{d0}} \cdot e^{-\Delta T / T_\eta}. \tag{3.5}$$

Then, the thermal dependencies of the transparent carrier density and the loss can be derived as:

$$N_{\mathrm{tr}}(T) = N_{\mathrm{tr0}} \cdot e^{\Delta T / T_{\mathrm{N}}}, \tag{3.6}$$

$$\alpha_{\mathrm{tot}} = \langle \alpha_{\mathrm{i}} \rangle + \alpha_{\mathrm{m}} = \alpha_{\mathrm{tot,0}} (1 + \Delta T / T_\alpha), \tag{3.7}$$

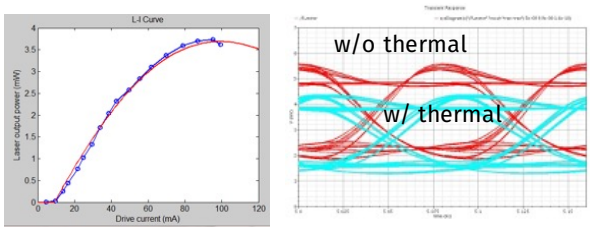


Figure 3.4: Laser thermal effects modeling: (a) extraction of thermal-dependent coefficients from measured LI roll-over; and (b) simulated degradations in the eye diagram.

where

$$T_\alpha = T_\eta, \tag{3.8}$$

$$\frac{2}{T_\mathrm{N}} + \frac{2g_\mathrm{th}}{g_{0\mathrm{N}}} \cdot \frac{1}{T_\alpha} = \frac{1}{T_0}. \tag{3.9}$$

Finally, the degradations in the laser eye diagram with thermal effects can be simulated by substituting the thermal-dependent carrier density and loss into the laser models described in this section. Fig. 3.4b shows the simulation of the previously modeled DFB laser under the thermal effect, where T_0 and T_η are derived to be 41.5 K and 50 K, respectively, and other model parameters the same as listed in Table 3.1.

3.2.2 Compact Modeling for Modulators

For application scenarios where a higher modulation speed is required, an architecture that employs CW lasers and high-speed external modulators becomes more desirable. To this end, we present the compact modeling methodology for silicon modulators developed based on an electroabsorption modulator (EAM) fabricated as part of an ONoC reported in [30].

Electroabsorption Modulator

A lumped EAM, relying on a strong quantum-confined stark effect (QCSE), has the advantages of a smaller footprint and lower energy consumption compared to traveling-wave EAMs MZMs. Lumped EAM design on the heterogeneous Si platform has demonstrated significant improvement in modulation bandwidth from 10 GHz to 30 GHz [23, 218]. The 3-dB bandwidth of an integrated DFB-EAM transmitter (Tx) on Si fabricated by quantum well intermixing has achieved 2 GHz [213]. The EAM modeled in this chapter is similar to the one described in [23]. The device is 100 µm–long and has 12 quantum wells centered at 1485 nm [30].

For DC optical transmission characteristics of the EAM, the dependence of the transmission spectrum on bias voltage can be described by a logistic equation:

$$T_{\text{opt}}\left(V_{\text{j}}\right) = L\left(\frac{1-b}{1+\exp\left(-k\left(V_{\text{j}}-V_0\right)\right)}+b\right), \tag{3.10}$$

where V_{j} is the voltage on the PIN junction; L is the insertion loss; V_0 is the transition voltage; and b is the residual optical transmission at a large bias voltage. Fig. 3.5a shows a close fit between measured spectrum and the proposed model. Similar curves are observed from 1550 nm to 1580 nm, reflecting a wide optical bandwidth. By applying a bias around V_0 and a modulation driving signal, the EAM can perform on-off keying (OOK) modulation of the optical signal. The EAM modeled in this chapter achieves a bias voltage lower than 2 V, which is more energy-efficient than the previously reported bias voltage of 3.5 V in [23].

The electrical characteristics of the EAM can be described by the equivalent quasi-static circuit model shown in Fig. 3.5b, where C_{p} is the parasitic capacitance introduced from the metal pads, R_{sm} the device series resistance, L_{m} the device series inductance, C_{im} the junction capacitance, V_{j} the junction voltage that determines the optical absorption and transmission in Eq. (3.10), I_{p} the photocurrent generated by the absorbed optical power

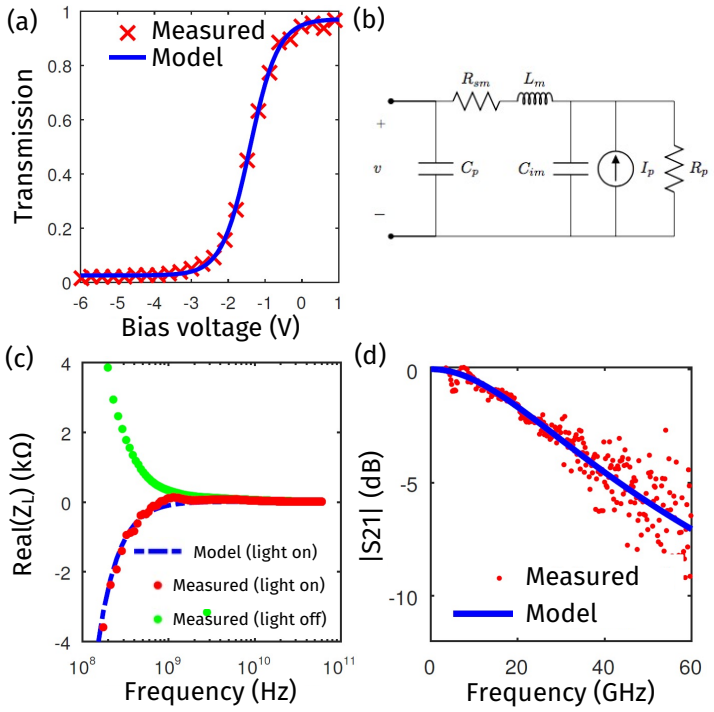


Figure 3.5: (a) EAM transmission characteristics plotted as a function of bias voltage; (b) quasi-static circuit model for the EAM; (c) real part of the EAM load impedance as a function of modulation frequency; and (d) S21 (EO) response of the EAM.

and fed back into the electrical circuit, and $R_\mathrm{p} = \mathrm{d}V_\mathrm{j} / \mathrm{d}I_\mathrm{p}$ the equivalent resistance of the photocurrent source [219, 220]. The circuit parameters can be extracted from static and high-speed measurements under different bias conditions. A lightwave component analyzer (LCA) is used for collecting the high-speed scattering parameters. The extracted values for a 150 µm–long device are shown in Table 3.2.

The alternating current (AC) in the junction is composed of the displacement current, related to the junction capacitance C_im, and the AC photocurrent, generated from optical absorption [219]. At a very low modulation frequency and under high optical power, the EAM behaves principally as a photodetector (PD), which sources current. Such behavior manifests as a negative real impedance. At a higher frequency ($1/(\omega C_\mathrm{m}) \ll R_\mathrm{p}$), almost all the AC current passes through the junction capacitance C_im, making the EAM appear

Table 3.2: EAM parameter list.

Symbol	Value	Unit	Source
L	30	dB	Transmission curve (fitting Eq. (3.10) to Fig. 3.5a)
b	2.54×10^{-2}		Transmission curve (fitting Eq. (3.10) to Fig. 3.5a)
k	2.63	V^{-1}	Transmission curve (fitting Eq. (3.10) to Fig. 3.5a)
V_0	−1.41	V	Transmission curve (fitting Eq. (3.10) to Fig. 3.5a)
C_p	15	fF	Metal pad test structure
R_{sm}	17	Ω	Z_L at high frequency under forward bias
L_m	21	pH	Z_L at high frequency under forward bias
C_{im}	71	fF	Z_L at high frequency under reverse bias
R_p	27.5	kΩ	Slope of IV curve under reverse bias

as a voltage-controlled device whose impedance has a positive real component. The EAM impedance as a function of frequency is shown in Fig. 3.5c.

As the displacement current dominates the AC current at moderately high frequency, the transit time effect, which only affects the AC photocurrent, will have little impact on the microwave property of the EAM. As a result, the EAM performance is mainly RC-limited and the extracted electrical parameters can be used for predicting the EO response (S_{21}) of the EAM, as shown in Fig. 3.5d. The 3-dB modulation bandwidth of this EAM is estimated to be 30 GHz.

The EAM compact model is constructed by implementing both the equivalent circuit and the optical transmission characteristics. When a modulating voltage is applied to the EAM, it is first filtered by the circuit network. The resulting voltage across the junction is then used by Eq. (3.10) for calculating the optical transmission $T_{opt}(V_j)$. The validity of this model is supported by the close agreement between the measured and simulated curves in Fig. 3.5.

3.3 Link-Level Simulation and Model Validation

Besides the compact models developed for the lasers and modulators described above, we also characterized and implemented compact models for other photonic devices that are key to ONoCs, including traveling-wave MZMs based on the method described in [86], silicon microring modulators based on the method described in [88], as well as PIN photodetectors, arrayed waveguide gratings (AWGs), and silicon waveguides based on the method described in [221]. The modeling effort leads to an enriched library of photonic devices, which enables circuit-level simulation of integrated photonic links and networks. In addition to individually validate the compact models at the device level, simulation of optical links comprising the modeled devices could further validate the modeling methodologies by circuit-level measurement results. In this section, we demonstrate the simulation of an ONoC fabricated on a heterogeneous Si platform, as reported in [30], using the compact models developed for individual devices.

Overview of the Target ONoC

As illustrated in Fig. 3.6, the fabricated ONoC consists of eight WDM transceiver (TRx) nodes connected by a reconfigurable ring bus, with eight high speed TRx channels in each node. Within each WDM node, on the Tx end, there are eight single-mode DFB lasers, eight high-speed EAMs, and eight monitoring photodetectors (MPDs), one for each channel. On the receiver (Rx) end, there are eight InGaAs/Si PIN PDs, one for each channel. Some of the receivers use semiconductor optical amplifiers (SOAs). AWGs are used for signal (de)multiplexing (MUX/DEMUX) at the Tx and Rx ends.

Although each optical link works at a different wavelength, the building blocks share similar designs. Among them, the EAM and the PD are broadband, while the lasing wavelength of the DFB laser can be effectively tuned. Because the crosstalk between channels is

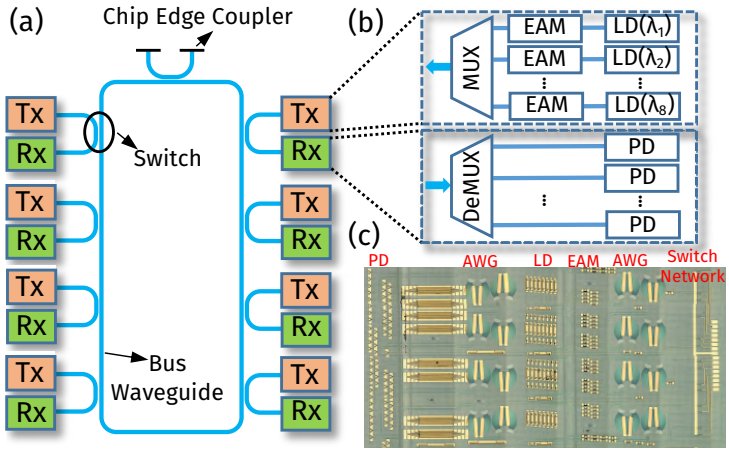


Figure 3.6: (a) Target ONoC architecture; (b) enlargements of a Tx and a Rx; and (c) microscopic image of the fabricated ONoC including 8 TRx nodes.

below −25 dB, provided by the AWG, the characterization of a single-wavelength optical link can be used for evaluating the overall network performance.

Full-Link Simulation and Analysis

An end-to-end single-channel transceiver link is simulated based on our compact device models. Fig. 3.7 demonstrates the schematic view in Cadence Virtuoso environment, which follows the measurement setup. The EAM is driven by a PRBS source swinging between −1 V and −2 V. The PD is biased at −3 V. The broadband switches have a wavelength operation range from 1550 nm to 1570 nm (2.5 THz). For this reason, an insertion loss is introduced for the switch without imposing any bandwidth limitation on the link. In addition, optical dispersion is not simulated for this on-chip link, as our laser model does not include chirp, and the waveguide model featured in [221] only captures the propagation loss.

First, the dynamic performance of the full transceiver link is studied. The frequency response of the link can be obtained by multiplying the individual frequency responses of the bandwidth-limiting components, namely the EAM, PD and AWGs. In Fig. 3.8, the simulated frequency response reasonably captures the link behavior indicated by the measure-

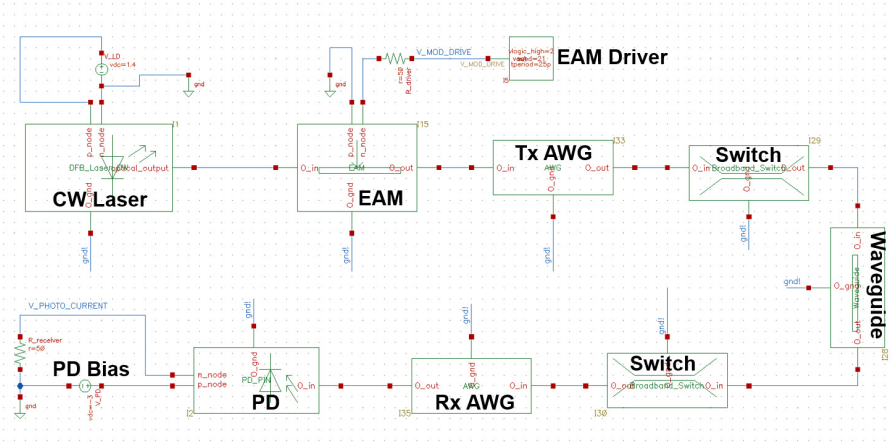


Figure 3.7: Schematic view of the transceiver link in Cadence Virtuoso. The 50 Ω resistors in the EAM driver and after the PD represent the internal resistance of the pattern generator and the oscilloscope.

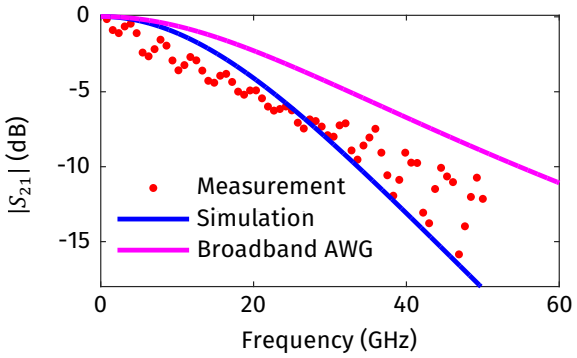


Figure 3.8: Simulated and measured frequency responses of the modeled single-channel transceiver link. The pink line represents the predicted frequency response of the transceiver link integrated with broadband AWGs.

ment data. Next, transient simulations are performed to obtain the eye diagrams of the received signal. In the experiment reported in [30], the data transmission test was applied with a 40 Gb/s pattern generator and $(2^7 - 1)$ bits of PRBS signal, which is reproduced in our simulation setup. The simulated eye diagram in Fig. 3.9c is in reasonable agreements with the measurement shown in Fig. 3.9a. The quantified eye characteristics are summarized in Table 3.3. The discrepancies between the simulated eye and the measured eye can be explained by the additional noise introduced by the electronic components in the measure-

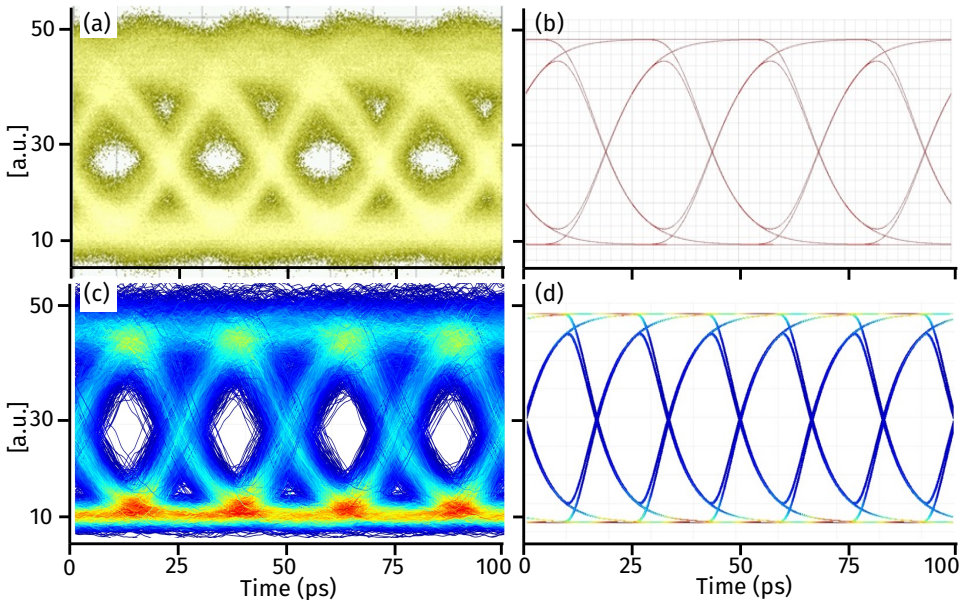


Figure 3.9: (a) Measured eye diagram of the single-channel transceiver link at 40 Gb/s; (b) noise-free full-link eye diagram simulated in Cadence Virtuoso at 40 Gb/s; (c) full-link eye diagram simulated in Lumerical INTERCONNECT at 40 Gb/s, where the power spectral density (PSD) of the noise source at the laser and the modulator are set to 1×10^{-17} W/Hz, and the PD thermal noise variance is 3.328×10^{-22} A^2/Hz; (d) simulated noise-free transceiver link eye diagram at 60 Gb/s in Lumerical INTERCONNECT with the AWGs removed, where the software-calculated extinction ratio is the same as the noise-free eye diagram with AWGs at 40 Gb/s.

Table 3.3: Characteristics of measured and simulated eye diagrams of the TRx link.

Data source	**Eye width** (ps)	**Rise time** (ps)	**Fall time** (ps)
Measurement (Fig. 3.9a)	~10	~11	~11.5
Simulation (Fig. 3.9c)	10.1	10.8	11.2

ment setup. These electronic components, such as the function generator, transimpedance amplifiers (TIAs), as well as coaxial cables and connectors, are not included in the simulation.

If the ONoC is integrated with its driver electronics, the nominal resistance can be customized to be smaller than 50 Ω. In this case, the RC-limited bandwidth of the EAM and PD can be improved, leaving the AWGs as the major bandwidth-limiting components. For this reason, the impact of the spectral filtering effect, introduced by the AWG transmission

spectrum, on the overall speed of the transceiver link needs to be investigated. This is accomplished by removing the AWGs from the transceiver link in the simulation. The data rate is then gradually increased so that the software-calculated extinction ratio is the same as in a simulated full transceiver link. The extinction ratio is chosen as the figure of merit because it directly reflects the eye openness. Based on this procedure, the transceiver having sufficiently broadband AWGs is predicted to have the same eye openness at 60 Gb/s, shown in Fig. 3.9d, as the simulated eye with AWGs at 40 Gb/s, shown in Fig. 3.9b. Such improvement in the AWG performance will scale the transmission capability of the ONoC up to 3.84 Tb/s.

The energy consumption of the transceiver link is calculated by adding up the power usage of individual active components. DC power is found by calculating the product of the bias voltage and bias current. The DFB laser and PD consume direct current (DC) power only, while the EAM consumes both AC and DC power. The energy E dissipated in an EAM in 1-bit cycle is expressed as:

$$E = CV_{\mathrm{PP}}^2/4 + I_{\mathrm{avg}}V_{\mathrm{bias}}/B, \tag{3.11}$$

where C is the junction capacitance; V_{PP} is the peak-to-peak voltage swing; and B is the bit rate. For the EAM, C is ~70 fF and the voltage swing is 1 V. The first and second summands in Eq. (3.11)) correspond to the AC and DC power, respectively. For this EAM modulated at 40 Gb/s, the AC power is 18 fJ/bit. The power consumption of the entire transceiver link is 678 fJ/bit, 90 % of which is from the on-chip laser. To make the ONoC competitive as an off-chip interconnect, it is necessary to reduce the photonic device energy further to the order of tens of fJ/b [2, 222]. Such a tight power budget for an on-chip transmitter calls for a different type of laser with lower threshold while still being temperature– and surface-recombination–insensitive when the device is made sufficiently small. Quantum-dot (QD) lasers grown or bonded on Si have shown promising results in each aspect [15, 223].

3.4 Model-Based Design Optimization

In this section, we demonstrate the application of our validated device models for ONoC design space exploration and optimization.

3.4.1 Design Space Exploration of PD Design

The fabricated photonic chip includes a large variety of PD test structures with different device lengths and widths, only a subset of which are used in the modeled transceiver links. Table 3.4 shows the key parameters of several PD designs, where PD #1 is used in the full transceiver link introduced in Section 3.3. The bandwidth (BW) is calculated by $(R+50\,\Omega)\cdot C$, considering a 50 Ω receiver. The optical modulation amplitude (OMA) is calculated as $I_1 - I_0$, where I_1 and I_0 are current for detected 1 and 0 levels of the eye diagrams in link simulation. Device parameters are extracted at −3 V bias.

Due to the nature of the waveguide photodetector design, the responsivity increases while the bandwidth decreases with the device length. From Table 3.4, PDs #2–4 have a higher responsivity but a lower bandwidth than PD #1. Therefore, it is possible to replace PD #1 with each of PDs #2–4 to trade bandwidth for responsivity to obtain a higher OMA. Then, PD #1 is swapped with each of PDs #2–4 in the full link simulation to get the corresponding link eye diagram at 40 Gb/s. From these eye diagrams, the OMA values are extracted as shown in the last column in Table 3.4. The simulation results show that PD #4

Table 3.4: PD design space and simulated OMA.

PD #	**Width** (μm)	**Length** (μm)	**Resp.** (A/W)	**Cap.** (fF)	**Res.** (Ω)	**BW** (GHz)	**OMA** (μA)
1	4	30	0.45	38.9	38.9	46.0	29.3
2	2	50	0.48	38.6	52.0	40.4	30.6
3	2	75	0.56	54.8	36.0	33.8	34.4
4	2	100	0.66	71.5	29.4	28.0	38.3

leads to the highest OMA mainly because of its high responsivity. A higher OMA will result in a better signal-to-noise ratio (SNR) and, in turn, a lower bit error rate (BER). This study of PD designs demonstrates that, with these models, the designers can now explore photonic device designs for link and system optimization.

3.4.2 Optimization of EAM Driving Voltage

In the preceding link simulations, the driving voltage for the EAM is set at 1 V_{PP}, which saves energy but compromises the OMA. Such phenomenon can be seen from the EAM transmission curve in Fig. 3.5a, where 1 V_{PP} swing is not enough to reach the maximum and minimum of the optical transmission. The EAM driving voltage swing is then increased, and the eye-diagrams are re-simulated at 40 Gb/s. The extracted OMA and modulation energy consumption are plotted in Fig. 3.10. Other device parameters and driving conditions are the same as those in Section 3.3. The simulation results show that the OMA could be enhanced by increasing the EAM driving voltage swing at the cost of consuming greater transmitter power. A higher OMA reduces the requirement for the receiver sensitivity, which in turn reduce the receiver power consumption [224]. Therefore, with these photonic models and simulation methodologies, the transmitter and receiver could be co-optimized for minimizing the overall power consumption while meeting the transmission quality requirement.

3.5 Concluding Remarks

In this chapter, accurate circuit-level models for silicon photonic devices are developed based on their electrical and optical properties. These models have been validated in multiple aspects based on the measurement data from fabricated devices and circuits. The models are described in standard hardware description languages suitable for SPICE-compatible

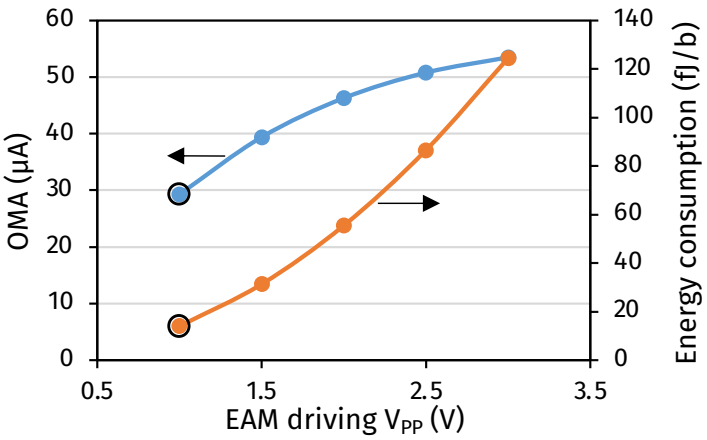


Figure 3.10: OMA and modulation energy consumption w.r.t. different EAM driving voltages, the black circles indicating the operating points of the EAMs in the measured and simulated ONoC.

simulators as well as in commercial photonic circuit simulators. In both cases, the simulation results of a full optical transceiver link has demonstrated reasonable agreement with the measurements. These simulation results provide insights to the overall performance of the link in terms of both speed and energy efficiency. Most importantly, the models assist and guide future PIC design through design space exploration and optimization. Ultimately, the modeling and simulation methodologies enable EO co-simulation, allowing designers to co-optimize photonic devices with electronic circuits seamlessly.

Chapter 4

Spatial Variation Modeling for Microring Resonators

Photonic integrated circuits (PICs) suffer from large process variations. Effective and accurate characterization of the variation patterns is a critical task for enabling the development of novel techniques to alleviate the variation challenges. In this chapter, we present a hierarchical approach that effectively decomposes the spatial variations of silicon microring–based optical transceivers into wafer-level, intra-die, and inter-die components. We then demonstrate that the characterized variation model can be used for generating credible synthetic data of silicon microring–based optical transceivers, which are indispensable for the development and validation of link- and system-level solutions for variation alleviation. We further demonstrate the utility of our variation characterization method for accurate yield prediction based on partial measurement data.

4.1 Introduction

Integrated silicon photonics has emerged as a promising alternative to traditional CMOS electronics [11, 12]. To substantially improve the bandwidth capacity and energy efficiency of future high-performance computing (HPC) systems, optical interconnects have been proposed to host the increasingly intensive communication traffic [1, 13]. However, due to the

intrinsically high sensitivity of silicon photonic devices to fabrication imperfection, optical interconnects have been suffering from large process variations and relying on post-fabrication compensation techniques to alleviate the variation challenges [105,173,174,175, 176,207,225,226]. Therefore, effectively and accurately characterizing the variation patterns is a critical task for enabling the development of link- and system-level solutions for variation alleviation [227].

Process variations of semiconductor manufacturing can exist in hierarchies. Spatial variations often manifest as a combination of wafer-level, intra-die, and inter-die patterns, while temporal variations may present at wafer-to-wafer and lot-to-lot levels [97]. Limited by the fabrication quantity, prior studies on variation characterization of photonic devices have mainly focused on sub-wafer levels. The methods proposed in [98, 99, 100] extract wafer-level variation patterns by modeling the variation magnitude as a function of the device location on the wafer. However, these methods only capture a smooth trend of the variations across the wafer and leave high-frequency components in the residuals that may not be entirely random. As the measurement data reported in [101, 102, 103] indicate that the variation magnitude between two devices on the same wafer is roughly proportional to their physical distance, the authors of [104] and [105] accordingly propose to characterize the intra-die and inter-die variations separately. Nevertheless, the limited number of dies per wafer and devices per die in these studies prohibited further extraction of location dependencies of the variations. Instead, they assume independent Gaussian distributions for both the intra-die and inter-die variations, resulting in oversimplified characterization. In [106] and [107], variation characterization methods encompassing both the wafer-level and intra-die components are presented but still limited by the small number of fabricated devices.

On the other hand, there have been extensive studies on the variation characterization of electronic manufacturing by leveraging mass fabrication and rich measurement data.

The extraction of spatial variation models can take approaches that roughly fall into two categories, using either decomposition-based methods at various levels [108, 109, 110, 111, 112, 113, 114] or estimation-based methods characterizing the entire wafer [115, 116, 117, 118, 119, 120]. However, the effectiveness of these methodologies requires a large number of devices per wafer. As the photonic devices tend to be bulkier than electronic ones and thus less can be fabricated on a single wafer, they are usually less effective if directly applied for characterizing the process variations of photonic wafers.

In this chapter, we present an effective method for accurately characterizing the spatial variations of silicon microring–based optical transceivers, and to explore its applications to link- and system-level design optimization. We propose to adopt a hierarchical approach incorporating domain-specific knowledge, spatial-frequency analyses, and low-rank tensor factorization methods to decompose the variations of optical test items into wafer-level, intra-die, and inter-die components. We also provide analyses on the residuals from each step to evaluate their randomness. Wafer-scale measurement data of 66 transceivers containing 3000+ microrings are used for validating our variation characterization method.

We then demonstrate that the characterized variation model can be used to generate credible synthetic data of microring-based optical transceivers for the development of novel variation alleviation techniques. The synthetic data generated by our approach closely resemble the actual data and can estimate the system performance with substantially higher accuracy than existing methods. We further explore the utility of our variation characterization method for yield prediction, and demonstrate that a reasonably accurate prediction of the transceiver yield can be achieved with only 60 % of measurement data of a wafer.

The rest of this chapter will be organized as follows. In Section 4.2, we introduce the fabricated microring-based optical transceivers, and describe the preprocessing steps to extract the variations of individual test items from raw measurement. In Section 4.3, we provide implementation details of our variation characterization method and residual analyses.

In Section 4.4, we demonstrate the applications of our variation characterization method for credible synthetic data generation and accurate yield prediction. And finally, in Section 4.5, we make the concluding remarks.

4.2 Overview and Data Preprocessing

4.2.1 Background of Microring-Based Optical Transceivers

A silicon microring resonator (MRR), as illustrated in Fig. 4.1a, is a highly wavelength-selective device that can modulate or filter optical signals at its specific resonance wavelength [25]. The transmission spectrum of an MRR can be modeled as a Lorentzian function parameterized by three key properties, namely λ_r, ER, and Q:

$$T(\lambda) = 1 - \frac{1 - 1/\mathrm{ER}}{1 + (2\mathrm{Q} \cdot (\lambda - \lambda_r)/\lambda_r)^2}, \tag{4.1}$$

where λ_r, ER, and Q are the resonance wavelength, extinction ratio, and quality factor of the spectrum (Fig. 4.1b). A silicon microring–based optical transceiver (TRx) achieves dense wavelength-division multiplexing (DWDM) by deploying cascaded MRRs alongside a com-

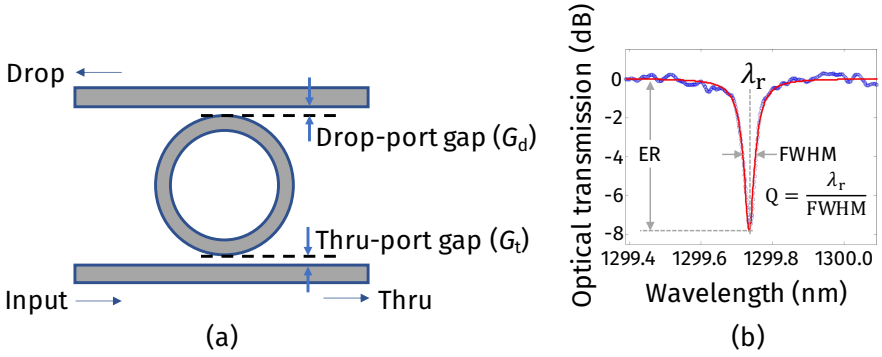


Figure 4.1: Illustrations of the typical (a) geometry design and (b) thru-port transmission spectrum of a microring resonator, where FWHM denotes the full width at half maximum.

mon waveguide [26,228,229]. At the transmitter (Tx) side, voltages applied to the microring modulators can slightly shift their resonance wavelengths to perform on-off keying (OOK) modulation of the optical signal. At the receiver (Rx) side, corresponding microring filters can couple the signal out for detection. The overall transmission spectrum of a Tx/Rx can be modeled as the product of all transmission functions of individual microrings:

$$T_n(\lambda) = \prod_{i=1}^{n} \left(1 - \frac{1 - 1/\mathrm{ER}_i}{1 + \left(2\mathrm{Q}_i \cdot (\lambda - \lambda_{\mathrm{r},i})/\lambda_{\mathrm{r},i}\right)^2}\right), \tag{4.2}$$

where n is the number of DWDM channels, which is also the number of microrings in the Tx/Rx.

The variations of λ_r, ER, and Q have a significant impact on the energy efficiency of the TRx. First of all, the resonance wavelengths of the Tx/Rx channels must be actively tuned to align with the carrier wavelengths of the laser source, and this tuning power is non-trivial [230]. Besides, variations of both ER and Q affect the loss and crosstalk noise of the optical link, and thus the optical power required to attain a specific data rate will vary. Therefore, we focus on these three key parameters in this chapter for variation characterization.

4.2.2 Overview of Fabricated Devices

Comprehensive measurement data of the transmission spectra of 24-channel microring-based transceivers are collected from a fabricated wafer containing 66 of them (Fig. 4.2a). The fabrication was carried out by a commercial foundry (STMicroelectronics [231]) on a 300 mm silicon-on-insulator (SOI) wafer. Each die consists of a TRx block (marked in Fig. 4.2b). In each TRx block (Fig. 4.2c), a 24-channel Tx and a 24-channel Rx were measured for optical transmission spectra. Thus, the total number of microrings involved is $66 \times 2 \times 24 = 3168$. Each group of 24 microrings starts with a 5 µm radius and ramps up

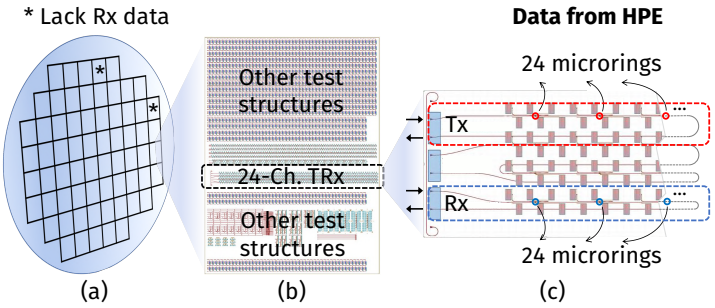


Figure 4.2: Organization of measured devices: (a) a wafer of 66 dies; (b) each die consisting of one TRx; and (c) each Tx/Rx consisting of 24 microrings.

to 5.046 µm radius with a step size of 2 nm. All microrings share the same drop-port gap width (G_d) of 200 nm. The thru-port gap widths (G_t) differ for Tx and Rx microrings and are 175 nm and 200 nm, respectively.

4.2.3 Model-Based Data Preprocessing

We preprocess the measured transmission spectra to extract the values and variations of λ_r, ER, and Q for each microring.

Extraction of λ_r, ER, and Q

The resonance wavelengths are extracted by detecting significant peaks of the measured spectra. In practice, a resonance peak may split into two adjacent ones due to back-reflection in the microring [232]. Techniques suggested in [233] are employed to effectively recognize the peak splitting as a single resonance. Then, ER and Q are extracted for all microrings by fitting Eq. (4.2) to the measured spectra via nonlinear least squares (an example shown in Fig. 4.3). We plot the extracted values of λ_r, ER, and Q in three wafer maps (Figs. 4.4a–c). Each wafer map is 16×240, in which each 2×24 block corresponds to a die. Odd rows are for Tx and even rows are for Rx. The ranges of the color bars are obscured to protect the manufacturer. Note that some microrings are found without signif-

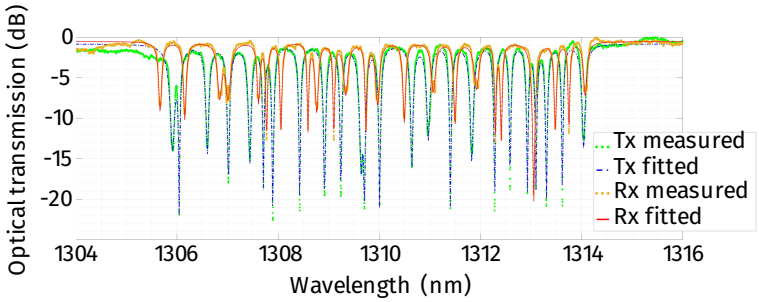


Figure 4.3: Measured and fitted transmission spectra of a 24-channel transceiver.

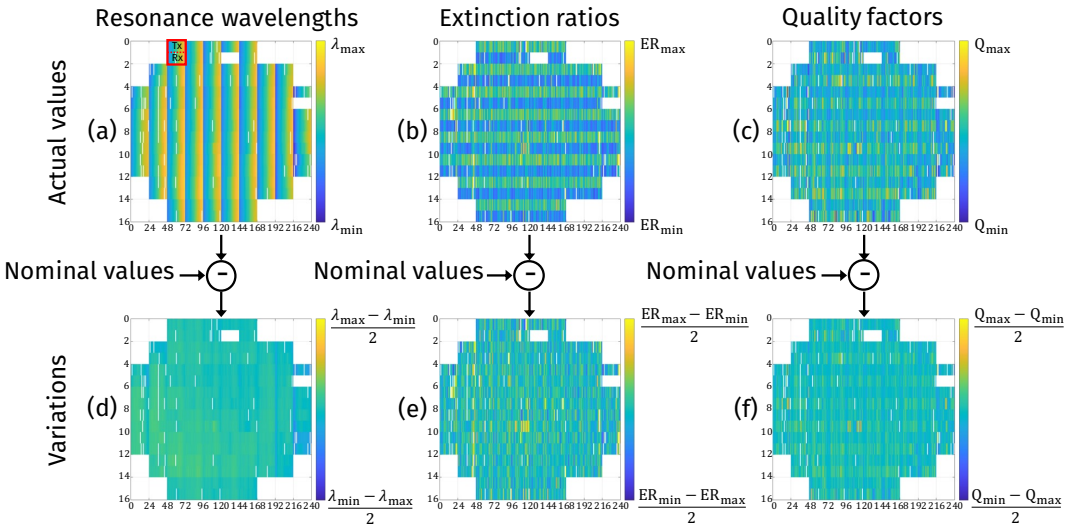


Figure 4.4: Wafer maps for the (a–c) actual values, and (d–f) variations of λ_r, ER, and Q.

icant resonance and represented by blank pixels in the wafer maps.

Computation of Nominal Values

The variations of λ_r, ER, and Q can be computed as the actual values subtracted by the nominal values. However, the nominal values are related to the microring geometry, and thus are different for microrings with various radii and gap widths. Figs. 4.4a–c clearly depict the incremental resonance wavelengths within each transceiver, and the differences between Tx/Rx rows for ER and Q.

Models for microring resonators [92] specify that

$$\lambda_{\mathrm{r}} = \frac{2\pi n_{\mathrm{eff}}}{m} \cdot r = c_0 \cdot r, \tag{4.3}$$

$$\mathrm{ER} = \left(\frac{\delta_{\mathrm{t}} + \delta_{\mathrm{d}}}{\delta_{\mathrm{t}} - \delta_{\mathrm{d}}} \right)^2, \tag{4.4}$$

$$\mathrm{Q} = \frac{2\pi \lambda_{\mathrm{r}}}{\mathrm{FSR} \cdot (\delta_{\mathrm{t}} + \delta_{\mathrm{d}})}, \tag{4.5}$$

where FSR = 13.09 nm is the average free spectrum range for these fabricated microrings, and

$$\delta_{\mathrm{t}} = c_1 \cdot \exp\left(-c_2 \cdot G_{\mathrm{t}}\right) + c_3, \tag{4.6}$$

$$\delta_{\mathrm{d}} = c_4 \cdot \exp\left(-c_5 \cdot G_{\mathrm{d}}\right) + c_6 \tag{4.7}$$

are the coupling ratios of the thru port and the drop port, respectively. We define an auxiliary function

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{\lambda}_{\mathrm{r}}^* - c_0 \cdot \boldsymbol{r} \\ \mathbf{ER}^* - \left(\frac{\boldsymbol{\delta}_{\mathrm{t}} + \boldsymbol{\delta}_{\mathrm{d}}}{\boldsymbol{\delta}_{\mathrm{t}} - \boldsymbol{\delta}_{\mathrm{d}}} \right)^2 \\ \mathbf{Q}^* - \frac{2\pi \boldsymbol{\lambda}_{\mathrm{r}}^*}{\mathrm{FSR} \cdot (\boldsymbol{\delta}_{\mathrm{t}} + \boldsymbol{\delta}_{\mathrm{d}})} \end{bmatrix}, \text{ where } \begin{cases} \boldsymbol{\delta}_{\mathrm{t}} = c_1 \cdot \exp\left(-c_2 \cdot \boldsymbol{G}_{\mathrm{t}}\right) + c_3 \\ \boldsymbol{\delta}_{\mathrm{d}} = c_4 \cdot \exp\left(-c_5 \cdot \boldsymbol{G}_{\mathrm{d}}\right) + c_6 \end{cases}. \tag{4.8}$$

Here, $\boldsymbol{\lambda}_{\mathrm{r}}^*$, $\mathbf{ER}^*$, and $\mathbf{Q}^*$ are the actual values extracted from the measured spectra, and $\boldsymbol{r}$, $\boldsymbol{G}_{\mathrm{t}}$, and $\boldsymbol{G}_{\mathrm{d}}$ are the design values of the microring radii and thru/drop-port gap widths. The coefficients c_0–c_6 can then be solved as $\left(\arg\min \left\| \boldsymbol{y} \right\|_2^2 \right)$ using existing nonlinear minimization algorithms [234]. The nominal values of λ_{r}, ER, and Q for each microring can then be computed based on Eqs. (4.3)–(4.7).

Figs. 4.4d–f plot the variations of λ_{r}, ER, and Q in three wafer maps after subtracting the nominal values from the actual values. The color bars are also shifted to have zero mean.

In the next section, we describe our hierarchical characterization method for the spatial variations of λ_r, ER, and Q.

4.3 Hierarchical Variation Characterization

We present our method for effectively decomposing the spatial variations of microring-based optical transceivers into wafer-level, intra-die, and inter-die components. As categorized in [110], wafer-level variation components capture a low-frequency trend of the variations across the wafer. They are usually assumed to be related to the fabrication process and independent of the layout. Intra-die components capture the device variations within a die and manifest themselves as repetitive patterns across the wafer. Inter-die components further capture the die-to-die differences that still appear more systematic than random.

4.3.1 Wafer-Level Variation Components

The authors of [98] decompose the spatial variations of optical devices into leveling and radial components. The two basis functions are written as

$$f_{\text{leveling}}(x, y) = a_1 x + a_2 y, \tag{4.9}$$

$$f_{\text{radial}}(x, y) = a_3 \cdot \cos\left(a_4 \sqrt{x^2 + y^2} + a_5\right), \tag{4.10}$$

where x and y are the coordinates of the device on the wafer. We further introduce two basis functions, namely wafer-edge and wafer-center, to capture the different behavior of

dies located near the edge/center of the wafer [113]:

$$f_{\text{edge}}(x,y) = \begin{cases} a_6, & (x,y) \in E \\ 0, & \text{otherwise} \end{cases}, \tag{4.11}$$

$$f_{\text{center}}(x,y) = \begin{cases} a_7, & (x,y) \in C \\ 0, & \text{otherwise} \end{cases}, \tag{4.12}$$

where E and C are predefined sets of edge dies and center dies, respectively. The coefficients are derived by robust regression [235] that fits the sum of the basis functions to the spatial variations. It is noteworthy that some alternating patterns between the Tx and Rx rows can be observed in the variation wafer maps for ER and Q (Figs. 4.4e and f), and the authors of [114] suggest extra basis functions dedicated to those. However, we assume that the alternating patterns largely relate to the different thru-port gap widths (G_t) of Tx/Rx microrings, but that the wafer-level variation components are layout independent. Later we show that these alternating patterns can be effectively extracted as part of the intra-die variation components.

In our analysis, we first remove some outliers from the wafer maps by identifying those more than 1.5 interquartile ranges [236] above the 75 % quartile or below the 25 % quartile. This method works generally well when the data are not normally distributed. We then scale the variations into the range of $[-0.5, 0.5]$ before decomposition. Fig. 4.5 shows the scaled variations for λ_r with the outliers removed and the wafer-level decomposition results. For simplicity, the three components corresponding to Eqs. (4.10)–(4.12) are merged into a single plot for the radial component. The significance of each component is computed as the variance of the component divided by that of the total variations. It can be observed that over 40 % of the variations are left in the residuals where systematic patterns still exist, despite the extracted wafer-level components.

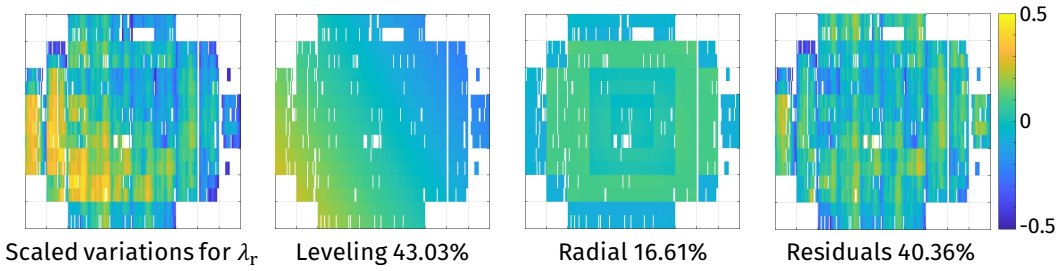


Figure 4.5: Wafer-level variation components for λ_r.

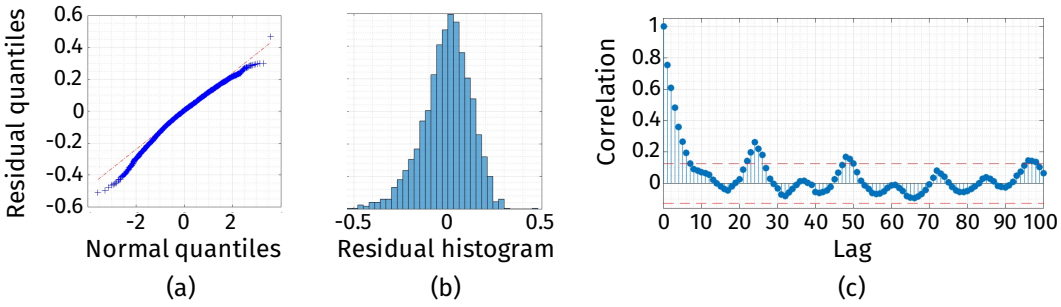


Figure 4.6: Analyses of wafer-level residuals for λ_r.

We provide two analyses of the wafer-level residuals for λ_r. First, we compare the distribution of the residuals to a standard normal distribution using a quantile-quantile plot (QQ-plot). As shown in Fig. 4.6a, the quantiles of the wafer-level residuals fall below a normal distribution at both high and low ends. This can also be verified by the histogram of the residuals which has a heavier tail to the left (Fig. 4.6b). We further investigate the autocorrelation of each row/column of the wafer-level residuals. Fig. 4.6c plots the autocorrelation of a center row and the 95 %–confidence intervals for a white noise process. Major peaks can be observed at every period of 24. For columns, a period of 2 is also observed. These findings indicate that a repetitive pattern with the size of a die (2×24) exists across the wafer.

Wafer-level decomposition is conducted for variations of ER and Q as well, and the existence of repetitive variation patterns is also confirmed among the wafer-level residuals. We summarize the wafer-level decomposition results for λ_r, ER, and Q in Table 4.1 for later

revisit. Next, we demonstrate the decomposition of intra-die variation components.

4.3.2 Intra-Die Variation Components

In this chapter, we employ a spatial-frequency-domain analysis proposed in [108] to extract the intra-die variation components. In contrast to [106] which simply takes an average of all dies as the intra-die variation pattern, this approach leverages the periodicity observed from the wafer-level residuals. Specifically, we convert the wafer-level residuals into the frequency domain using a 2-dimensional fast Fourier transform (FFT). The transformed matrix has the same size as the input matrix, i.e., 16×240. Theoretically, a periodic pattern in the spatial domain with the size of 2×24 will correspond to frequency-domain elements at every $16/2$ and $240/24$ intervals in the y- and x-directions. Therefore, the intra-die variation pattern can be extracted by first downsampling the frequency-domain matrix at these intervals, followed by an inverse FFT.

In Fig. 4.7, we show the intra-die variation patterns extracted for λ_r and the analyses of the residuals (hereinafter the *intra-die residuals*). Visually, the high end of the QQ-plot is closer to a standard normal distribution than that of the wafer-level residuals. The peaks in the autocorrelation plot are also reduced and mostly fall within the 95 %–confidence intervals for a white noise process.

In Fig. 4.8a, we plot the wafer-level residuals of Q for all microrings vs. their locations

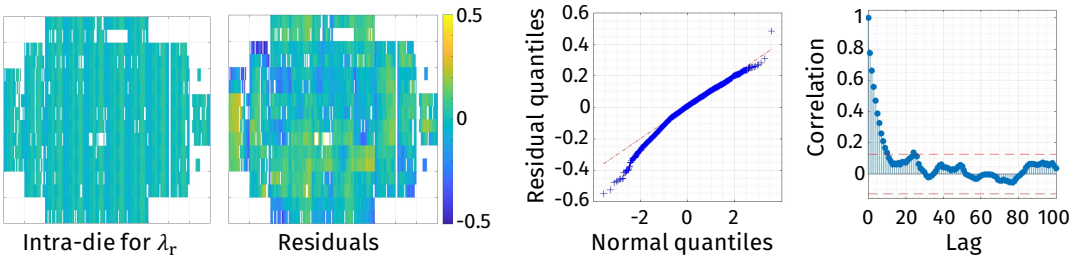


Figure 4.7: Intra-die variation patterns and residual analyses for λ_r.

within a die. The two rows in a die correspond to a Tx and an Rx. We also plot the extracted intra-die variation patterns in thick lines. We observe that the extracted patterns closely follow the general trends of the data groups. The differences between the Tx and Rx rows are also captured by the intra-die variation components (Fig. 4.8b), accounting for the alternating patterns mentioned in Section 4.3.1. The significance of the intra-die variation components to the total variations are also summarized in Table 4.1 for λ_r, ER, and Q.

4.3.3 Inter-Die Variation Components

Discontinuities at die borders can still be observed from the intra-die residuals, which is more likely due to systematic than random cause. We propose to further extract inter-die variation components from the intra-die residuals. Previous methods such as the one proposed in [110] are based on the availability of multiple wafers. In this chapter, however, we present a method based on low-rank tensor factorization for inter-die variation characterization.

Tensor factorization is a generalization of matrix factorization, which decomposes a tensor into another tensor of a smaller size and several factor matrices [237]. Existing applications of tensor factorization, however, have been focusing on data compression or accurate reconstruction of data with missing entries. Algorithms developed for these purposes usu-

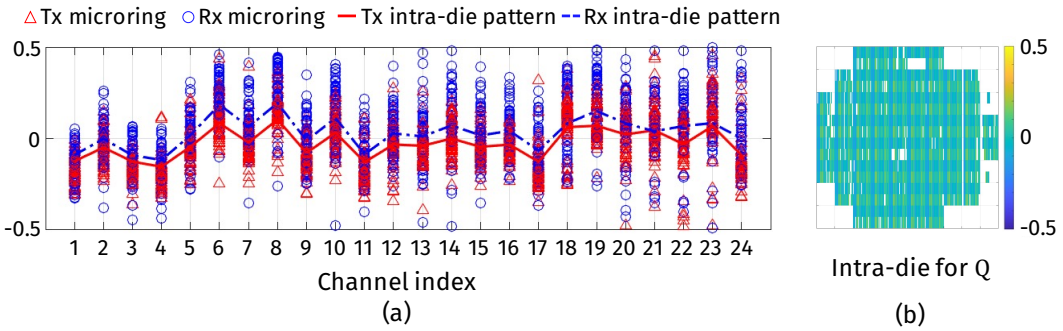


Figure 4.8: Intra-die variation patterns for Q.

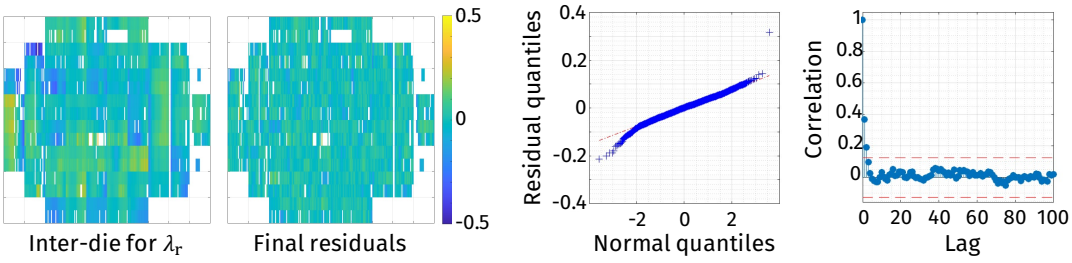


Figure 4.9: Inter-die variation characterization and final residual analyses for λ_r.

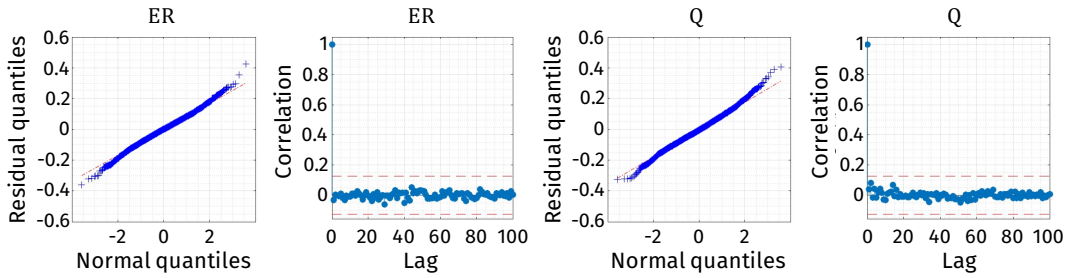


Figure 4.10: Final residual analyses for ER and Q.

ally leverage the low-rank properties of the data [238, 239, 240, 241], but cannot guarantee their effectiveness if the data are inherently not low-rank. For extracting the inter-die variation patterns, on the other hand, we aim to learn the hidden low-rank factors of the data disregarding the noise. In this context, we force a very low maximum rank to be explored by the algorithm in order not to capture the noise into the systematic variation components.

We first reshape the intra-die residuals into a tensor by splitting the wafer map at die borders and stacking all dies in a third dimension. Then, we apply tensor factorization to approximate the tensor with a low-rank version. This low-rank tensor is reshaped back to a matrix as the inter-die variation pattern, and the approximation errors are the final residuals of our whole decomposition workflow. In practice, we find that several existing tensor factorization algorithms all work considerably well with a maximum rank as low as 5. Fig. 4.9 shows the inter-die variation component for λ_r extracted by smooth parallel factor (PARAFAC) decomposition [240] and the final residual analyses. The linearity of the QQ-

Table 4.1: Summary of spatial variation decomposition.

Test item	Wafer-level		Intra-die	Inter-die	Residual	Residual distribution
	Leveling	Radial				
λ_r	43.03 %	16.61 %	5.12 %	27.45 %	7.79 %	$\mathcal{N}(0, 0.0407)$
ER	0.04 %	1.60 %	32.67 %	18.61 %	47.08 %	$\mathcal{N}(0, 0.0891)$
Q	1.59 %	2.31 %	29.47 %	21.72 %	44.91 %	$\mathcal{N}(0, 0.0940)$

plot indicates the resemblance of the residuals to a standard normal distribution within the range of ±2 standard deviations. The flat autocorrelation curve also confirms the successful extraction of periodic variation components. The residual analyses for ER and Q (Fig. 4.10) reach the same conclusion.

In Table 4.1, we summarize the hierarchical variation components extracted for λ_r, ER, and Q. We observe that the resonance wavelengths of the microrings are more prone to wafer-level variations, while ER and Q present more significant intra-die (layout-dependent) patterns. Considerable amounts of inter-die variations exist for all three test items. Overall, the variations of ER and Q appear more random than that of λ_r.

4.4 Applications of the Variation Model

In addition to the implications of variation sources, we further explore the applications of our characterized variation model for credible synthetic data generation and accurate yield prediction for microring-based optical transceivers.

4.4.1 Synthetic Data Generation

Synthetic data are often used to complement measurement data for the development and validation of link- and system-level variation alleviation techniques [105,225,227]. However, inaccurate synthetic data generated by oversimplified variation models may result in

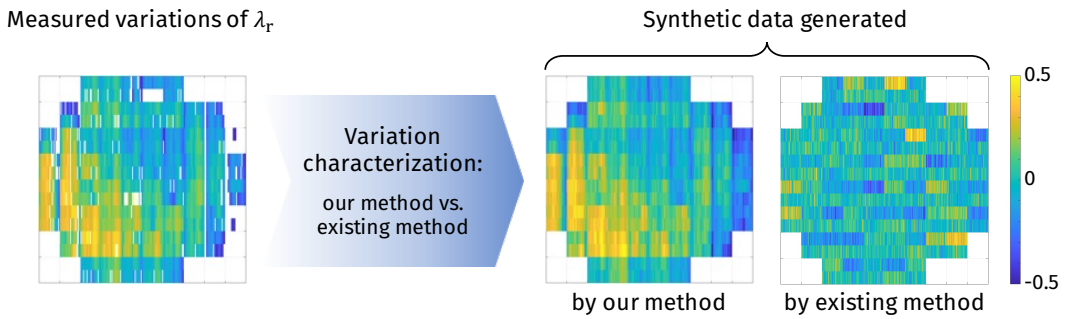


Figure 4.11: Comparison of the synthetic data generated by our method and the existing method from [105].

the misestimation of the system performance metrics. We generate synthetic wafer maps for the variations of λ_r, ER, and Q by adding a random term to the characterized systematic components. The random terms follow the normal distributions specified in Table 4.1. As a comparison, we characterize the same measurement data using the method featured in [105] and generate another set of synthetic wafer maps. In [105], the variations of the microrings are decomposed into global, local, and Tx-Rx offset terms. All three terms are directly approximated by independent normal distributions without characterizing their location dependencies. Fig. 4.11 compares the synthetic data for λ_r generated by the two methods. Visually, our synthetic data can better capture the spatial patterns of the measurement data.

We conduct two experiments to further evaluate the credibility of our synthetic data, namely to simulate 1) the microring tuning power for variation rectification and 2) the energy efficiency of the transceivers. The microring tuning power is the power required to thermally shift the resonance wavelengths of the Tx and Rx toward a common set of laser wavelengths, and is related to the variations of λ_r. The tuning efficiency assumed in the experiments is 0.15 nm/mW [242]. Fig. 4.12a plots the average tuning power per channel in ascending order computed for all measured and synthetic transceivers. Ten synthetic wafers are generated using either our method or the method from [105]. The results clearly

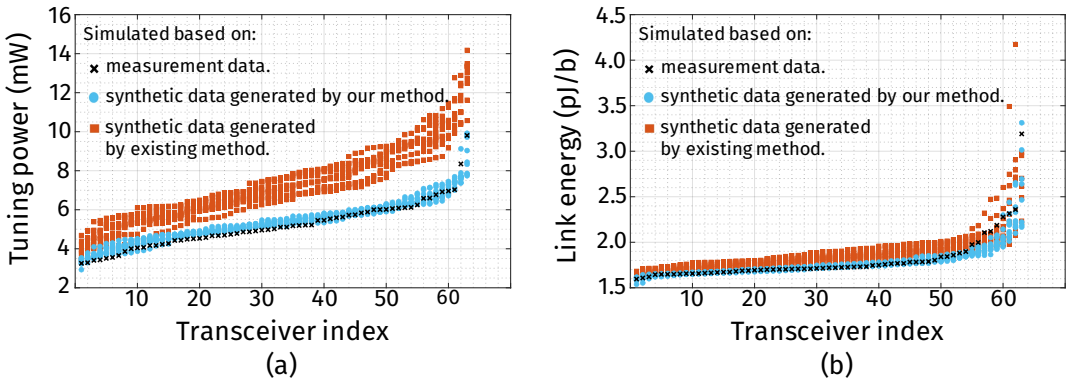


Figure 4.12: Simulations of tuning power and link energy efficiency show better quality of our synthetic data compared to that of [105].

indicate that our synthetic data closely resemble the measurement data and can estimate the tuning power with substantially higher accuracy than that of [105].

The simulation of the transceiver energy efficiency involves variations of λ_r, ER, and Q. It is computed as the total power consumption divided by the data rate. Apart from the microring tuning power, the power consumption of electrical driving circuitry has been modeled in [230] as a function of the data rate. As for the laser power, it has to be sufficiently high to overcome the various power losses within the TRx link and eventually satisfy the receiver sensitivity requirement:

$$P_{\text{laser}} \cdot \text{WPE} \cdot \prod_i \text{PL}_i \geqslant P_{\text{sensitivity}}. \tag{4.13}$$

Here, the wall-plug efficiency (WPE) is the efficiency with which the laser converts the electrical power into optical power. The various power losses, i.e., PL_i in Eq. (4.13), are determined by the variations of ER and Q [243]. Finally, by specifying a target data rate, the communication energy per bit can be computed for each transceiver with distinct variation profiles. Fig. 4.12b shows the simulated energy per bit for measured and synthetic transceivers at a target data rate of 20 Gb/s per channel and a laser WPE of 20 %. It is further

confirmed that our variation characterization method generates more realistic synthetic data than that of [105]. With credible synthetic data enabling variation-aware analyses of link power budget, novel techniques for variation alleviation can thus be evaluated with improved accuracy.

4.4.2 Yield Prediction

Our hierarchical spatial variation model can also be learned from incomplete wafer maps. We further explore the opportunity for transceiver yield prediction based on a partially measured wafer. We define a *working* transceiver as one that can achieve a bit error rate (BER) of 10^{-12} at a given data rate. We then define the yield as the number of working transceivers on a wafer divided by the total number of transceivers. The authors of [243] model the receiver sensitivity requirement for a BER of 10^{-12} as a function of the target data rate. Due to the optical nonlinearities of the microrings and the silicon waveguides, there exists a maximum optical power that can be injected into the transceiver [244]. Based on Eq. (4.13), a maximum optical power at the receiver end can be computed for each transceiver with distinct variation profiles. If this maximum power at the Rx end is still lower than the sensitivity required for a BER of 10^{-12}, the transceiver is considered *not working* at this data rate.

In the experiments, we characterize the spatial variation model based on the measurement data of randomly sampled transceivers and predict a complete wafer map (Fig. 4.13a). Yield is computed based on this predicted wafer map and compared to that of a fully measured wafer. In Fig. 4.13b, we plot the predicted yield for various sampling ratios and target data rates. The dashed lines correspond to the yield computed from full measurement data. As can be seen, a sampling ratio too low tends to produce over-optimistic predictions of the yield, which indicates that the severity of the variations is underestimated. However, a sam-

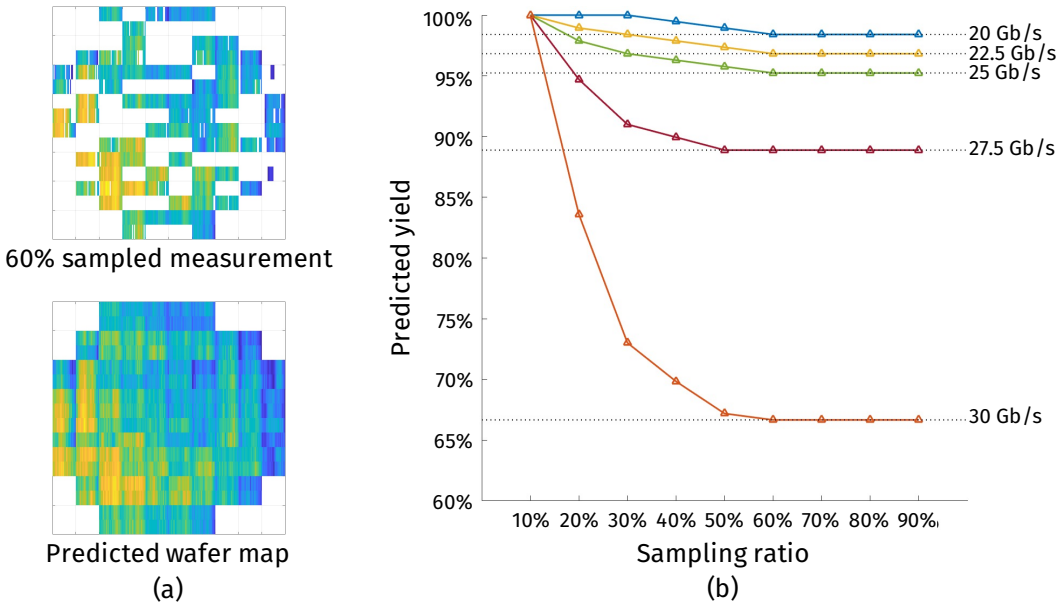


Figure 4.13: Predicted yield at various sampling ratio.

pling ratio of 60 % is high enough for the characterized variation model to accurately predict the yield of the whole wafer. Our variation characterization method can thus be used to investigate the measurement order of the transceivers in order to provide an early estimation of the wafer yield.

4.5 Concluding Remarks

In this chapter, we present a hierarchical method for characterizing the spatial variations of microring-based optical transceivers. The extracted wafer-level, intra-die, and inter-die variation components effectively capture the location dependencies of the variations. We demonstrate the applications of the characterized variation model for credible synthetic data generation and accurate yield prediction. If more fabricated wafers become available in the future, topics that are worth exploring include calibration of the spatial variation model, characterization of wafer-to-wafer temporal variations, and broad applications of

the variation model in system-level variation-aware design optimization for optical interconnects.

Part III

Link-Level Variation Alleviation

Chapter 5

Redundant Laser Comb Lines for Microring-Based Optical Transceivers

The comb laser–driven microring-based dense wavelength-division multiplexing (DWDM) silicon photonics is one of the promising candidates for next-generation high-bandwidth energy-efficient optical interconnects. However, existing solutions for exploring the power-performance trade-off of such systems have been restricted to a limited design space and a naïve channel alignment scheme. These restrictions result from the unnecessary assumptions of using identical spacing for laser comb lines and microring channels and utilizing consecutive laser comb lines for data transmission. In this chapter, we propose an energy-efficient channel alignment scheme that aligns the microring channels to a subset of laser comb lines that are non-uniformly distributed in the free spectrum range (FSR) of the microrings. The proposed scheme takes into account the device variations in fabricated components and selects the channel wavelengths for alignment with greater flexibility. Based on a well-established process variation model, our simulations show that the proposed scheme significantly reduce the microring tuning power in the presence of denser comb lines. The power saved from microring tuning has the potential to improve the overall link energy efficiency despite some power wasted in unused laser comb lines. Thus, the laser comb spacing can be added as an additional dimension for exploration in the design space of DWDM sil-

icon photonics. We further conduct a case study for design space exploration (DSE) using the proposed channel alignment scheme, seeking the most energy-efficient configuration in order to achieve a target aggregated data rate.

5.1 Introduction

The continuing traffic growth in data centers and high-performance computing (HPC) systems calls for the next generation of high-throughput, low-latency, and energy-efficient optical interconnects [1]. DWDM silicon photonics has been proposed as a cost-effective and scalable solution to the above applications by taking advantage of large-scale CMOS-compatible integration [13]. A promising approach combines innovations in quantum-dot (QD) comb lasers and silicon photonic microring resonators (MRRs) to achieve concurrent multi-channel DWDM [228, 229]. A multi-wavelength QD comb laser [21] is preferred to an array of single-wavelength lasers [19] due to its ease of temperature control, wavelength tracking and packaging at large scale [245]. Cascaded microring modulators and filters are one of the promising candidates for short-reach DWDM solutions due to their compact footprints, low power consumption, and (de)multiplexer-free architecture [26, 27, 28].

In microring-based optical links, the microring radii are typically designed to provide a set of discrete resonance wavelengths with a constant channel spacing [26, 27, 28, 228, 229]. Due to process variations, the resonance wavelengths of the fabricated microrings can deviate from the design values for up to ±10 nm across a wafer [173]. As a result, the resonance wavelengths of the microrings in the transmitter (Tx) and the receiver (Rx) need to be actively tuned to align with the carrier wavelengths provided by the laser comb lines.

The microring tuning power has been identified as one of the greatest contributors toward the link power consumption, thus the priority of further optimization [174, 230]. The authors of [172] provide device-level solutions to low-power wavelength tuning. However,

the unoptimized quality factor (Q) of the device design hinders it from immediate DWDM applications. Techniques such as channel shuffling/remapping and sub-channel redundant microrings [175, 176] are proposed to reduce the average tuning power of a multi-channel microring-based optical link. However, these studies are based on oversimplified assumptions, such as allocating up to 3 channels per transceiver (TRx), and lack validation from actual measurement data. The solution proposed in [105] mitigates the average tuning power of microring-based transceivers by optimally assigning Tx-Rx pairs among a number of fabricated devices. However, it only ensures the channel alignment between the Tx and the Rx, but fails to account for the misalignment of the TRx channels and the laser comb lines. Moreover, all of the mentioned techniques treat microring tuning as a standalone problem; none have studied the link-level implications of the proposed strategies due to inter-component interactions within the optical link.

The laser comb lines, as the provider of the carrier wavelengths, are an indivisible part of the microring tuning problem. The inclusion of laser comb lines into the design space would intrigue the need for addressing the allocation and alignment schemes of the microring channels and the carrier wavelengths. Recent studies on design space exploration of DWDM silicon photonics assume comb lasers to provide just enough laser comb lines for the on-chip microring channels [243, 246]. The spacing of the laser comb lines, on the other hand, is often left out of the design space. Since the microring channels are designed to be equally spaced in frequency, it seems intuitive that to utilize a comb laser with the same spacing as the microrings would statistically lead to the minimum expected tuning power [173], provided that the laser comb lines used for data transmission are consecutive (hereinafter referred to as the *consecutive channel alignment scheme*). However, our simulations based on a well-established process variation model for multi-channel microring-based transceivers show that, in the presence of denser comb lines, a more flexible channel alignment scheme where the usage of consecutive laser comb lines are not mandatory may

further reduce the power consumption of microring tuning. Our proposed scheme aligns the microring channels to a subset of laser comb lines that are non-uniformly distributed in the free spectrum range of the microrings (hereinafter referred to as the *non-uniform channel alignment scheme*). The simulations reveal that, despite some power wasted in unused laser comb lines, the overall link energy efficiency can potentially benefit from the reduction of the microring tuning power. Our non-uniform channel alignment scheme therefore expands the design space of DWDM silicon photonics to incorporate the laser comb spacing as one of the design parameters to be explored.

The rest of this chapter is organized as follows. Section 5.2 overviews the architecture of the target DWDM silicon photonic transceiver and presents the process variation model used in synthetic data generation. Section 5.3 introduces the power model of the transceiver link as a function of the link configuration and discusses typical design considerations for the QD comb laser. In Section 5.4, the mechanism of the proposed non-uniform channel alignment scheme is illustrated. Simulation results are also provided to show that the proposed scheme allows the utilization of denser comb lines without compromising the link energy efficiency. In Section 5.5, a case study for design space exploration of DWDM silicon photonic transceivers is conducted using our non-uniform channel alignment scheme, seeking the most energy-efficient configuration in order to achieve a target aggregated data rate. Finally, Section 5.6 draws the conclusion of this chapter.

5.2 Overview and Data Preparation

5.2.1 Measurement Data of Microring-Based Transceivers

In order to model the process variations of multi-channel microring-based transceivers, the resonance wavelengths of a batch of fabricated devices are extracted from the mea-

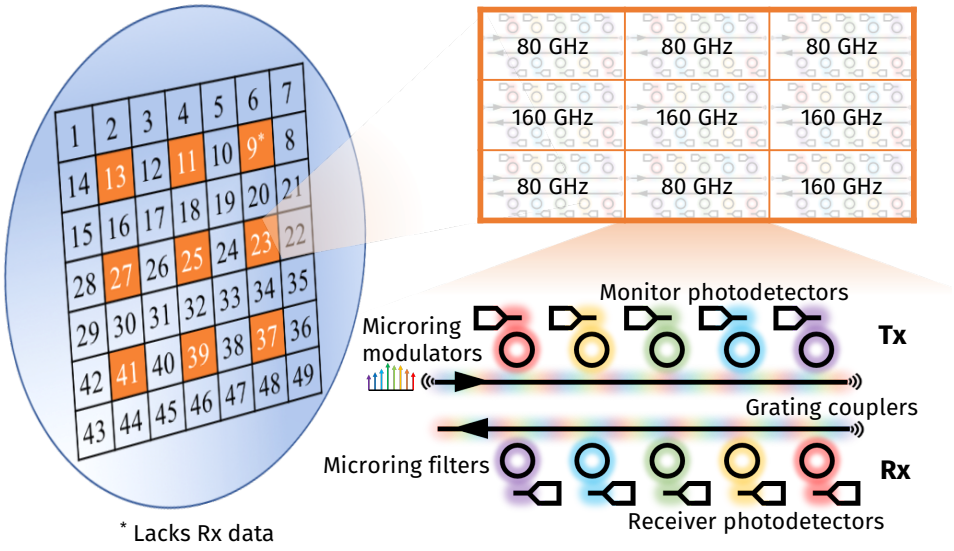


Figure 5.1: Illustration of the fabricated transceivers from which the process variation model for resonance wavelengths is extracted.

surement data. The fabricated wafer consists of 49 dies, of which nine representative ones (highlighted in Fig. 5.1) are measured. Each die consists of nine transceivers, where five are designed with 80 GHz channel spacing and four with 160 GHz. Each TRx is designed with five high-speed microring modulators on the Tx side for on-off keying (OOK) modulation and five corresponding microring filters on the Rx side for demultiplexing. The microring radii are designed to be ~ 10 µm, and the FSR is measured to be ~ 13.9 nm, which can accommodate up to 30 channels at 80 GHz spacing or 15 channels at 160 GHz spacing near 1310 nm. Due to the lack of Rx measurement data on Die #9 and the absence of clear resonance dips in one of the 160 GHz transceivers, 40 out of 45 transceivers with 80 GHz channel spacing and 31 out of 36 transceivers with 160 GHz channel spacing are used for resonance wavelength extraction. An example of the measured optical spectrum of the transceivers is shown in Fig. 5.2a, where variations exhibit in the resonance wavelengths of the microrings.

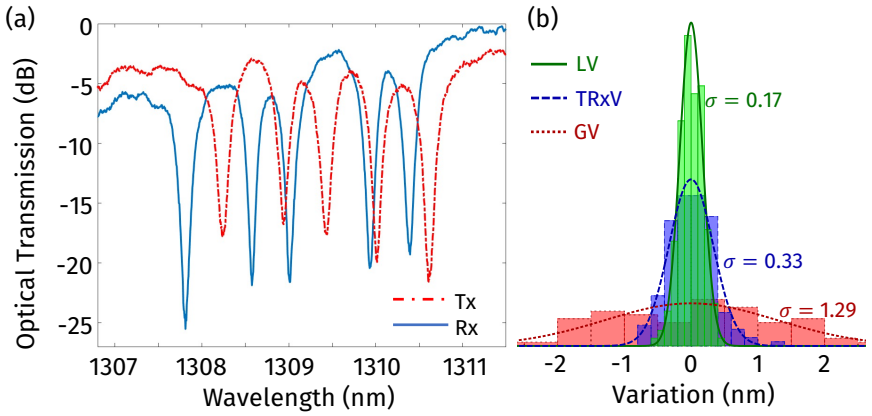


Figure 5.2: (a) Measured optical spectrum of a five-channel transceiver with 80 GHz channel spacing, and (b) distributions of the variation components and the fitted Gaussian curves for transceivers with 80 GHz channel spacing.

5.2.2 Variation Modeling and Synthetic Data Generation

The process variation model for resonance wavelengths featured in this chapter is based on the one proposed in [105], where the resonance wavelength of each microring is decomposed into the design value and several independent variation components. However, the study conducted in [105] converted the channel spacing of the microrings from the frequency domain to the wavelength domain (80 GHz ~ 0.46 nm in the 1310 nm regime). This approximation is acceptable for small channel count (five in [105]), but the error accumulates as the number of channels increases. A modified version of the process variation model is used in this chapter. Specifically, the jth channel of the ith transceiver has the resonance wavelength $\lambda_{\mathrm{Tx/Rx}}(i,j)$ expressed as Eq. (5.1), where i is the transceiver index, j is the channel index, λ_0 is the design value of the first channel, Δf is the channel spacing in the frequency domain, and c is the speed of light. The global variation term (GV) and the Tx-Rx offset term (TRxV) are shared by all microrings within one transceiver, while the local

variation term (LV) is independent for each microring:

$$\begin{aligned}\lambda_{\mathrm{Tx}}(i,j) &= \frac{c}{\left(c/\lambda_0-(j-1)\,\Delta f\right)}+\mathrm{GV}_i+\mathrm{LV}_j,\\ \lambda_{\mathrm{Rx}}(i,j') &= \underbrace{\frac{c}{\left(c/\lambda_0-(j'-1)\,\Delta f\right)}}_{\text{Design value}}+\underbrace{\mathrm{GV}_i+\mathrm{LV}_{j'}+\mathrm{TRxV}_i}_{\text{Variation components}}.\end{aligned} \tag{5.1}$$

The results of the resonance wavelength extraction show that all of the three variation components follow normal distributions with zero mean, as shown in Fig. 5.2b. The developed process variation model is then used to generate synthetic data of microring-based transceivers for validating the study conducted in this chapter.

5.3 Power Models and Assumptions

5.3.1 Power Model of the Transceiver Link

The power consumption of a comb laser–driven microring-based optical transceiver can be broadly decomposed into the laser power, the microring tuning power, and the dynamic power consumption of the driving circuitry, namely the modulator drivers and the receiver transimpedance amplifier (TIA). In order to achieve a low bit error rate (BER), the optical power budget provided by the comb laser must be sufficiently high enough to overcome the power penalties along the transceiver link and eventually meet the sensitivity requirement of the receiver. The authors of [246] have proposed a fairly comprehensive model for various power penalties present in a microring-based DWDM optical transceiver. The power penalties can be data rate–dependent (e.g., the receiver sensitivity), channel spacing–dependent (e.g., the microring crosstalk), channel count–dependent (e.g., the insertion loss of the microring array), in addition to some constant terms (e.g., losses of the silicon waveguides and grating couplers), and thus a function of the link configuration. For the study featured

in this chapter, lookup tables of power penalties are made for different channel spacings, channel counts, and per-channel data rates, based on the values reported in [246]. The minimum required laser power is obtained by adding up the power penalties of the optical link and the sensitivity of the receiver. The minimum power budget of the laser is computed as 1.03 mW per comb line at 10 Gb/s channel rate, and 2.30 mW per comb line at 25 Gb/s channel rate, which is consistent with both actual measurements [229, 245] and technology projections [21, 247]. The power consumption of the modulator driver and the receiver TIA at different data rates are sampled from [243]. The microring tuning power, as one of the optimization targets in this chapter, is detailed in Section 5.4.

5.3.2 Design Considerations for the Comb Laser

The multi-wavelength emission capability of the QD comb laser renders it popular for DWDM silicon photonic applications. Four important characteristics of the QD comb laser will interact with the design of microring-based DWDM silicon photonic transceivers: 1) the spacing of the laser comb lines; 2) the spectrum range of the laser optical output; 3) the wall-plug efficiency (WPE) of the comb laser; and 4) the maximum optical power supported by each comb line. In this section, typical design considerations regarding these characteristics of comb lasers are discussed, and corresponding assumptions are made for simulation purpose.

The spacing and the spectrum range of the laser comb lines together determine the carrier wavelengths for the transceiver channels to align to. WPE refers to the efficiency with which the laser diode converts electrical power into optical power. The choice of the laser comb spacing is a result of the optimization of the laser design, especially in terms of WPE. In practice, a cavity length of 500–1000 µm (40–80 GHz comb spacing) usually leads to the optimum, depending on the actual design of the device epi-structure [245]. The spectrum

range of the comb laser must demonstrate little relative intensity noise (RIN) of each longitudinal mode to achieve a low BER in data transmission [20]. Under certain bias and temperature, the low-noise operating range falls into the 1310 nm (O-band) or the 1550 nm (C-band) regime, which is suitable for fiber optics. In other words, once the QD comb laser is stabilized at its optimal operating point, further tuning of the comb lines to better align with the transceiver channels is undesirable for avoiding the introduction of excessive noise into the operating band. Therefore, in this chapter, the laser comb lines are assumed to be fixed at their designed wavelengths in each simulation, and the microring channels are unilaterally tuned toward the carrier wavelengths determined by the laser comb lines.

5.4 Channel Alignment Schemes

The refractive index of the silicon microring changes with temperature due to thermal-optical (TO) effects and causes the resonance wavelength to drift [88]. On-chip resistive heaters are fabricated inside the silicon microrings to thermally tune the resonance wavelengths toward the allocated channels. Due to the nature of thermal tuning, the spectrum of the microrings can only be shifted in the direction of longer wavelengths (redshift). Fig. 5.3a illustrates the mechanism of the commonly used consecutive channel alignment scheme, where the circles represent the resonance wavelengths of the microrings as fabricated, and the dashed lines represent the allocated channels. The group of consecutive channels are selected such that the total tuning distance of the microrings is minimized, and each resonance wavelength is only redshifted. Under the consecutive channel alignment scheme, the utilization of laser comb lines that are denser than the microring channels (as shown in Fig. 5.3b) would pile the allocated channels to the right in order to satisfy both the consecutive channel allocation and the redshift requirement, thus causing the tuning distance to increase. As a result, the consecutive channel alignment scheme by default assumes a comb

laser with the same channel spacing as the microrings to minimize the expected tuning distance.

Restrictions on the design space induced by the consecutive channel alignment scheme hinder the exploration of more energy-efficient design of DWDM silicon photonics. In order to eliminate the need for the identical spacing of laser comb lines and microring channels, we propose a non-uniform channel alignment scheme where the usage of consecutive laser comb lines are not mandatory. As illustrated in Fig. 5.3c, the proposed scheme aligns each pair of microrings in the Tx and the Rx to the next available laser comb line to their right. The result is a non-uniform distribution of the channels which allows the existence of unused

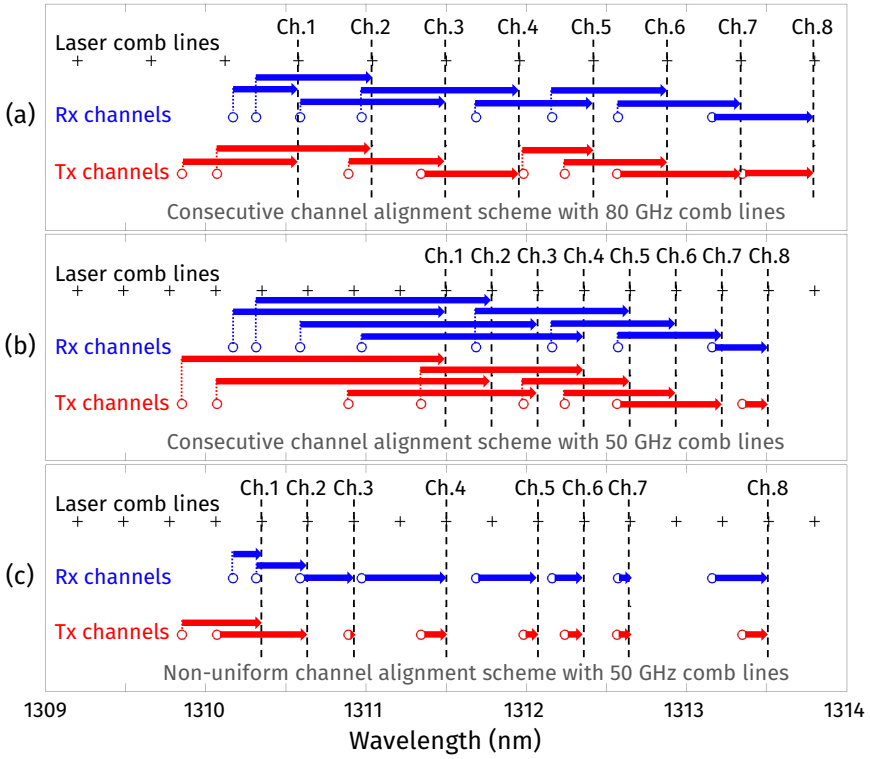


Figure 5.3: Illustration of the tuning distance computed for (a) the consecutive channel alignment scheme with 80 GHz–spaced laser comb lines, (b) the consecutive channel alignment scheme with 50 GHz–spaced laser comb lines, and (c) the proposed non-uniform channel alignment scheme with 50 GHz–spaced laser comb lines. The microring channels are of 80 GHz spacing with variations added on top.

laser comb lines.

Fig. 5.4 illustrates the average tuning distance required for each microring at various channel counts using the traditional consecutive channel alignment scheme and our proposed non-uniform channel alignment scheme. Synthetic data of 1000 transceivers with 80 GHz channel spacing are generated for each channel count configuration to model the process variations. The simulation results show that, for microring-based transceivers designed with 80 GHz channel spacing, our proposed non-uniform channel alignment scheme with 40–80 GHz–spaced laser comb lines always reduces the average tuning distance compared to the consecutive channel alignment scheme. It is also observed that the average tuning distance per microring increases with the number of channels in the presence of an 80 GHz–spaced comb laser, owing to the redshift-only tuning mechanism. This channel count dependency of the tuning distance can be eliminated by using our non-uniform channel alignment scheme with denser comb lines. The availability of extra comb lines between channels buffers the variations of the resonance wavelength of individual microrings, and mitigates their influence on the channel selection of neighboring microrings.

Our non-uniform channel alignment scheme enables greater flexibility in channel allo-

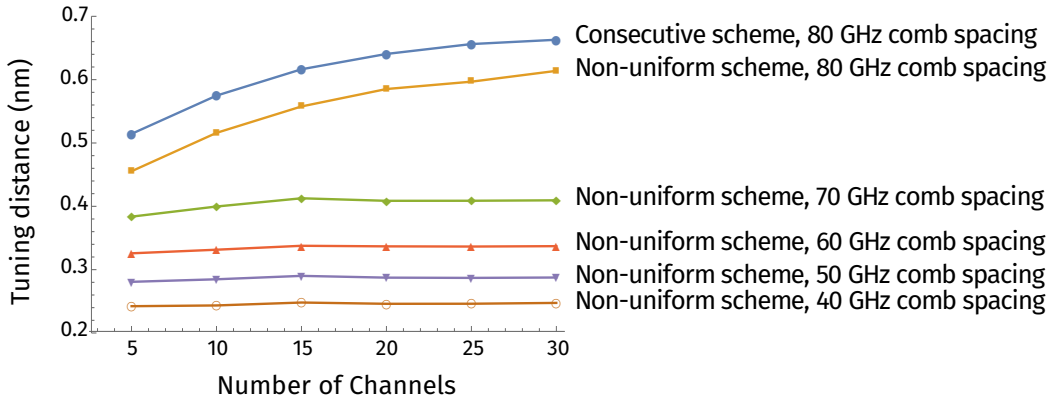


Figure 5.4: Average tuning distance per microring for various channel counts using the traditional consecutive channel alignment scheme and our proposed non-uniform channel alignment scheme.

cation and reduces the required tuning distance of the microrings at the cost of some power wasted in unused laser comb lines. An interesting trade-off therefore arises in terms of the overall energy efficiency of the DWDM optical link. A case study of the link energy efficiency w.r.t. different channel counts and laser comb spacings is conducted to demonstrate the trade-off introduced by the proposed channel alignment scheme. The average tuning power per TRx is computed assuming 0.15 nm/mW tuning efficiency [242]. Due to the limited range of the laser comb spacing (40–80 GHz), the microring channel spacing is set at 80 GHz in order to compare with the consecutive channel alignment scheme. The power consumption of each component in the optical link is computed at a 10 Gb/s per-channel data rate. Finally, the overall energy efficiency of the link is computed as the total power consumption over the aggregated data rate for each configuration. The energy efficiency is measured in pJ/b, for which the smaller the value the better.

Fig. 5.5 shows the saving of the overall energy per bit of the link using our proposed channel alignment scheme compared to using the consecutive one. It is noteworthy that the proposed scheme with 80 GHz comb spacing is always beneficial to the link energy efficiency regardless of the channel count, as no extra laser comb lines are introduced. With denser comb lines, both the laser comb spacing and the number of channels play their parts in the aforementioned trade-off. For a larger channel count (and thus a higher aggregated

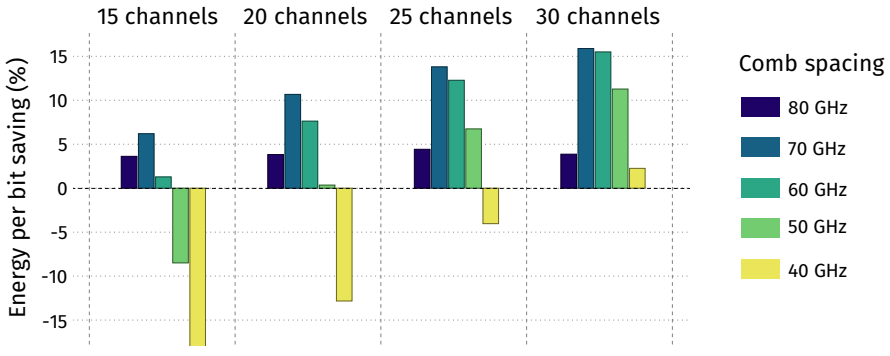


Figure 5.5: Saving of the link energy per bit under different channel counts using the proposed channel alignment scheme compared to using the consecutive scheme.

data rate), the energy per bit contributed by the laser is reduced, and denser comb lines can be utilized without compromising the overall link energy efficiency. The biggest energy per bit saving, however, does not require aggressively dense comb lines. In this particular example, a comb laser with 70 GHz comb spacing best exploits the proposed scheme in all channel count configurations for microrings designed with 80 GHz channel spacing.

5.5 Design Space Exploration

Adding the laser comb spacing as a design parameter to be explored, our proposed non-uniform channel alignment scheme expands the design space of DWDM silicon photonic transceivers by an additional dimension. A case study for design space exploration is conducted in this section, seeking the most energy-efficient link configuration to achieve a target aggregated data rate. As the microrings modeled in this chapter are based on fabricated devices with a moderate channel spacing of 80 GHz and a 13.9 nm FSR, the purpose of the exploration is not to push the limit of the highest attainable data rate, but rather to observe the pattern of the optimal design w.r.t. various link configurations and provide useful guidelines to the designers of DWDM silicon photonics. The laser comb spacing and the channel count of the transceiver are two of the major design parameters explored in this chapter. Once the channel count is decided, the per-channel data rate is calculated as the target aggregated data rate over the number of channels. The microring channel spacing is fixed at 80 GHz in accordance to the developed process variation model. In addition to the laser comb spacing and the channel count, this chapter also takes into consideration the projected future advances in microring and comb laser designs by varying the microring tuning efficiency and the laser WPE.

Simulations are conducted for various aggregated data rate ranging from 50 Gb/s to 300 Gb/s, and the exploration results for 100 Gb/s and 200 Gb/s are summarized in Fig. 5.6

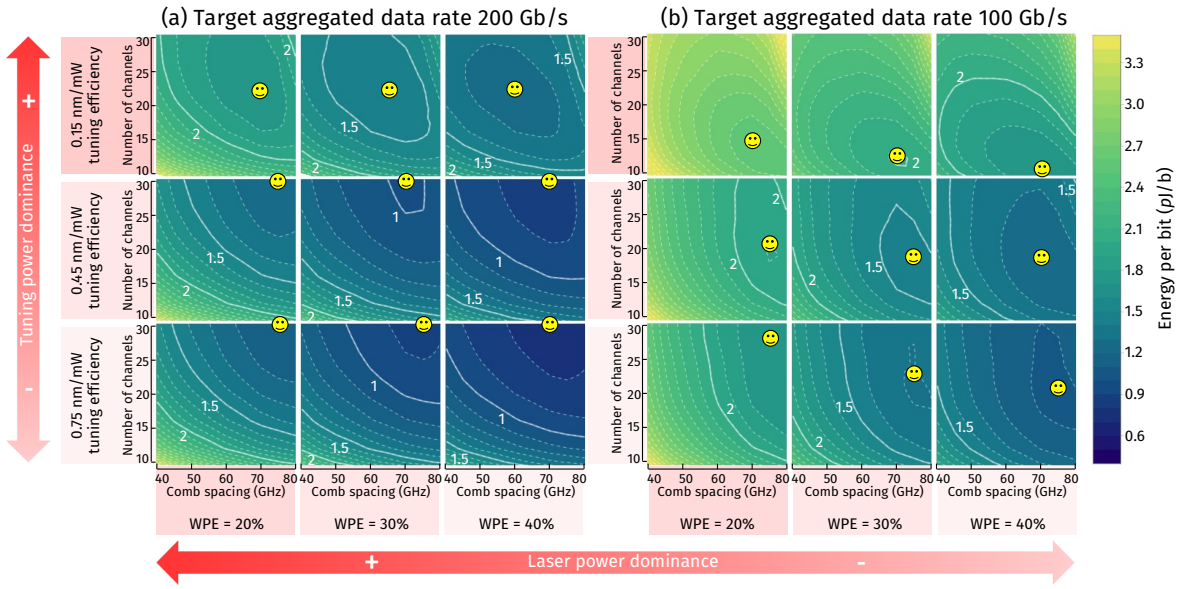


Figure 5.6: Design space exploration using our non-uniform channel alignment scheme with a target aggregated data rate of (a) 200 Gb/s, and (b) 100 Gb/s. The smiley sign in each subplot stands for the configuration that leads to the lowest energy per bit value.

to demonstrate the pattern of the optimal link configurations. The overall energy per bit of the TRx link is illustrated with contour plots color coded from yellow (highest pJ/b) to blue (lowest pJ/b). For each combination of the tuning efficiency and the laser WPE, the contours of the link energy per bit are drawn for different laser comb spacings and TRx channel counts. The link configuration that leads to the lowest energy per bit value is marked with a smiley sign in each contour plot.

The first observation made from the plots is that the optimal designs in all cases utilize laser comb lines that are denser than the microring channels. The effectiveness of our proposed non-uniform channel alignment scheme is therefore validated. Our proposed scheme scales well with the target aggregated data rate of the TRx link as well as the projected future advances in the microring tuning efficiency and the laser WPE.

As the link configuration changes, the optimal design that leads to the lowest energy per bit occurs at different combinations of the laser comb spacing and the channel count, reflected by the coordinate of the smiley sign in each subplot. In order to provide general

guidelines to the designers of DWDM silicon photonic transceivers, we attribute the various factors affecting the optimal design to two generalized properties of the optical link, namely the *tuning power dominance* and the *laser power dominance*. A low tuning efficiency of the microrings is a contributor to the tuning power dominance of the link energy per bit, while a low WPE causes the link energy per bit to depend more heavily on the laser power. Increasing the target aggregated data rate requires a larger power budget provided by the comb laser, which also contributes to the laser power dominance of the link. As can be observed from Fig. 5.6, the reduction of tuning power dominance causes the optimal design to drift to the upper right, which encourages the usage of a higher channel count (lower date rate per channel) and a larger comb spacing. On the other hand, the reduction of laser power dominance renders the optimal design to drift down, which encourages a lower channel count in the TRx design. In short, the case study demonstrates how the addition of laser comb spacing into the design space can lead to a more energy-efficient design of DWDM optical transceivers by manipulating the trade-off between the tuning power and the laser power.

5.6 Concluding Remarks

In this chapter, we present a non-uniform channel alignment scheme enabled by sub-channel redundant comb lines for comb laser–driven microring-based optical transceivers. The proposed scheme reduces the average microring tuning power in the presence of laser comb lines that are denser than the microring channels. By exploring the trade-off between the tuning power and the laser power, our proposed scheme is capable of improving the overall energy efficiency of the transceiver compared to the traditional consecutive channel alignment scheme. The utilization of the non-uniform channel alignment scheme allows the laser comb spacing to be added to the design space of DWDM silicon photonic transceivers. A case study for design space exploration is conducted to investi-

gate the most energy-efficient design in order to achieve a target aggregated data rate. The results of the exploration validate the effectiveness and scalability of the proposed channel alignment scheme. Patterns of the optimal design under various system configurations are observed and summarized to guide the energy-efficient design of future DWDM silicon photonic transceivers.

Chapter 6

Bidirectional Tuning of Microring Resonance Wavelengths

Microring-based silicon photonic transceivers are promising to resolve the communication bottleneck of future high-performance computing (HPC) systems. To rectify the process variations in microring resonance wavelengths, thermal tuning is usually preferred over electrical tuning due to its preservation of extinction ratios and quality factors. However, the low energy efficiency of resistive thermal tuners results in nontrivial tuning cost and overall energy consumption of the transceiver. In this chapter, we propose a hybrid tuning strategy which involves both thermal and electrical tuning. Our strategy determines the tuning direction of each resonance wavelength with the goal of optimizing the transceiver energy efficiency without compromising signal integrity. Formulated as an integer programming problem and solved by a genetic algorithm, our tuning strategy yields 32–53 % savings of overall energy per bit for measurement data of 5-channel transceivers at 5–10 Gb/s per channel, and up to 24 % saving for synthetic data of 30-channel transceivers, generated based on the process variation model built upon the measurement data. We further investigate a polynomial-time approximation method which achieves over 100x speedup in tuning scheme computation, while still maintaining considerable energy-per-bit savings.

6.1 Introduction

With the benefits in throughput and energy efficiency brought by dense wavelength-division multiplexing (DWDM), optical interconnects have been proposed to accommodate traffic-intensive applications in future high-performance computing systems [1]. Silicon photonics is emerging as a cost-effective and scalable solution to optical interconnects by taking advantage of large-scale CMOS-compatible integration [13]. A promising architecture incorporates innovations in quantum-dot (QD) comb lasers [21] and silicon photonic microring resonators (MRRs) [25] to achieve concurrent DWDM [228, 229].

Microring resonators are highly wavelength-selective devices that can be used to modulate or filter optical signals at their resonance wavelengths [26, 27]. Due to process variations, the fabricated resonance wavelengths can deviate significantly from the design values [173] and thus require active post-fabrication tuning to align with the carrier wavelengths. For p-i-n junction–based, carrier-injection microring resonators [25], the electro-optical (EO) and thermal-optical (TO) effects shift the resonance wavelengths in opposite directions [88, 248], which provides such devices with inherent capability of bidirectional tuning. Despite better energy efficiency of electrical tuning as compared to thermal tuning, the increase of electron-hole pair density introduced into the cavity by electrical tuning results in the degradations of extinction ratios (ERs) and quality factors (Qs) of the transmission spectra [25, 88, 249], which makes electrical tuning less preferable. However, if the required tuning distance is small and thus the degradations of ER and Q are limited, such degradations can be compensated by the increase of laser power to maintain the transmission quality, if the overall power consumption of the transceiver (TRx) is still lower than that of thermal tuning. Nevertheless, such opportunities have not been explored in microring-based DWDM applications which are currently dominated by thermal tuning schemes [28, 250, 251, 252, 253].

One of the greatest challenges faced by thermal tuning schemes is the nontrivial power consumption [230]. Techniques based on channel remapping [173, 174] and sub-channel redundancies [175, 176, 225] are proposed to mitigate the expected tuning power. Wafer-level variations are also exploited in [105] to further reduce the average tuning power when a large number of transceivers are available. However, the inherently low energy efficiency of resistive thermal tuners limits the effectiveness of these techniques as long as all-thermal tuning schemes are adopted.

In this chapter, we present a hybrid tuning strategy for carrier-injection microring-based transceivers which involves both thermal and electrical tuning. Our strategy accounts for the potential degradations of ER and Q induced by electrical tuning and determines the tuning direction of each resonance wavelength with the goal of optimizing the overall energy efficiency of the transceiver link. We solve for the optimal tuning scheme using a genetic algorithm (GA) and a polynomial-time approximation method. Tuning schemes given by both methods are evaluated on measurement data from a batch of 5-channel transceivers, as well as synthetic data, generated based on the process variation model built from the measurement data, of 5–30-channel transceivers, showing considerable savings of the overall energy per bit.

The rest of this chapter is organized as follows. Section 6.2 overviews the architecture of the target multi-channel microring-based transceivers and the process variation model. Section 6.3 introduces the transceiver link power models and assumptions. In section 6.4, we formulate the problem of bidirectional tuning and adapt it to a genetic algorithm–based optimizer with linear constraints. Section 6.5 evaluates the proposed strategy on measured and synthetic data of multi-channel transceivers. In Section 6.6, a polynomial-time approximation method is introduced to speed up the tuning scheme computation. Finally, Section 6.7 draws the conclusions of this chapter.

6.2 Overview and Data Preparation

6.2.1 Measurement Data of Microring-Based Transceivers

Our bidirectional tuning strategy targets the optical transceivers made of p-i-n junction–based carrier-injection microring resonators. The transceivers that we used for variation characterization were fabricated by CEA-Leti [254] on a 200 mm silicon-on-insulator (SOI) wafer. As illustrated in Fig. 6.1, each fabricated transceiver consists of five high-speed microring modulators in the transmitter (Tx) for on-off keying (OOK) modulation, and five corresponding microring filters in the receiver (Rx) for demultiplexing. The channel spacing of the transceiver is designed to be 80 GHz in the 1310 nm regime (O-band). The free spectrum range (FSR) of the microring is measured as ~13.9 nm, which can accommodate up to 30 multiplexed channels. In total, 45 transceivers located on 9 dies were sampled for measurement, in which 5 transceivers showed no clear resonance dips, and another 6 transceivers failed in either Tx or Rx. Thus, 34 transceivers were successfully measured and characterized.

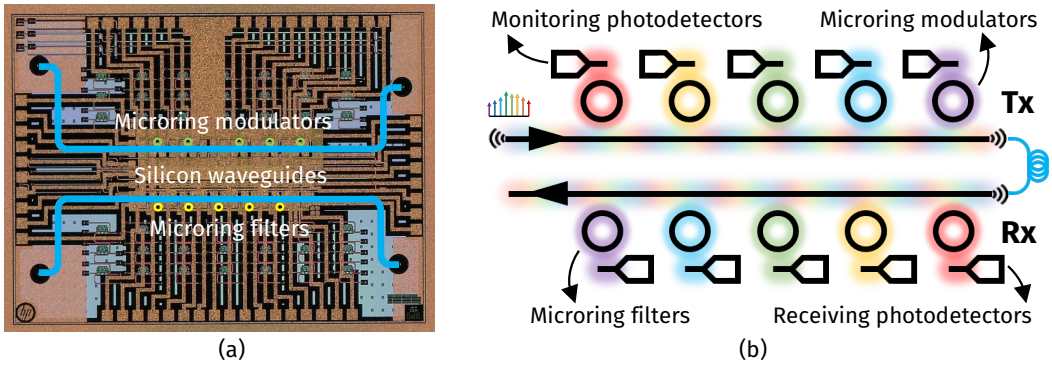


Figure 6.1: (a) Microscopic image and (b) architectural illustration of a 5-channel microring–based transceiver fabricated by CEA-Leti.

6.2.2 Variation Modeling and Synthetic Data Generation

The optical signal travels through the *through port* of each cascaded microring in the Tx, and is coupled out at the *drop port* of a microring in the Rx, before sensed by the receiving photodetector (PD). Process variations of the through/drop-port transmission spectra are characterized as follows.

Through-Port Spectrum

We model the through-port transmission spectrum of microring #i with a Lorentzian function–based equation, characterized by its resonance wavelength $\lambda_{r,i}$, extinction ratio $\mathrm{ER}_{\mathrm{thru},i}$, quality factor $\mathrm{Q}_{\mathrm{thru},i}$, and a through-port loss $\alpha_{\mathrm{thru},i}$ [88]:

$$T_{\mathrm{thru},i}(\lambda) = \alpha_{\mathrm{thru},i} \cdot \left(1 - \frac{1 - 1/\mathrm{ER}_{\mathrm{thru},i}}{1 + \left(2\mathrm{Q}_{\mathrm{thru},i} \cdot \left(\lambda - \lambda_{r,i}\right)/\lambda_{r,i}\right)^2}\right). \tag{6.1}$$

The through-port spectrum of an n-channel Tx/Rx is modeled as the product of the through-port spectrum of each cascaded microring:

$$T_n(\lambda) = \prod_{i=1}^{n} T_{\mathrm{thru},i}(\lambda) = \alpha \cdot \prod_{i=1}^{n} \left(1 - \frac{1 - 1/\mathrm{ER}_{\mathrm{thru},i}}{1 + \left(2\mathrm{Q}_{\mathrm{thru},i} \cdot \left(\lambda - \lambda_{r,i}\right)/\lambda_{r,i}\right)^2}\right), \tag{6.2}$$

where $\alpha = \prod_{i=1}^{n} \alpha_{\mathrm{thru},i}$. Fig. 6.2a shows an example of the measured and fitted TRx spectra, with variations observed in the resonance wavelengths, extinction ratios, quality factors, as well as the overall through-port losses.

We adopt the process variation model for resonance wavelengths proposed in [225], where each resonance wavelength is decomposed into the design value and three independent variation components, namely the global variation (GV), the local variation (LV), and the Tx-Rx offset (TRxV). All of the three variation components are approximated by normal distributions with zero mean, as shown in Fig. 6.2b. The loss factor α is also approximated

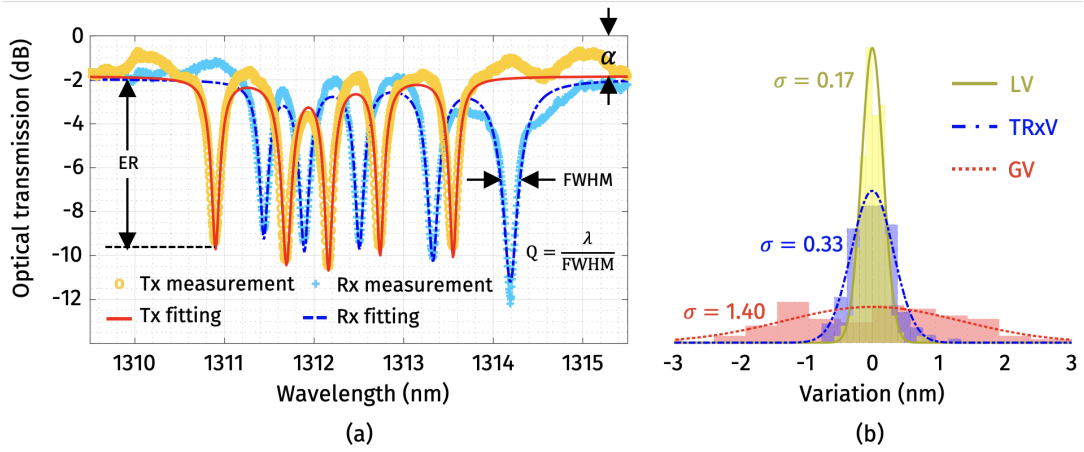


Figure 6.2: (a) Measured and fitted spectra of a 5-channel transceiver, where FWHM denotes the full width at half maximum, and (b) variation characterization for resonance wavelengths.

by a normal distribution with a mean of 0.68 and a standard deviation of 0.06.

Contrarily, location dependencies are observed in the process variations of ER and Q, prohibiting them from being modeled with independent variation components. Instead, we employ Virtual Probe (VP) [116, 119] to capture the spatial patterns from the measurement data and predict the values at other locations as our synthetic data for ER and Q. Fig. 6.3 shows the application of VP on measured ER as an example.

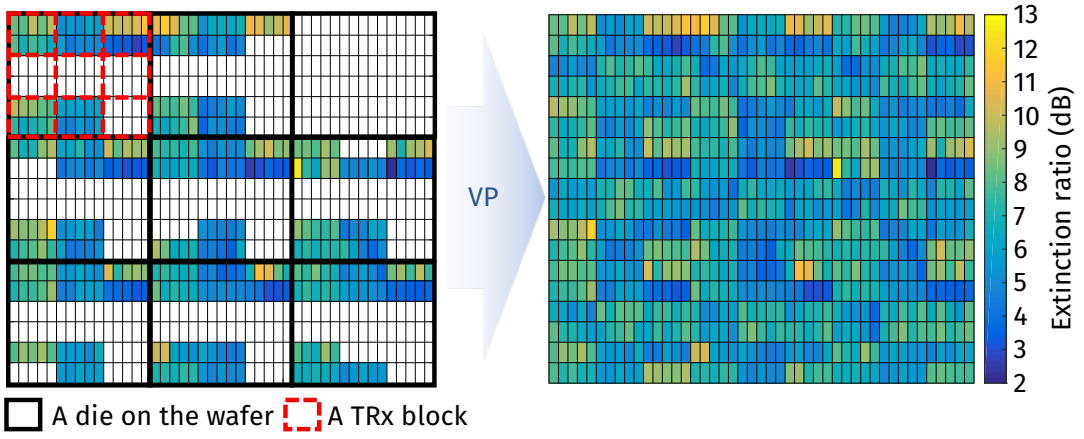


Figure 6.3: Relative locations of measured ER (left) and predicted values for unmeasured microrings by Virtual Probe (right).

Drop-Port Spectrum

The drop-port transmission spectrum of microring $\#i$ is also modeled as a Lorentzian function characterized by its resonance wavelength, extinction ratio, quality factor, and a drop-port loss $\alpha_{\text{drop},i}$:

$$T_{\text{drop},i}(\lambda) = \alpha_{\text{drop},i} \cdot \frac{1 - 1/\text{ER}_{\text{drop},i}}{1 + \left(2\text{Q}_{\text{drop},i} \cdot \left(\lambda - \lambda_{r,i}\right)/\lambda_{r,i}\right)^2}. \tag{6.3}$$

Process variations for each parameter are characterized using the same method described in Section 6.2.2. Finally, synthetic data of transceivers with up to 30 channels are generated and used, together with the measurement data, to evaluate our bidirectional tuning strategy.

6.3 Power Models and Assumptions

The total power consumption of a transceiver link for data communication includes those consumed by the comb laser, resonance wavelength tuning, modulator drivers, receiver transimpedance amplifier (TIA), and serializer/deserializer (SerDes) circuitry. The energy efficiency of the TRx is measured as the power consumption divided by the aggregated data rate, in the unit of pJ/b, for which the smaller the value the better.

6.3.1 Laser Power Budget

The QD comb laser is capable of generating a group of evenly-spaced frequency combs. However, the optical power of each comb line is usually different [21]. A Gaussian-shaped comb spectrum is assumed with a spectrum efficiency $\eta = P_{\text{usable}}/P_{\text{total}} \approx -3.2\,\text{dB}$ [18], as illustrated in Fig. 6.4. The optical power provided at the laser output should be high enough

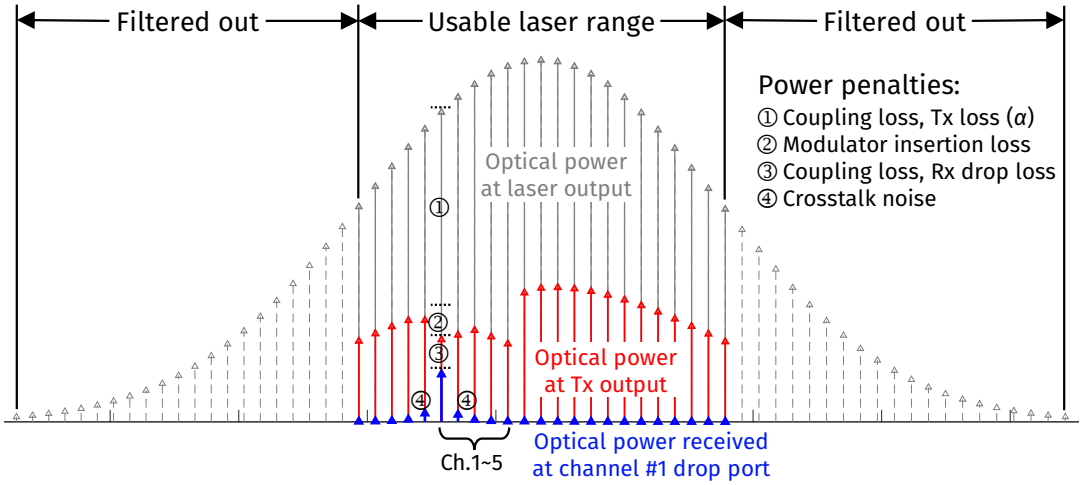


Figure 6.4: Illustration of the laser spectrum and power losses in a 5-channel transceiver.

so that the following power budget equation holds for any TRx channel:

$$P_{\text{comb line}} \cdot \prod_i \text{PL}_i \geqslant P_{\text{sensitivity}}. \tag{6.4}$$

Here, $P_{\text{comb line}}$ is the optical power of the laser comb line for a particular channel; $\text{PL}_i \in (0,1)$ denotes the various power losses along the TRx link, including the coupling loss, transmitter through-port loss (α in Eq. (6.2)), microring insertion loss, microring drop-port loss, crosstalk penalty, etc (Fig. 6.4); $P_{\text{sensitivity}}$ is the receiver sensitivity requirement, of which a relationship with the target data rate is given in [243]. The laser power budgets computed from Eq. (6.4) at various data rates show good consistency with both actual measurements [229, 245] and technology projections [21, 247].

Recent high-efficiency QD comb lasers have achieved up to 7–9 dBm per comb line at a ~20 % wall-plug efficiency (WPE) [21]. It is noteworthy that the optical nonlinearities of the microrings and the silicon waveguides limit the optical power that can be injected into the transceiver [244]. In this chapter, any transceiver that requires over 5 dBm optical power per channel or over 20 dBm optical power per waveguide is considered unusable.

6.3.2 Microring Wavelength Tuning

The microring tuning power is computed according to the tuning direction of the channel. For thermal tuning, a tuner efficiency of 0.15 nm/mW is assumed [242]. For electrical tuning, the EO effect model for carrier-injection microrings proposed in [88] is utilized to convert the tuning distance into current. Then, the electrical tuning power can be computed based on the microring DC model [88]. Meanwhile, the degradations of ER and Q of the microring can also be derived, and compensated by the increase of laser power budget based on the inequality of Eq. (6.4) described in Section 6.3.1.

6.3.3 Power Models for Other Components

The power consumptions of the modulator drivers, receiver TIAs, and SerDes circuitry mainly depend on the data rate per channel. In this chapter, lookup tables of power consumptions of these components are made for different data rates based on the values reported in [230].

6.4 Problem Formulation

Our bidirectional tuning strategy selects the carrier wavelengths for the TRx channels to align to. The difference between the resonance wavelengths before and after tuning determines the degradations of ER and Q and, ultimately, the shape of the TRx spectra after tuning. Therefore, given the spectrum-related parameters of a TRx before tuning, the overall energy-per-bit consumption of the TRx after tuning can be seen as a function of the selected carrier wavelengths:

$$\text{Energy per bit} = E\left(\vec{\lambda}_{\text{carrier}}\right). \tag{6.5}$$

The pool of candidate carrier wavelengths is a set of comb lines provided by the QD comb laser:

$$\text{Combs} = \{\lambda_1, \lambda_2, ..., \lambda_L\}, \tag{6.6}$$

where L is the number of usable comb lines. For an n-channel transceiver, the overall energy-per-bit value can only be computed once n carrier wavelengths are all determined. The goal is then to find $\vec{x} = [x_1, x_2, ..., x_n]$ where $x_i \in \mathbb{Z} \cap [1, L]$, such that

$$E\left([\lambda_{x_1}, \lambda_{x_2}, ..., \lambda_{x_n}]\right) \text{ is minimized.}$$

The vector $\vec{x}$, with all integer elements, is in fact the indices of the laser comb lines that are selected as carrier wavelengths. Our strategy preserves the order of TRx channels before and after tuning by forcing $x_i < x_{i+1}$ during the optimization.

To shrink the size of the problem space, we set the lower and upper bounds of the candidate carrier wavelengths for each TRx channel, as shown in Fig. 6.5. For the ith channel, the lower (upper) bound of its candidate carrier wavelengths lb_i (ub_i) is the index of the nearest laser comb line to the left (right) of both resonance wavelengths of the Tx and Rx microrings. Finally, our bidirectional tuning strategy is formulated as an integer program-

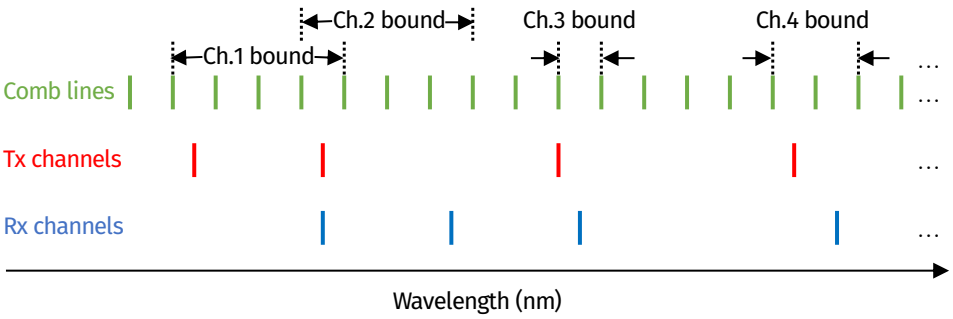


Figure 6.5: Bounds of candidate carrier wavelengths for each TRx channel.

ming problem of finding $\vec{x}$ subject to a lower bound $\overrightarrow{\text{lb}}$, an upper bound $\overrightarrow{\text{ub}}$, and a linear inequality constraint $A\vec{x} < \vec{b}$ (Eq. (6.7)), with the goal of minimizing the overall energy per bit E.

$$\overbrace{\begin{bmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & -1 \end{bmatrix}}^{A} \overbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}}^{\vec{x}} < \overbrace{\begin{bmatrix} -1 \\ -1 \\ -1 \\ \vdots \\ -1 \end{bmatrix}}^{\vec{b}}. \tag{6.7}$$

The NP-completeness of integer programming problems has been proven in [255]. In this chapter, we employ an implementation of the genetic algorithm in MATLAB [256] to solve for the optimal wavelength tuning scheme.

6.5 Evaluation

6.5.1 Evaluation on Measurement Data

We first evaluate our bidirectional tuning strategy on the measurement data of 5-channel transceivers. The tuning scheme given by our strategy is compared to all-electrical and all-thermal tuning schemes, as shown in Figs. 6.6a&b. Due to severe degradations of ER and Q and thus explosive increase in laser power budgets, all-electrical tuning results in fewer usable transceivers (and thus a lower yield) and a higher energy-per-bit consumption compared to all-thermal tuning, especially at a higher data rate per channel. On the contrary, our bidirectional tuning scheme does not compromise the yield of the TRx at any data rate per channel compared to all-thermal tuning. As all-thermal tuning is essentially a special

case of bidirectional tuning and is automatically chosen by our strategy when all other tuning schemes result in either worse energy-per-bit consumption or device failure, it serves as the worst case of our bidirectional tuning strategy.

The application of our strategy on the measurement data also brings significant energy-per-bit savings, as shown in Fig. 6.6b. For commonly used channel rates in the range of 5–10 Gb/s, which lead to relatively low energy-per-bit consumptions, our strategy (the right bar of each 3-bar group) results in 32–53 % savings of overall energy per bit compared to all-thermal tuning (the middle bar), with ~90 % of measured transceivers identified to benefit from bidirectional tuning.

Non-intuitively, our tuning strategy does not necessarily lead to a higher laser power consumption compared to all-thermal tuning, despite some degradations of ER and Q induced by electrical tuning in selected channels. The reason lies in the non-uniform laser power spectrum, where laser comb lines near the center inherently provide higher optical power than that of the outer ones. Fig. 6.6c shows an example of bidirectional tuning opportunities identified by our strategy. Electrical tuning on channels 1 and 2 aligns them to laser comb lines with higher optical power, with only minor degradations of ER and Q. It is noteworthy that our bidirectional tuning strategy has no knowledge of the Gaussian-shaped laser spectrum, and should be able to capture such opportunities for other laser spectrum

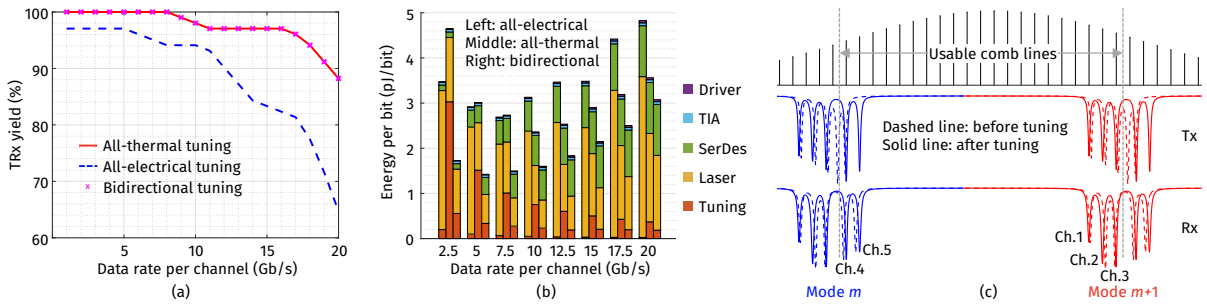


Figure 6.6: Evaluation of bidirectional tuning on measurement data: (a) yield comparison; (b) energy-per-bit comparison; and (c) bidirectional tuning opportunities identified by our strategy.

models (e.g., the flat-comb model in [18]).

6.5.2 Evaluation on Synthetic Data

We further evaluate our bidirectional tuning strategy on synthetic data of transceivers with 5–30 channels. Synthetic spectra of 81 transceivers, based on the variation model derived from the measurement data, are generated for each configuration. Fig. 6.7 summarizes the energy-per-bit savings of our bidirectional tuning strategy over all-thermal tuning. Up to 56 % saving can be attained for 5-channel transceivers at 5–10 Gb/s per channel, showing consistency with the simulation results on measured data. As the channel count increases, the free spectrum range of the microring is packed with more channels, which reduces the flexibility in carrier wavelength selection due to the inequality constraint expressed in Eq. (6.7) of Section 6.4. As a result, the attainable energy-per-bit saving of our bidirectional tuning strategy decreases for transceivers with higher channel counts. However, ~24 % energy-per-bit saving is still observed for 30-channel transceivers at 5 Gb/s per channel when bidirectional tuning is applied.

The problem space of bidirectional tuning grows exponentially with the number of channels, and thus requires a longer time for the genetic algorithm to converge. One of the stop-

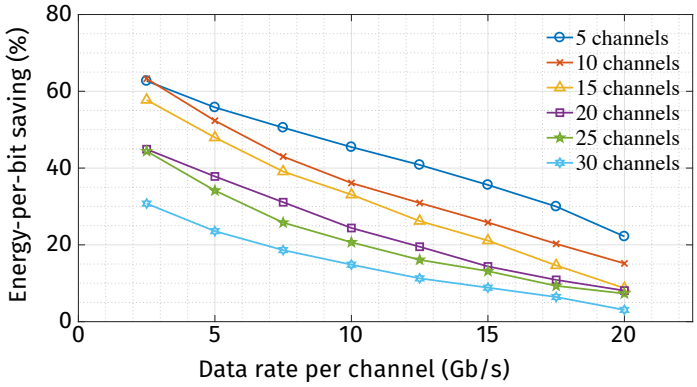


Figure 6.7: Evaluation of energy-per-bit savings of bidirectional tuning vs. all-thermal tuning on synthetic data of 5–30-channel transceivers.

ping criteria is the `MaxStallGenerations`, which sets the maximum number of generations over which the best value of the objective function stalls before the algorithm terminates. In Fig. 6.8, we explore the attainable energy-per-bit saving of our strategy and the computation time of the algorithm for various channel counts and values for `MaxStallGenerations`. The data rate is set at 7.5 Gb/s per channel. It is observed that a `MaxStallGenerations` of ~10 is large enough to approximate the maximum energy-per-bit saving without incurring unnecessarily long computation time.

It is worth mentioning that in practice, thermal tuners are fabricated for every microring, and electrical tuning is essentially a DC bias applied through the driver. As the optimal tuning scheme can be computed once the transceiver is fabricated and measured, our bidirectional tuning scheme does not increase the design complexity of the transceiver compared to all-thermal tuning, nor does it require sophisticated runtime reconfiguration support from the driving circuitry.

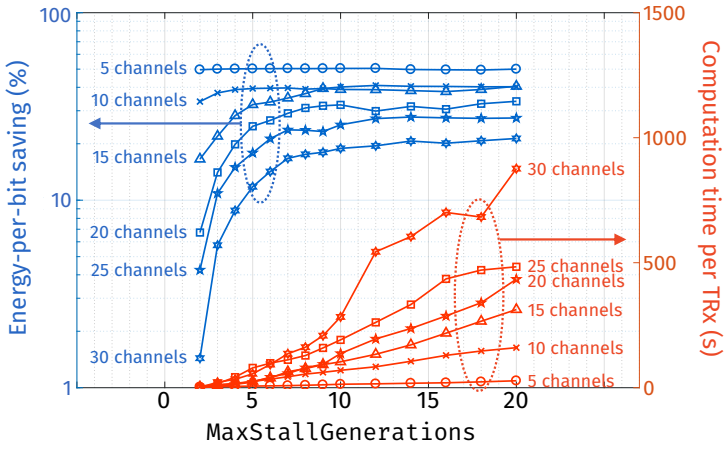


Figure 6.8: Energy-per-bit savings vs. computation time for bidirectional tuning.

6.6 Polynomial-Time Approximation

The computation of the optimal tuning scheme adds to the already long testing time, for which any extra time cost, if nontrivial, should be minimized. In this section, we investigate a polynomial-time approximation method to shorten the computation time of our bidirectional tuning strategy. Instead of computing all carrier wavelengths as a whole, our approximation method simplifies the problem by assuming that wavelength tuning of an individual channel has minor influence on its neighboring channels and thus determines the carrier wavelength for each channel one at a time. Algorithm 6.1 outlines the main steps of the approximation method. The energy-per-bit computation takes $O\left(n^2\right)$ time, resulting in an overall time complexity of $O\left(n^3\right)$ for this method.

Note that the approximation method greedily selects the carrier wavelength for each channel, which may lead to a local minimum. In order to explore a larger solution space,

Algorithm 6.1: A polynomial-time approximation method for bidirectional tuning.

Input : resonance wavelengths of the Tx and Rx, candidate carrier wavelengths of the comb laser.
Output : an array of selected carrier wavelengths for TRx channels to align to.
Goal : to minimize the overall energy per bit of the TRx.

// Initialization
Find $\overrightarrow{\text{lb}}$ and $\overrightarrow{\text{ub}}$ for all channels; // Fig. 6.5
selectedWavelengths ← empty array; // Indices of selected carrier wavelengths
// Selecting the carrier wavelength one channel at a time
for $j \leftarrow 1$ ***to*** n **do** // n is the number of channels
    $E_{\text{min}} \leftarrow +\infty$; // Best energy per bit so far
    for $i \leftarrow \overrightarrow{lb}_j$ ***to*** $\overrightarrow{ub}_j$ **do** // Try each wavelength within the bounds for channel # j
        Compute energy per bit E if comb line # i is selected for channel # j;
        if $E \leqslant \varepsilon \cdot E_{min}$ **then**
            $E_{\text{min}} \leftarrow E$; // Update best energy per bit
            Add i to selectedWavelengths;
            if $j < n$ **then** // Update the bounds for the next channel
                $\overrightarrow{\text{lb}}_{j+1} \leftarrow \max\left(\overrightarrow{\text{lb}}_{j+1}, i+1\right)$;
                $\overrightarrow{\text{ub}}_{j+1} \leftarrow \max\left(\overrightarrow{\text{ub}}_{j+1}, \overrightarrow{\text{lb}}_{j+1}\right)$;
Compute the link energy per bit based on the final selectedWavelength;
If worse than all-thermal tuning, use all-thermal tuning;

we introduce a parameter ε which, intuitively, is the tendency of the algorithm to choose thermal tuning over electrical tuning for a channel when both schemes have comparable energy-per-bit consumptions. An ε smaller than 1 encourages the algorithm to consider electrical tuning even if it is sub-optimal for a channel, yet to make room for subsequent channels on its right which may actually benefit from electrical tuning.

Fig. 6.9 shows the energy-per-bit savings of the tuning schemes given by this approximation method compared to all-thermal tuning, evaluated on the measurement data of 5-channel transceivers at various data rates from 5–10 Gb/s per channel. The energy-per-bit savings at optimal ε for this range of data rates are identified to be 28–48 %, which is quite comparable to the 32–53 % savings derived by the genetic algorithm in Section 6.5.1. However, the average computation time for the tuning scheme of each TRx reduces drastically compared to the GA method.

We further evaluate the approximation method on synthetic data of transceivers with higher channel counts to explore the trade-off between the speedup and the attainable energy-per-bit saving. A data rate of 7.5 Gb/s per channel is assumed in the experiments. The computation time of the GA method is measured with a `MaxStallGenerations` of 10. Table 6.1 lists the comparison between the approximation method and the GA method eval-

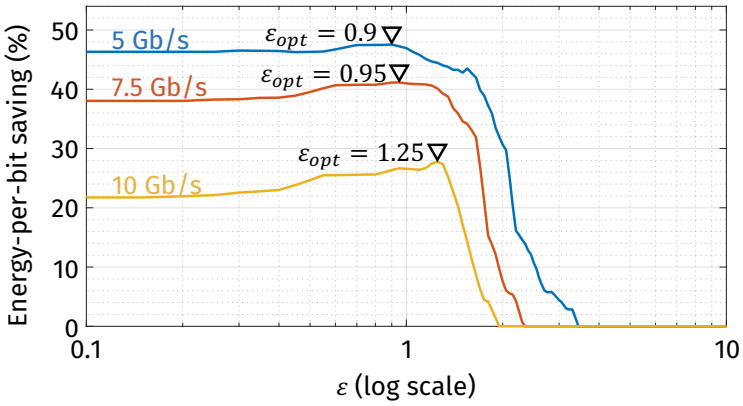


Figure 6.9: Effect of ε on the energy-per-bit savings of the polynomial-time approximation method.

Table 6.1: Polynomial-time approximation vs. genetic algorithm evaluated on synthetic data at 7.5 Gb/s per channel for energy-per-bit saving and computation time per TRx.

# of channels	**Energy-per-bit saving**		**Computation time** (s)		**Speedup**
	Genetic algorithm	Approximation	Genetic algorithm	Approximation	
5	51 %	46 %	12.817	0.107	119 x
10	43 %	37 %	67.362	0.304	221 x
15	39 %	35 %	101.195	0.495	204 x
20	31 %	27 %	190.273	0.893	213 x
25	26 %	23 %	108.929	1.027	106 x
30	17 %	15 %	310.796	2.805	110 x

uated across a wide range of channel counts. As observed from the simulation results, the polynomial-time approximation method is able to derive a tuning scheme over 100x faster than the GA method, while losing only 2–6 percentage points of the attainable energy-per-bit saving.

6.7 Concluding Remarks

To break the inherent bottleneck of energy-inefficient thermal tuners in microring-based optical links, we propose a bidirectional tuning strategy for carrier-injection microring-based transceivers which incorporates a hybrid of both thermal and electrical tuning. Our strategy determines the tuning direction of each TRx channel with the goal of minimizing the overall energy-per-bit consumption. The potential degradations of ER and Q are well accounted for, thus incurring no compromise in signal integrity. By formulating our strategy as an integer programming problem, we apply a genetic algorithm to solve for the optimal tuning schemes under various system configurations. The derived schemes lead to significant savings of overall energy per bit for measured and synthetic data of transceivers with various channel counts, based on commonly used data rates ranging from 5 to 10 Gb/s per channel. We further explore a polynomial-time approximation method which speeds up the

computation of bidirectional tuning schemes for over 100x, which adds negligible overhead on device testing time while losing only minor percentage points of attainable energy-per-bit savings.

Chapter 7

Transceiver Grouping for Microring-Based Optical Interconnects

Optical interconnects enabled by silicon microring–based transceivers offer great potential for short-reach data communication in future high-performance computing (HPC) systems. However, microring resonators (MRRs) are prone to process variations that harm both the energy efficiency and the yield of the fabricated transceivers. Especially in the application scenario where a batch of transceivers are fabricated for assembling multiple optical networks, how the transceivers are mixed and matched can directly impact the average energy efficiency and the yield of the networks assembled. In this chapter, we propose *transceiver grouping* for assembling communication networks from a pool of fabricated transceivers, aiming to optimize the network energy efficiency and the yield. We evaluate our grouping algorithms by wafer-scale measurement data of microring-based transceivers, as well as synthetic data generated based on an experimentally validated variation model. Our experimental results demonstrate that the optimized grouping schemes achieve significant improvement in the network energy efficiency and the yield across a wide range of network configurations, compared to a baseline strategy that randomly groups the transceivers.

7.1 Introduction

Optical interconnects are promising alternatives to electrical ones in modern HPC systems to accommodate traffic-intensive applications [1]. Silicon photonics has emerged as a scalable and cost-effective enabler of the optical interconnects by taking advantage of a CMOS-compatible fabrication process [13]. Silicon microring–based optical transceivers are one of the most popular implementations that achieve dense wavelength-division multiplexing (DWDM) with a compact footprint [228,229]. Various conceptual designs and prototypes of optical network-on-chips (ONoCs) have been reported [257,258,259] to leverage a microring-based architecture.

Despite great potential demonstrated, silicon microrings often suffer from significant process variations due to fabrication imperfection. As a result, the optical links and networks comprising these imperfect devices must be actively tuned to compensate for the process variations, for which the tuning power is non-trivial [230]. The variation issues become more prominent in the application scenario where a batch of transceivers are fabricated for assembling multiple optical networks. Specifically, some transceivers with straggling variation magnitudes may produce networks that either 1) demand excessive power for variation compensation or 2) fail to support a target data rate, thus worsening the average energy efficiency, the product uniformity, and the yield of the networks assembled. Nevertheless, network-level variation alleviation techniques that exploit wafer-scale fabrication of microring-based transceivers have been lacking. Techniques based on channel shuffling [173, 174] and sub-channel redundancies [175, 176, 225] are proposed to reduce the expected power for thermally tuning the resonance wavelengths of the microrings. A hybrid strategy employing both thermal and electrical tuning is proposed in [226]. However, these techniques are limited to the link-level, rather than the network-level, and only target a single pair of transmitter (Tx) and receiver (Rx). Considering wafer-scale fabrica-

tion, an optimal pairing scheme for a batch of fabricated transceivers could further reduce the average tuning power required for pairs formed from the batch [105]. Nevertheless, all of the above techniques are restricted to the mitigation of the wavelength tuning power, while the overall energy efficiency and the yield of the transceivers are also impacted by the variations of other parameters, such as the extinction ratios and the quality factors of the microrings. Moreover, none have encompassed the application scenario where the fabricated transceivers are used for assembling communication networks of multiple nodes.

We observe from wafer-scale measurement data of microring-based transceivers that, due to the distinct variation profile of each transceiver (TRx), optical networks assembled from different transceivers will have different energy efficiency. Therefore, when a batch of fabricated transceivers are available for assembling several networks, there is an opportunity to group the transceivers in a way that the average energy efficiency of the networks assembled is optimized. Meanwhile, it is also desirable, from the perspective of quality control, that the energy efficiency of the networks assembled is uniform. Moreover, some networks assembled may fail to support a target data rate, thus lowering the yield. Therefore, the grouping of the transceivers should also be optimized for the objective of meeting the target data rate. In this chapter, we propose *transceiver grouping* which mixes and matches a pool of fabricated transceivers to assemble networks of equal size, aiming to optimize the average energy efficiency, the uniformity, and the yield of the networks assembled.

We present two algorithms inspired by simulated annealing (SA) to address this multi-objective optimization problem. The proposed algorithms are evaluated by wafer-scale measurement data of microring-based transceivers, as well as synthetic data generated by an experimentally validated variation model. Our experimental results demonstrate that the proposed grouping algorithms achieve significant improvement in all three objectives, namely the average energy efficiency, the uniformity, and the yield of the networks assembled, compared to a baseline strategy that randomly groups the transceivers.

The rest of the chapter is organized as follows. In Section 7.2, we review the background of this chapter. In Section 7.3, we formulate transceiver grouping as an optimization problem and present our algorithms. In Section 7.4, we elaborate the measurement and the synthetic data of microring-based transceivers for evaluating our algorithms. We also introduce the power models of the optical devices used in our simulations. In Section 7.5, we evaluate our grouping algorithms for a wide range of network configurations. Finally, in Section 7.6, we draw the conclusion of this chapter.

7.2 Background

7.2.1 Microring-Based Optical Interconnects

An optical network is a collection of optical links that provides data communication among processing nodes. Fig. 7.1 illustrates an exemplar architecture of an optical network with a generic ring topology [257], where silicon microring–based transceivers are utilized to send and receive optical signals at each node. A silicon microring resonator is a highly wavelength-selective device [25], whose transmission spectrum can be characterized by a Lorentzian function:

$$T(\lambda) = 1 - \frac{1 - 1/\mathrm{ER}}{1 + (2\mathrm{Q} \cdot (\lambda - \lambda_r)/\lambda_r)^2}, \tag{7.1}$$

where λ_r, ER, and Q are the resonance wavelength, the extinction ratio, and the quality factor of the microring. A microring-based transceiver, as shown in Fig. 7.1, achieves DWDM by deploying cascaded microring resonators alongside a shared waveguide. At the Tx side, applying voltages to the microring modulators can slightly shift their resonance wavelengths to perform on-off keying (OOK) modulation of the optical signal. At the Rx side, corresponding microring filters can couple the signal out for detection. The overall transmission spec-

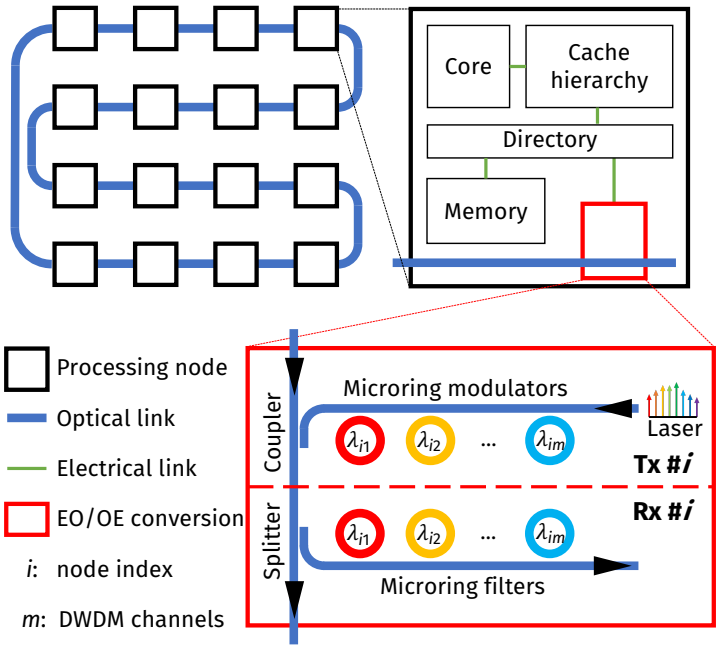


Figure 7.1: Illustration of an optical network with a ring topology.

trum of a Tx/Rx can thus be modeled as

$$T_m(\lambda) = \prod_{i=1}^{m} \left(1 - \frac{1 - 1/\mathrm{ER}_i}{1 + \left(2\mathrm{Q}_i \cdot (\lambda - \lambda_{r,i})/\lambda_{r,i}\right)^2} \right), \tag{7.2}$$

where m is the number of DWDM channels. The cascaded microrings of a Tx/Rx are usually designed with incremental radii to provide a set of evenly-spaced resonance dips. However, as shown in Fig. 7.2, the fabricated transceivers often suffer from significant process variations that manifest themselves as the deviation of λ_r, ER, and Q from their design values.

7.2.2 Impact of Process Variations on Energy Efficiency

The energy efficiency of an optical network largely depends on the energy efficiency of the links it comprises, which, in turn, is impacted by the process variations of the microrings. First of all, the resonance wavelengths of the Tx/Rx must be tuned and aligned to a mutual set of carrier wavelengths. Besides, the variations of ER and Q affect the loss and

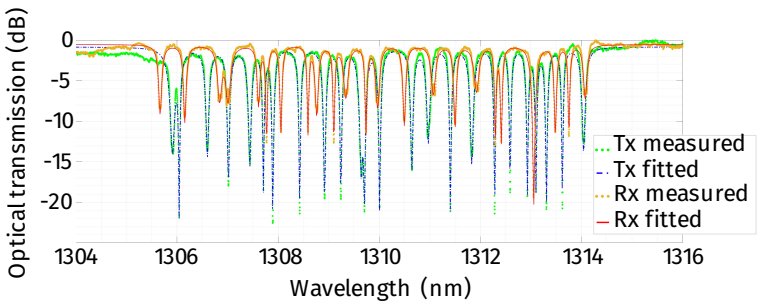


Figure 7.2: Measured and fitted transmission spectra of a 24-channel TRx.

the crosstalk noise within the optical channel, which must be compensated by an increased laser power to maintain a target data rate. As the variation magnitudes are different from device to device, optical networks comprising different transceivers will have different energy efficiency. Therefore, when a batch of fabricated transceivers are available for assembling such networks, how the transceivers are grouped can directly impact the energy efficiency of each network assembled.

7.2.3 Optimization Objectives for Transceiver Grouping

In the study featured in this chapter, we focus on the application scenario where a pool of fabricated transceivers are grouped to assemble several optical networks, as shown in Fig. 7.3. We assume the networks to be assembled are of a multiple-reader-multiple-writer (MWMR) architecture. For networks of other architectures, our proposed approach would still apply except that some specifics need to be adjusted. As illustrated in Fig. 7.1, each network node in MWMR has both write and read access to the optical ring bus achieved by its Tx and Rx, respectively. With a proper arbitration scheme [260], any two nodes can establish point-to-point communication without the need for relay nodes. Based on this assumption, we propose the following optimization objectives for transceiver grouping.

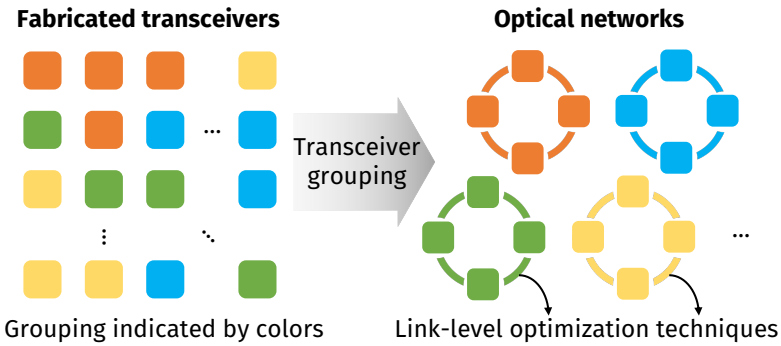


Figure 7.3: Illustration of transceiver grouping and its relationship with existing link-level optimization techniques [173, 174, 175, 176, 225, 226].

Energy Efficiency

We propose to optimize the average energy efficiency of the networks assembled. We first quantify the energy efficiency of an optical link as its power consumption divided by its data rate. Measured in pJ/b, the smaller the value, the better the energy efficiency. Now, consider a total of N transceivers to be grouped into G networks, each with n nodes ($G \leqslant N/n$). The energy efficiency of a network is thus a weighted sum of the energy efficiency of all its links:

$$e = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} p_{ij} \cdot \epsilon_{ij}, \tag{7.3}$$

where ϵ_{ij} is the energy efficiency of the unidirectional link from Tx $\#i$ to Rx $\#j$ (hereafter *link* (i, j)), and p_{ij} is the portion of the network traffic carried out by this link. The average energy efficiency of all networks assembled is thus

$$E = \frac{1}{G} \sum_{g=1}^{G} e_g = \frac{1}{G} \sum_{g=1}^{G} \left(\sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} p_{ij} \cdot \epsilon_{g,ij} \right), \tag{7.4}$$

where $\epsilon_{g,ij}$ denotes the energy efficiency of link (i, j) in the g^{th} network.

For a specific application, p_{ij} can be recorded by executing the application within a network simulator [186, 201] and can be different for each link. However, in this chapter, we assume that the network traffic results from the execution of various applications and is uniformly distributed to each link. Therefore, p_{ij} is considered identical for all links.

Note that the microring tuning schemes proposed in [173,174,175,176,225,226] are dedicated to improving ϵ_{ij} of a specific link, as shown in Fig. 7.3. However, regardless of which technique adopted at the link level, we can always apply transceiver grouping to further optimize the average energy efficiency of the networks.

Product Uniformity

Product uniformity is another victim of the process variations, as the energy efficiency can be vastly different for each network assembled. The authors of [100] suggest binning, a widely adopted technique after the testing stage, to categorize the transceivers based on the variation magnitudes. However, different bins may end up having different performance specifications, such as the maximum data rate. On the contrary, our transceiver grouping can improve the uniformity of the energy efficiency of the networks assembled without compromising the target data rate, thus delivering products with similar performance specifications. Specifically, we propose to reduce the standard deviation of the energy efficiency across the networks assembled:

$$\sigma = \sqrt{\frac{1}{G}\sum_{g=1}^{G}\left(e_g - E\right)^2}, \tag{7.5}$$

where all networks still target the same data rate. The transceiver pairing technique proposed in [105] is a special case of our transceiver grouping with $n = 2$. However, our study accounts for the overall energy efficiency for communication, in contrast to [105] that only targets the microring tuning power. Moreover, we further introduce a third optimization

objective for transceiver grouping, i.e., the network yield.

Network Yield

Apart from producing defective devices, the process variations could harm the yield in a way that some networks assembled cannot support a target data rate. Specifically, due to the optical nonlinearities of the silicon material, we assume a maximum optical power of 7 dBm per channel [244], which limits the highest data rate that an optical link can attain. We then propose to optimize

$$Y = G \Big/ \left\lfloor \frac{N}{n} \right\rfloor \cdot 100\,\%, \tag{7.6}$$

where G is the number of networks determined capable of supporting the target data rate. Note that in contrast to E and σ, Y is expected to be maximized.

As suggested by Eqs. (7.4) and (7.5), both E and σ can be computed from ϵ_{ij}. Therefore, for N transceivers available for grouping, it is desirable to prepare a cost matrix $\mathscr{E} \in \mathbb{R}^{N\times N}$ so that every possible ϵ_{ij} is computed beforehand for fast look-up. It is also noteworthy that ϵ_{ij} is computed as the link power consumption divided by the target data rate. During the computation of ϵ_{ij}, if the required optical power is found to exceed the maximum allowed value, the link and the network to which it belongs should be marked as not supporting the target data rate. The preparation of the cost matrix is detailed in Section 7.4.3 with a description of the device power models involved.

7.3 Problem Formulation

Consider a complete directed graph with N vertices and $N(N-1)$ directed edges, as illustrated in Fig. 7.4. Suppose that each vertex denotes a transceiver, and each edge is weighted

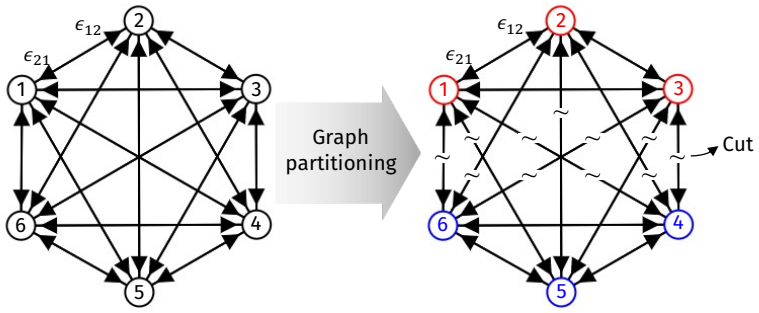


Figure 7.4: Transceiver grouping as a graph partitioning problem.

by ϵ_{ij}, the energy efficiency of link (i, j). Then, the objective of minimizing E, as suggested by Eq. (7.4), can be converted to finding a partition of the graph into equally sized blocks such that the sum of in-block edge weights is minimized. It is further equivalent to finding a partition of the graph with the maximum cut weights [261] with a balance constraint on the sub-graphs [262]. The NP-completeness of the graph partitioning problem has been proven [263], and several heuristic methods have been proposed for balanced partitioning [264, 265, 266]. However, the balance constraint in these algorithms is often formulated as a penalty to the cost function and might not be strictly satisfied. Directly applying them to transceiver grouping can result in groups of different sizes. Moreover, there exists no algorithm for balanced partitioning with multiple objectives. Therefore, we develop our customized heuristics for transceiver grouping.

7.3.1 Grouping Scheme Representation

To strictly enforce groups of equal size, we encode a grouping scheme of N transceivers into a vector $\boldsymbol{s}$, where $\boldsymbol{s}$ is a permutation of $\{1, 2, \ldots, N\}$. Every n elements of $\boldsymbol{s}$ are automatically grouped. For example, a grouping scheme for 16 transceivers into four 4-node

networks can be

$$\boldsymbol{s} = \left[\begin{array}{c:c:c:c} 6\,3\,16\,11 & 7\,14\,8\,5 & 15\,1\,2\,4 & 13\,9\,10\,12 \end{array}\right]. \tag{7.7}$$

It can be observed that any permutation of the elements within a group does not change the grouping scheme. However, such a representation allows us to easily generate new schemes by *shuffling* a current one:

$$\boldsymbol{s}' = [\ldots s'_u \ldots s'_v \ldots] = \text{shuffle}(\boldsymbol{s}) = [\ldots s_v \ldots s_u \ldots], \tag{7.8}$$

where u and v are randomly chosen from two different groups.

7.3.2 Proposed Algorithms

Simulated Annealing

Heuristics based on simulated annealing [267] can take advantage of the shuffling operation to explore various grouping schemes. We first present an SA-based algorithm (outlined in Algorithm 7.1) that aims to minimize a unified cost function:

$$Z = E + w_1 \cdot \sigma + w_2 \cdot (1 - Y). \tag{7.9}$$

Here, the objective E has a constant weight of 1, while the objectives σ and $(1 - Y)$ are weighted by w_1 and w_2, respectively. In other words, the energy efficiency of the networks assembled is always an optimization target, while the significance of the uniformity and the yield, as a second and a third optimization target, can be adjusted by the values of w_1 and w_2. At each SA iteration, a new grouping scheme $\boldsymbol{s}'$ is generated by shuffling the current $\boldsymbol{s}$, and its corresponding cost Z' is evaluated based on Eqs. (7.4)–(7.6) and Eq. (7.9).

Algorithm 7.1: simulated annealing for transceiver grouping.

Input : cost matrix $\mathscr{E}_{N\times N}$, initial grouping $\boldsymbol{s}_0$, w_1, and w_2.
Output : optimal grouping $\boldsymbol{s}^*$, E^*, σ^*, and Y^*.
Goal : to minimize $Z = E + w_1 \cdot \sigma + w_2 \cdot (1 - Y)$.

// Initialization
Set initial temperature T_0, cooling rate c, re-annealing interval $i_{\max}$, and rounds of annealing $r_{\max}$;
$r \leftarrow 1$, $\boldsymbol{s} \leftarrow \boldsymbol{s}_0$;
Compute Z from current $\boldsymbol{s}$; // Eqs. (7.4)–(7.6) and Eq. (7.9)
$Z^* \leftarrow Z$, $\boldsymbol{s}^* \leftarrow \boldsymbol{s}$;
// Simulated annealing
while $r \leqslant r_{max}$ **do**
    $i \leftarrow 1$, $T \leftarrow T_0$;
    while $i \leqslant i_{max}$ **do**
        $\boldsymbol{s}' \leftarrow \text{shuffle}(\boldsymbol{s})$; // Eq. (7.8)
        Compute Z' from $\boldsymbol{s}'$; // Eqs. (7.4)–(7.6) and Eq. (7.9)
        if $Z' < Z$ **then**
            $Z \leftarrow Z'$, $\boldsymbol{s} \leftarrow \boldsymbol{s}'$;
            if $Z' < Z^*$ **then**
                $Z^* \leftarrow Z'$, $\boldsymbol{s}^* \leftarrow \boldsymbol{s}'$;
        else
            $Z \leftarrow Z'$, $\boldsymbol{s} \leftarrow \boldsymbol{s}'$ with a probability of $p = P\left(Z, Z', T\right)$; // Eq. (7.10)
        $i \leftarrow i + 1$, $T \leftarrow T \cdot c$;
    $r \leftarrow r + 1$;

The algorithm decides whether to accept the new grouping scheme with a probability of $p = P\left(Z, Z', T\right)$:

$$P\left(Z, Z', T\right) = \begin{cases} 1 & Z' < Z \\ \frac{1}{1+\exp((Z'-Z)/T)} & Z' \geqslant Z \end{cases}, \tag{7.10}$$

where T is the current temperature. When Z' is no better than Z, there is still a probability between 0 and 1/2 to accept the new grouping scheme in order to avoid local minima.

The SA-based algorithm is seeded by an initial grouping scheme $\boldsymbol{s}_0$ which is produced by a greedy algorithm (outlined in Algorithm 7.2). At each iteration, the algorithm greedily groups n transceivers for the best network energy efficiency determined by Eq. (7.3), until $\lfloor N/n \rfloor$ groups are formed.

Algorithm 7.2: Greedy algorithm for initial grouping scheme.

Input : cost matrix $\mathscr{E}_{N\times N}$, group size n.
Output : initial grouping $\boldsymbol{s}_0$.
// Initialization
$\boldsymbol{s}_0 \leftarrow$ empty array, ungrouped $\leftarrow \{1,2,\ldots,N\}$;
// Greedy grouping
for $g \leftarrow 1$ ***to*** $\lfloor N/n \rfloor$ **do**
  $e^* \leftarrow \infty$, thisGroup $\leftarrow$ empty array;
  foreach $i \in$ *ungrouped* **do**
    forall $j \in$ *ungrouped* $\setminus \{i\}$ **do**
      Find $(n-1)$ transceivers with the smallest $(\epsilon_{ij}+\epsilon_{ji})$, indexed by $j_{i1}, j_{i2}, \ldots, j_{i(n-1)}$;
    Compute e_i for group $[i, j_{i1}, j_{i2}, \ldots, j_{i(n-1)}]$; // Eq. (7.3)
    if $e_i < e^*$ **then**
      $e^* \leftarrow e_i$, thisGroup $\leftarrow [i, j_{i1}, j_{i2}, \ldots, j_{i(n-1)}]$;
  Append thisGroup to $\boldsymbol{s}_0$;
  ungrouped $\leftarrow$ ungrouped $\setminus$ {elements of thisGroup};

Pareto Simulated Annealing

The SA-based algorithm allows the user to prioritize the three minimization targets, namely E, σ, and $(1-Y)$, by specifying w_1 and w_2. However, it presents another challenge to determine the proper values for w_1 and w_2. A straightforward approach is to sweep w_1 and w_2 within a given range. Alternatively, one may employ another optimization solver that takes w_1 and w_2 as input variables to explore their impact on the optimization results. Nevertheless, both methods involve an execution of the SA-based algorithm for each pair of w_1 and w_2 and thus can be time-consuming. To address this challenge, we further propose an algorithm based on Pareto simulated annealing (PSA) [268] to efficiently explore the trade-off between E, σ, and Y. Without the need to specify w_1 and w_2, the PSA-based algorithm directly targets

$$\boldsymbol{Z} = [z_1, z_2, z_3] = [E, \sigma, Y] \tag{7.11}$$

to find a Pareto front of the three optimization objectives where improving any objective will require sacrificing another. During the PSA iterations, a new solution ($\boldsymbol{Z}'$) is said to

Algorithm 7.3: Pareto simulated annealing for transceiver grouping.

Input : cost matrix $\mathscr{E}_{N\times N}$, initial grouping $\boldsymbol{s}_0$, population size n_p, and weight adjustment factor a.
Output : a Pareto front of E, σ, and $(1-Y)$.

// Initialization
Set initial temperature T_0, cooling rate c, re-annealing interval $i_{\max}$, and rounds of annealing $r_{\max}$;
Generate a set S of n_p grouping schemes by shuffling $\boldsymbol{s}_0$; // Eq. (7.8)
forall $\boldsymbol{s} \in S$ **do**
    Compute $\boldsymbol{Z}$ from $\boldsymbol{s}$ and update the Pareto front F; // Eqs. (7.4)–(7.6) and Eq. (7.11)
$r \leftarrow 1$;
// Pareto simulated annealing
while $r \leqslant r_{max}$ **do**
    $i \leftarrow 1$, $T \leftarrow T_0$, $\boldsymbol{\Gamma} \leftarrow [1/3, 1/3, 1/3]$;
    while $i \leqslant i_{max}$ **do**
        foreach $\boldsymbol{s} \in S$ **do**
            $\boldsymbol{s}' \leftarrow$ shuffle $(\boldsymbol{s})$; // Eq. (7.8)
            Compute $\boldsymbol{Z}'$ from $\boldsymbol{s}'$; // Eqs. (7.4)–(7.6) and Eq. (7.11)
            if $\boldsymbol{Z}'$ *dominates* $\boldsymbol{Z}$ **then**
                Update the Pareto front with $\boldsymbol{Z}'$ added;
                $\boldsymbol{s} \leftarrow \boldsymbol{s}'$;
            else
                for $i \leftarrow 1$ ***to*** *# of objectives* **do**
                    if $z'_i \geq z_i$ **then**
                        $\gamma_i \leftarrow \gamma_i \cdot a$;
                    else
                        $\gamma_i \leftarrow \gamma_i / a$;
                $\boldsymbol{s} \leftarrow \boldsymbol{s}'$ and update the Pareto front with $\boldsymbol{Z}'$ with a probability of $p = P\left(\boldsymbol{Z}, \boldsymbol{Z}', \boldsymbol{\Gamma}, T\right)$; // Eq. (7.12)
        $i \leftarrow i+1$, $T \leftarrow T \cdot c$;
    $r \leftarrow r+1$;

dominate an old one ($\boldsymbol{Z}$) if all three objectives of $\boldsymbol{Z}'$ are improved compared to that of $\boldsymbol{Z}$. Accordingly, the rule for deciding whether to accept a new grouping scheme is modified into

$$P\left(\boldsymbol{Z}, \boldsymbol{Z}', \boldsymbol{\Gamma}, T\right) = \begin{cases} 1 & \boldsymbol{Z}' \text{ dominates } \boldsymbol{Z} \\ \dfrac{1}{1+\exp\left(\sum_{i=1}^{3}\left(\gamma_i\left(z'_i - z_i\right)\right)/T\right)} & \text{otherwise} \end{cases}, \tag{7.12}$$

where $\boldsymbol{\Gamma}$ is a vector of weights associated with each optimization objective and automatically updated during the optimization. The larger is the weight of an objective, the lower is the

probability of accepting the new grouping scheme if it worsens the objective. At each PSA iteration, multiple new schemes can be generated and evaluated in parallel. Algorithm 7.3 outlines the main steps of our PSA-based algorithm.

7.4 Data Preparation

We now introduce the measured and synthetic data of microring-based transceivers for evaluating our algorithms. We also elaborate the computation of the cost matrix and the device power models involved.

7.4.1 Measurement Data

We measured the transmission spectra of the 24-channel microring-based transceivers fabricated by STMicroelectronics [231] on a 300 mm silicon-on-insulator (SOI) wafer. As illustrated in Fig. 7.5, the transceivers are organized into 66 dies, each die consisting of a transmitter and a receiver. The microrings in each Tx/Rx start with a 5 µm radius and ramp-up to a 5.046 µm radius with a step size of 2 nm. The Rx spectra of two dies were not measured correctly, as indicated in Fig. 7.5a. Thus, we have the measurement data of 64 fabricated transceivers for evaluating our grouping algorithms.

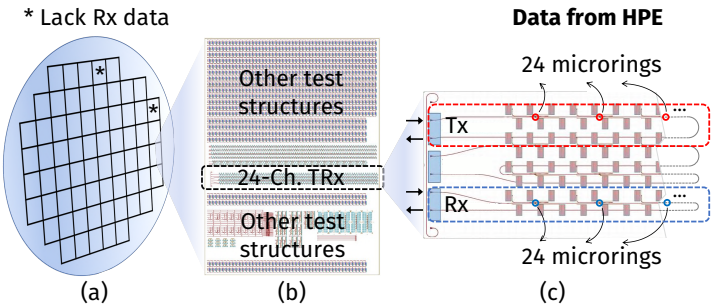


Figure 7.5: Organization of measured devices: (a) a wafer of 66 dies; (b) each die consisting of one TRx; and (c) each Tx/Rx consisting of 24 microrings.

7.4.2 Synthetic Data

To emulate situations where more transceivers are available for grouping, we generate synthetic data of transceivers to evaluate our grouping algorithms. We first extract the resonance wavelength (λ_r), the extinction ratio (ER), and the quality factor (Q) of each fabricated microring by fitting Eq. (7.2) to the measured spectra (Fig. 7.2). Then, we effectively characterize the spatial variations of λ_r, ER, and Q by applying our well-established variation modeling method [269]. Specifically, we attribute the location dependency of the variation magnitude on a wafer to three systematic components, namely wafer-level, intra-die, and inter-die components. This hierarchical method, detailed in [269], involves the usage of 1) robust regression [235] to fit the measurement data with several wafer-level basis functions, followed by 2) a spatial-frequency-domain analysis to extract the intra-die variation patterns, and 3) low-rank tensor factorization [270] to extract the inter-die variation patterns. Finally, we fit the residuals from this hierarchical decomposition process with a normal distribution $\mathcal{N}(\mu,\sigma)$ that is assumed spatially-stationary across the wafer. Fig. 7.6 visualizes the variation modeling process for λ_r as an example. The variations of ER and Q are modeled in the same manner, and the results are summarized in Table 7.1.

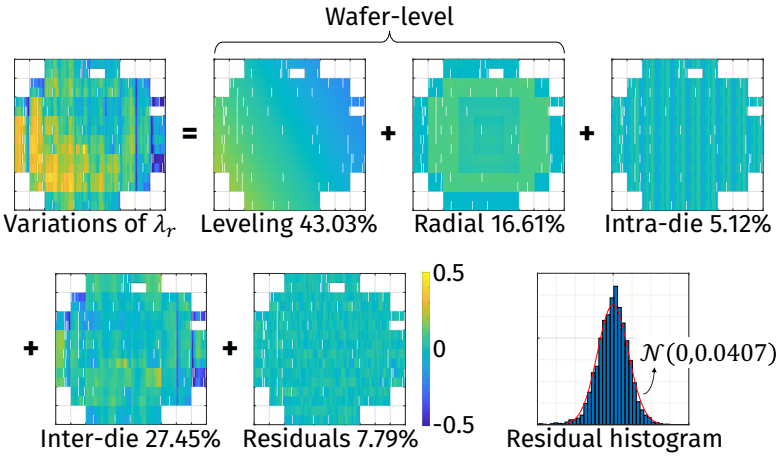


Figure 7.6: Wafer-scale variation characterization for λ_r, same process applied to ER and Q.

Table 7.1: Summary of spatial variation decomposition.

Test item	Wafer-level		Intra-die	Inter-die	Residual	Residual distribution
	Leveling	Radial				
λ_r	43.03 %	16.61 %	5.12 %	27.45 %	7.79 %	$\mathcal{N}(0, 0.0407)$
ER	0.04 %	1.60 %	32.67 %	18.61 %	47.08 %	$\mathcal{N}(0, 0.0891)$
Q	1.59 %	2.31 %	29.47 %	21.72 %	44.91 %	$\mathcal{N}(0, 0.0940)$

We generate wafer-level data for λ_r, ER, and Q following the variation model and synthesize them into transceiver spectra based on Eq. (7.2). To validate that our synthetic transceivers can closely resemble the fabricated ones in terms of power and energy estimation, we simulate the microring tuning power and the communication energy efficiency for the fabricated transceivers and ten wafers of synthetic transceivers. Fig. 7.7 plots the simulation results in ascending order for a data rate of 30 Gb/s per channel, showing a considerable resemblance of the synthetic transceivers to the fabricated ones. The power models used in these simulations are the same as those used for the computation of ϵ_{ij} and detailed in Section 7.4.3.

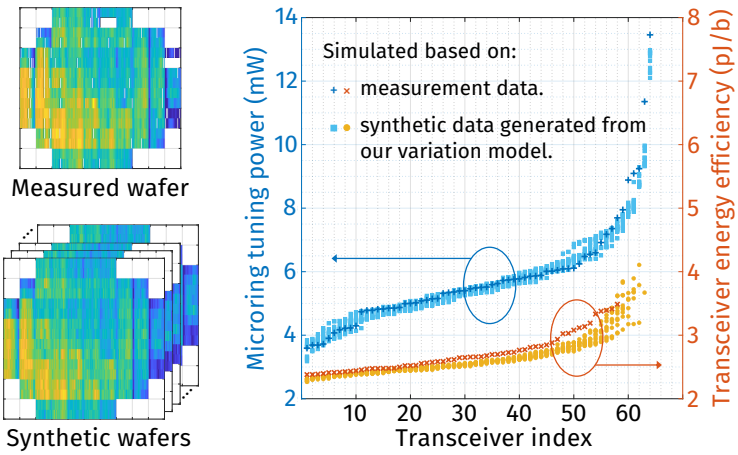


Figure 7.7: Synthetic data generation and validation.

7.4.3 Cost Matrix

For N transceivers available for grouping, a cost matrix $\mathscr{E} \in \mathbb{R}^{N\times N}$ is computed where the entry ϵ_{ij} is the energy efficiency of a unidirectional link from Tx $\#i$ to Rx $\#j$ at a given data rate, $i, j \in \{1,2,\ldots,N\}$. We compute ϵ_{ij} as the power consumption of the link divided by the aggregated data rate of all DWDM channels. The power consumption includes those of the laser, microring wavelength tuning, and Tx/Rx driver circuitry. Therefore, we have

$$\epsilon_{ij} = \frac{P_{\text{laser}} + P_{\text{tuning}} + P_{\text{driver}}}{m \cdot \text{DR}}, \tag{7.13}$$

where m is the number of DWDM channels, and DR is the target data rate per channel. The power models and assumptions are listed in Table 7.2 and explained as follows.

Laser Power

We assume a quantum-dot (QD) comb laser [21] that can generate a group of evenly-spaced frequency combs to cover the free spectrum range (FSR) of the microrings. We further assume a Gaussian-shaped comb spectrum, as illustrated in Fig. 7.8, with a spectrum efficiency $\eta = P_{\text{usable}}/P_{\text{total}} \approx -3.2\,\text{dB}$ [18]. The optical power provided at the laser output must be high enough so that the following power budget equation holds for any channel

Table 7.2: Models and assumptions for link power computation.

Laser			
Wall-plug efficiency	20 % [21]		
Data rate-dependent			
$P_{\text{sensitivity}}$	[243]	P_{driver}	[230]
Microring losses			
Passing	0.2 dB	Drop-port	1 dB
Insertion	Computed for a modulation distance of 0.2 nm		
Waveguide losses			
Coupling	1 dB [243]	Propagation	1 dB/cm [243]

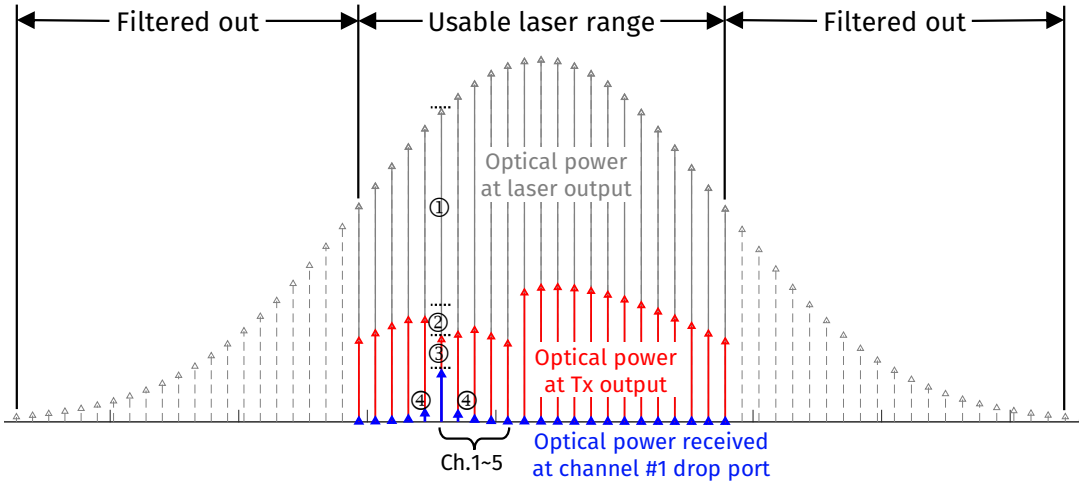


Figure 7.8: Power losses in a microring-based optical link, plotted for five channels for illustration purpose, including ① coupling loss and modulator passing loss; ② modulator insertion loss; ③ coupling loss, propagation loss, and Rx drop-port loss; and ④ crosstalk noise.

$k \in \{1, 2, \ldots, m\}$:

$$P_{\text{comb},k} \cdot \text{PL}_k \geqslant P_{\text{sensitivity}}. \tag{7.14}$$

Here, $P_{\text{comb},k}$ is the optical power of the k^{th} comb line; $\text{PL}_k \in (0,1)$ is the overall power loss of the k^{th} channel, which is the product of several losses (listed in Fig. 7.8) as the light travels; $P_{\text{sensitivity}}$ is the sensitivity requirement of the receiver and is modeled as a function of the data rate in [243]. The laser is characterized by the wall-plug efficiency (WPE) when converting the electrical power into the optical power:

$$P_{\text{laser}} \cdot \text{WPE} = \left(\sum_{k=1}^{m} P_{\text{comb},k} \right) / \eta. \tag{7.15}$$

Based on Eqs. (7.14) and (7.15), the laser power consumption can be computed for various data rates and is consistent with what reported in [21]. Note that if the required optical power for Eq. (7.14) to hold exceeds the maximum power allowed (7 dBm as per [244]), the

link is marked as not supporting the target data rate.

Microring Tuning

The P_{tuning} term in Eq. (7.13) is the tuning power required to align the microring resonance wavelengths of Tx $\#i$ and Rx $\#j$ to a mutual set of laser comb lines. We assume that thermal tuning is adopted to redshift the resonance wavelengths of the microrings with a tuning efficiency of 0.15 nm/mW [242]. If some resonance wavelengths fall out of the usable laser range, channel shuffling [173, 174] is applied to utilize a neighboring mode for alignment.

Driver Circuitry

We consider the modulator drivers, the receiver transimpedance amplifier (TIA), and the serializer/deserializer (SerDes) circuitry as the main components of the driver circuitry of an optical link, thus:

$$P_{\text{driver}} = P_{\text{mod}} + P_{\text{TIA}} + P_{\text{SerDes}}. \tag{7.16}$$

A decent analysis is provided in [230] that models the power of the driver circuitry as a function of the data rate. In this chapter, we made lookup tables for P_{driver} at various data rates for the computation of ϵ_{ij}.

Note that for network topologies other than the ring bus described in Section 7.2.3, one can adjust Eq. (7.13) accordingly for computing ϵ_{ij}, which is the energy efficiency of a unidirectional link from Tx $\#i$ to Rx $\#j$ including relay nodes (if there are any), so that the transceiver grouping algorithms proposed in Section 7.3.2 can be directly applied without modification.

7.5 Evaluation

We evaluate our SA- and PSA-based algorithms for transceiver grouping based on the data of 64 measured transceivers and up to 256 synthetic transceivers for a wide range of network configurations.

7.5.1 SA-based Grouping Algorithm

Effectiveness

We first present a few case studies to demonstrate the effectiveness of our SA-based algorithm (Algorithm 7.1). Fig. 7.9 shows an example for $N = 16$ and $n = 2$ at a target data rate of 30 Gb/s per channel. Several grouping schemes are illustrated in the form of graphs, including random grouping, local grouping, greedy grouping, and three grouping schemes produced by the SA-based algorithm with different w_1's and w_2's. The nodes in each graph represent the transceivers available for grouping (i.e., pairing when $n = 2$). The energy efficiency of each group (pair) is computed from the data of the first 16 measured transceivers.

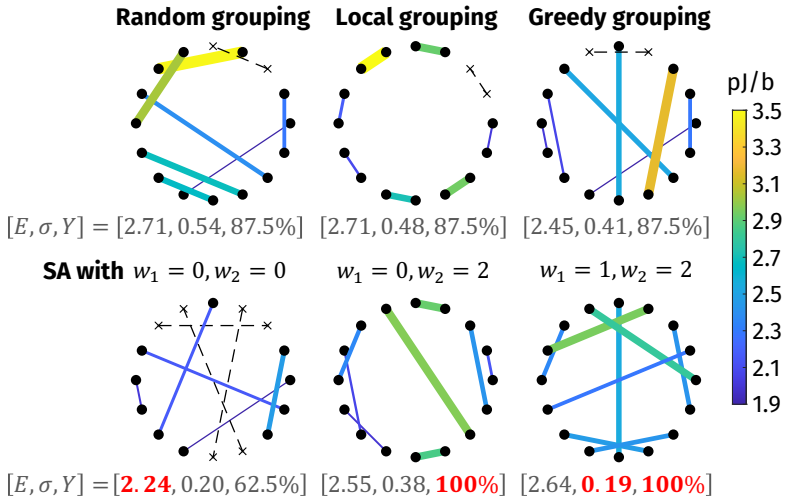


Figure 7.9: Illustration of grouping schemes for $N = 16$ and $n = 2$ at a target data rate of 30 Gb/s per channel.

The thinner an edge, the better the energy efficiency. A dashed line, however, indicates that the link cannot support the target data rate.

We observe from Fig. 7.9 that, compared to a random grouping scheme, the local grouping scheme that groups neighboring transceivers on a wafer only achieves marginal improvement in E and σ. It might seem non-intuitive, as local grouping should mitigate the impact of wafer-level variations. However, as suggested in Table 7.1, even transceivers that are close to each other still suffer from significant inter-die variations. The observation justifies the need for more sophisticated grouping algorithms. We further observe that

- the greedy algorithm is able to achieve considerable improvement in E but not σ, as the transceivers that lead to better energy efficiency are greedily grouped at earlier steps, leaving the remaining ones grouped at later steps incurring significantly worse energy efficiency;
- the SA-based algorithm, which starts the optimization by shuffling the greedy grouping scheme, can further improve E when $w_1 = w_2 = 0$, but may converge to a solution with a low yield;
- the SA-based algorithm can also improve σ and Y by increasing their corresponding weights, at the cost of less improvement in other objectives.

We then use the *energy-yield curves* to compare different grouping schemes for other network configurations. Fig. 7.10 provides two more cases for $N = 32$ and 64, $n = 4$, at a target data rate of 30 Gb/s per channel. Specifically, for each grouping scheme, we plot the energy efficiency of all networks assembled in ascending order, so that the average energy efficiency and the uniformity of the networks assembled can be visualized by the position and the slope of a curve. On the other hand, the horizontal axis of the plot, i.e., the network index $g \in \{1,2,\ldots,G\}$, is normalized by $\lfloor N/n \rfloor$. Then, as defined by Eq. (7.6), the network

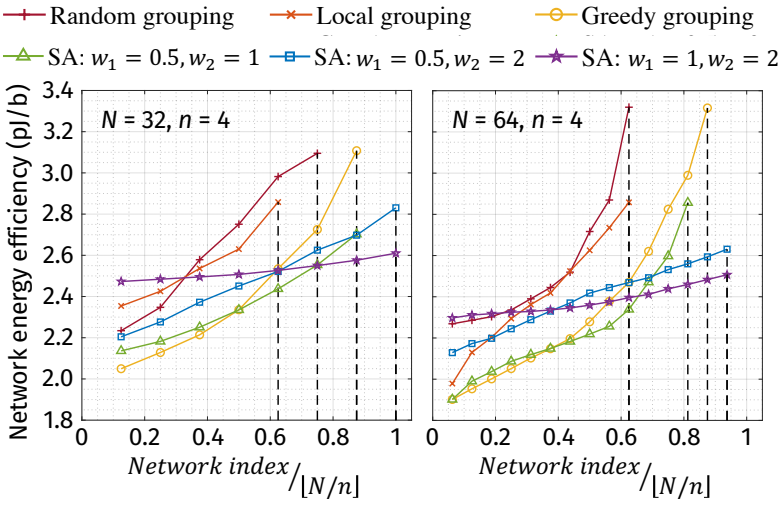


Figure 7.10: Comparison of grouping schemes for $N = 32, 64$ and $n = 4$ at a target data rate of 30 Gb/s per channel.

yield of a grouping scheme can thus be visualized by the x-coordinate of the ending point of the corresponding curve, as indicated by the vertical dashed lines in Fig. 7.10. The energy-yield curves again verify that our SA-based algorithm, with a proper assignment of w_1 and w_2, can achieve significant improvement in the average energy efficiency and the yield of the networks assembled, while drastically improving the uniformity compared to a random grouping scheme.

Scalability

We further evaluate our SA-based algorithm for a variety of network configurations that cover $N \in \{16, 32, 64, 128, 256\}$, $n \in \{2, 4, 8, 16\}$, and a target data rate ranging from 20 Gb/s to 30 Gb/s per channel. We compute the improvement in E, σ, and Y achieved by our SA-based algorithm over random grouping. Note that the improvement in E and σ is measured by the percentage of reduction compared to that of the random grouping scheme, while the improvement in Y is measured by the arithmetic difference of the yields (a.k.a. *percentage point* or *p.p.*) of the two grouping schemes. For example, improving the yield from 50 % to 80 % is considered as an increase of 30 percentage points, rather than a 60 % increase. Over-

all, our SA-based algorithm with $w_1 = 1$ and $w_2 = 2$ achieves up to 25 % improvement in the average energy efficiency of the networks assembled, up to 94 % reduction of the standard deviation of the energy efficiency, and up to 75 percentage points increase of the network yield, compared to a random grouping scheme for the network configurations evaluated. Furthermore, we observe several trends from the evaluation results that are noteworthy:

- As shown in Fig. 7.11a, for a given network size (n) and a target data rate, the energy efficiency improvement achieved by our SA-based algorithm increases with N, i.e., the total number of transceivers. In other words, with more transceivers available for grouping, there is a greater opportunity to optimize the average energy efficiency of the networks assembled.

- As shown in Fig. 7.11b, for a given number of transceivers available for grouping, the reduction of the standard deviation of the energy efficiency, achieved by our SA-based algorithm, is more significant for a larger n. In other words, when the networks to be assembled are of a larger size, there is a greater opportunity to group the transceivers in a way that the networks assembled have relatively similar energy efficiency.

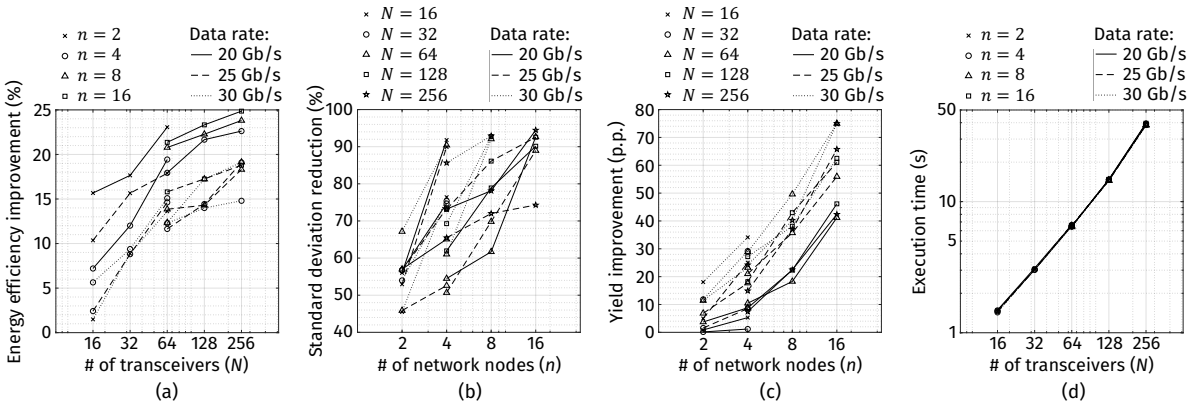


Figure 7.11: (a)–(c) Improvement achieved by our SA-based algorithm with $w_1 = 1$ and $w_2 = 2$ over the random grouping scheme in energy efficiency, yield, and uniformity, evaluated for various network configurations, and (d) execution time of our SA-based algorithm for various network configurations.

- As shown in Fig. 7.11c, for a given number of transceivers available for grouping, the yield improvement achieved by our SA-based algorithm is greater for a larger n and a higher data rate. It was observed that the network yield resulted from a random grouping scheme drastically decreases with the network size and the target data rate. Especially for $n = 16$, none of the randomly assembled networks could support a target data rate of 30 Gb/s. Nevertheless, our SA-based algorithm can maintain a reasonably high yield for all network configurations evaluated.

The execution time of our SA-based algorithm is recorded for an initial temperature of 100, a cooling rate of 0.95, a re-annealing interval of $(10 \times N)$ iterations, and 50 rounds of annealing. Thus, each optimized grouping scheme is produced from a total of $(500 \times N)$ annealing iterations. According to Fig. 7.12, this setting is empirically found adequate for Eq. (7.9) to converge to a steady value. As shown in Fig. 7.11d, the execution time of our SA-based algorithm grows polynomially with the number of transceivers and is largely independent of other network parameters. Limited within 40 s for $N = 256$, the execution time of our SA-based algorithm is considered a small overhead to the test time of the fabricated transceivers.

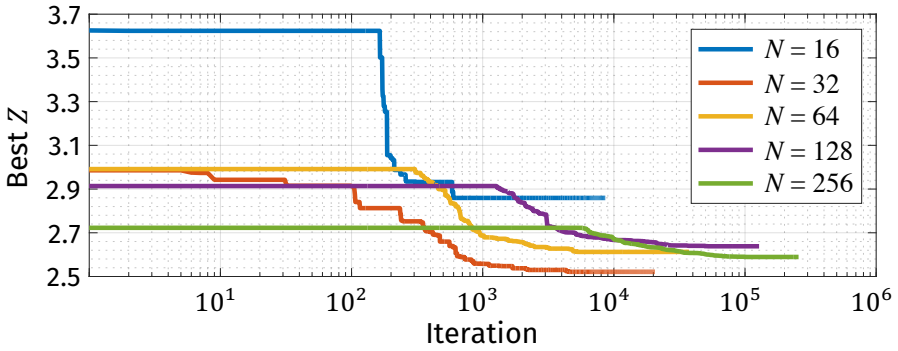


Figure 7.12: Evolution of the cost value with SA iterations, plotted for $n = 4$ and 30 Gb/s per channel as an example.

7.5.2 PSA-based Grouping Algorithm

The SA-based algorithm requires a proper combination of w_1 and w_2 to be specified. To avoid excessive trials only to determine the values for w_1 and w_2, the SA-based algorithm is best suited for situations where 1) either the uniformity or the yield of the networks assembled has an overriding priority over the other, so that having w_1 or w_2 equal to zero generally works well; or 2) the proper values for w_1 and w_2 are already learned from past runs for the network configuration of interest. For situations where the proper values for w_1 and w_2 are unknown, our PSA-based algorithm (Algorithm 7.3) can effectively and efficiently explore the trade-off between the three optimization objectives, namely the energy efficiency, the uniformity, and the yield of the networks assembled. By giving a set of Pareto-optimal solutions in a single run, our PSA-based algorithm allows one to select a desired grouping scheme without the need to specify w_1 and w_2. We compare our PSA-based algorithm to two other methods that explore the same trade-off by varying the combination of w_1 and w_2:

1) To sweep w_1 and w_2 within a given range (hereafter the *SWEEP* method). For each combination of w_1 and w_2, the SA-based algorithm is called to optimize Eq. (7.9). The Pareto front of E, σ, and Y is derived after the sweeping by eliminating the dominated solutions.

2) To employ another optimization solver that takes w_1 and w_2 as input variables. In this study, we modified an existing implementation of Multi-Objective Particle Swarm Optimization [271] (hereafter the *MOPSO* method). In each generation, the MOPSO method generates multiple combinations of w_1 and w_2 and calls the SA-based algorithm to optimize Eq. (7.9) for each combination. The Pareto front of E, σ, and Y is updated at the end of each generation, and new combinations of w_1 and w_2 are generated for the next generation based on the current Pareto front.

Effectiveness

For each network configuration, i.e., given N, n, and a target data rate, a Pareto front of E, σ, and Y is explored by SWEEP, MOPSO, and our PSA-based algorithm with the following settings, respectively:

SWEEP We sweep both w_1 and w_2 from 0.2 to 2 with a step size of 0.2. Thus, a total of 100 different combinations of w_1 and w_2 are explored. For each combination of w_1 and w_2, a grouping scheme is optimized through $(500 \times N)$ SA iterations.

MOPSO We specify a population size of 10 for the MOPSO method, i.e., ten combinations of w_1 and w_2 generated and evaluated in each generation. Thus, a total of 100 combinations of w_1 and w_2 are explored in 10 generations, each producing a grouping scheme optimized through $(500 \times N)$ SA iterations.

PSA We execute our PSA-based algorithm for $(500 \times N)$ iterations with a population size of 100, where each individual in the population is a candidate grouping scheme. In other words, 100 grouping schemes are simultaneously optimized through $(500 \times N)$ PSA iterations.

Fig. 7.13 shows the results for $N = 32, 64$, $n = 4$, and $N = 128, 256$, $n = 8$, at a target data rate of 30 Gb/s per channel. Specifically, each plotted point corresponds to a grouping scheme, whose E and σ can be read from its x- and y-coordinates, respectively. The value of Y is color-coded from light yellow (lowest) to dark blue (highest). Therefore, a grouping scheme is considered a better one if it is closer to the bottom left corner and darker in color. The random, local, and greedy grouping schemes are also marked in each plot. We compare the Pareto-optimal grouping schemes produced by SWEEP, MOPSO, and our PSA-based algorithm and make the following observations:

- The yield of the networks assembled, as suggested by Eq. (7.6), can only take a few discrete values. Thus, the Pareto front of E, σ, and Y appears as multiple curves that correspond to different yield values. Taking Fig. 7.13a as an example, for a network configuration of interest, one may pick a grouping scheme from the Pareto front by first specifying an acceptable yield value, then selecting a grouping scheme on the corresponding curve that reflects the desired trade-off between E and σ.

- In all four plots of Fig. 7.13, most of the Pareto-optimal solutions given by SWEEP and MOPSO are overlaid by solutions given by our PSA-based algorithm. In other words, our PSA-based algorithm can produce Pareto-optimal grouping schemes as good as those identified by SWEEP and MOPSO.

- For $N = 128$ and 256, both SWEEP and MOPSO tend to produce grouping schemes with a low yield. However, our PSA-based algorithm can still explore various grouping schemes with a reasonably high yield.

- Our PSA-based algorithm can always identify multiple grouping schemes that are significantly better than the random grouping scheme in all three optimization objectives, namely E, σ, and Y.

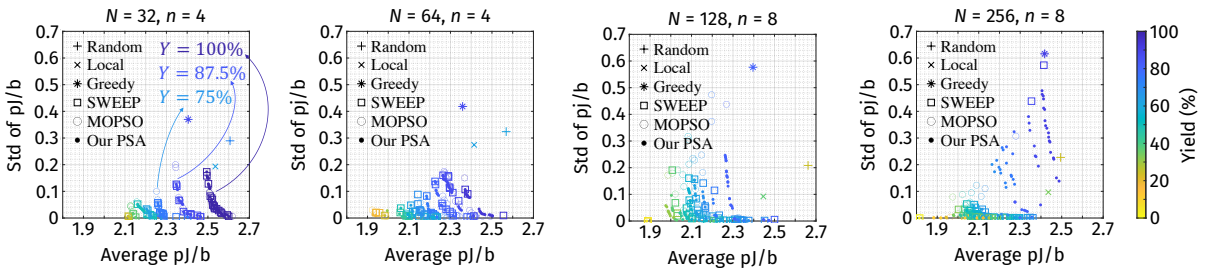


Figure 7.13: Comparison of the Pareto fronts of E, σ, and Y produced by SWEEP, MOPSO, and our PSA-based algorithm for (a) $N = 32$, $n = 4$, (b) $N = 64$, $n = 4$, (c) $N = 128$, $n = 8$, and (d) $N = 256$, $n = 8$, at a target data rate of 30 Gb/s per channel.

Efficiency

We define the efficiency of SWEEP, MOPSO, and our PSA-based algorithm as the number of Pareto-optimal grouping schemes that can be produced in unit time. Using the settings specified in Section 7.5.2, the number of candidate grouping schemes to be optimized by each method was 100, while some of the optimized grouping schemes ended up not on the Pareto front. Fig. 7.14 compares the efficiency of our PSA-based algorithm to that of SWEEP and MOPSO for various network configurations, and the following observations are made:

- The MOPSO method brings a minor increase in the number of Pareto-optimal grouping schemes compared to the SWEEP method, at the cost of longer execution time for the same amount of candidates evaluated. On average, the MOPSO method only achieves 0.97x efficiency compared to the SWEEP method.
- Our PSA-based algorithm, compared to both SWEEP and MOPSO, can produce significantly more Pareto-optimal grouping schemes within a shorter execution time for all network configurations evaluated. Overall, our PSA-based algorithm achieves 1.67x to 9.30x improvement in terms of efficiency with an average of 3.13x, compared to the SWEEP method.

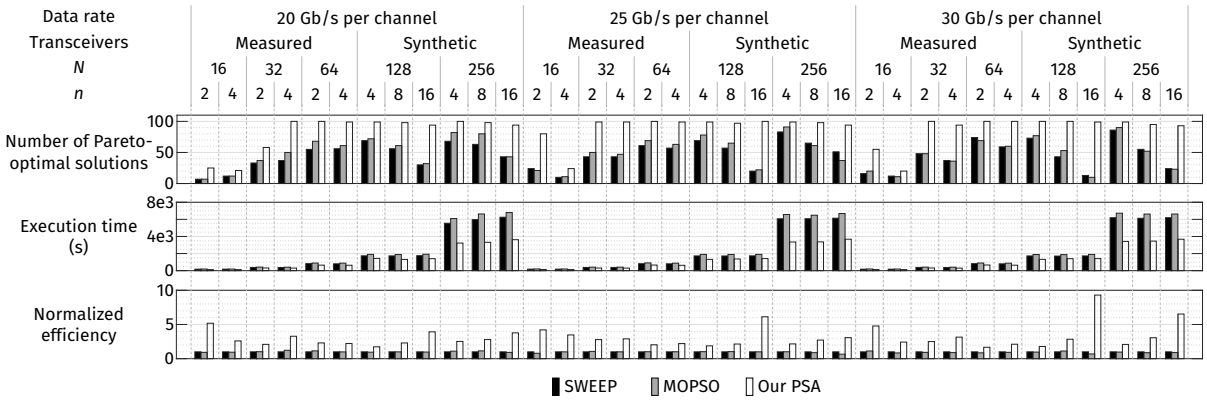


Figure 7.14: Number of Pareto-optimal solutions, execution time, and efficiency comparison of SWEEP, MOPSO, and our PSA-based algorithm.

In a nutshell, when a proper combination of w_1 and w_2 is unknown, our PSA-based algorithm can explore a larger solution space with better efficiency compared to SWEEP and MOPSO, producing more Pareto-optimal grouping schemes for selection.

7.6 Concluding Remarks

In this chapter, we target the application scenario where fabricated microring-based transceivers are grouped for assembling optical networks of multiple nodes. We propose two algorithms to mix and match the fabricated transceivers so that the three optimization objectives, namely the average energy efficiency, the uniformity, and the yield of the networks assembled, are optimized. We evaluate our proposed algorithms by wafer-scale measurement data of microring-based transceivers, as well as synthetic data generated from an experimentally validated variation model. Our first algorithm based on simulated annealing can achieve up to 25 % improvement in the average energy efficiency of the networks assembled, up to 94 % reduction of the standard deviation of the energy efficiency, and up to 75 percentage points increase of the network yield, compared to a baseline strategy that randomly groups the transceivers. Moreover, our second algorithm based on Pareto simulated annealing can efficiently produce multiple Pareto-optimal grouping schemes that significantly outperform the random grouping scheme in all three optimization objectives, namely the energy efficiency, the uniformity, and the yield of the networks assembled.

Part IV

System-Level Runtime Power Reconfiguration

Chapter 8

TMALPS: Task Mapping–Assisted Laser Power Scaling

Energy efficiency of an optical network-on-chip (ONoC) largely relies on an effective laser power management strategy. Addressing the limitations of existing techniques, we propose a Task Mapping–Assisted Laser Power Scaling (TMALPS) framework to optimize the energy consumption and the application execution time of an ONoC. Through the combination of task mapping exploration and runtime laser power reconfiguration applied to a wide range of application benchmarks, our TMALPS framework achieves an average of 66 % saving of the energy-delay product, compared to a baseline scenario where the optimization techniques are not applied. Significant improvement over existing techniques was also observed. The hardware overhead required to support our TMALPS framework is minimal with intelligent reuse of existing on-chip hardware resource.

8.1 Introduction

The multi-processor system-on-chip (MPSoC) has been proposed for broad adoption in modern high-performance computing (HPC) systems to accommodate traffic-intensive applications [272]. With the network complexity in modern MPSoCs ever growing, the optical network-on-chip was proposed to provide better bandwidth and energy efficiency than

that of all-electrical implementations [1]. However, such energy efficiency can be compromised if the laser power is not well managed. The authors of [207] propose to provide just enough laser power for each optical channel depending on its quality. Nevertheless, their solution lacks a reconfiguration mechanism under dynamic workloads. Recent advances in on-chip laser designs, and thus the improved laser power controllability [18], bring about traffic-adaptive reconfiguration strategies that switch the lasers `ON` and `OFF` at application runtime [197, 198, 273]. Some of these strategies are reviewed in [199] and categorized as *coarse-grained (`ON-OFF`)* strategies as the lasers have solely two configurable states. Due to the large turn-on delay of the lasers [274], the energy saving obtained from the `OFF` state may be counteracted by the energy wasted when switching the lasers back `ON`, and the application execution time may also be prolonged. To tackle this issue, a *fine-grained* strategy is proposed in [199], a.k.a. Dynamic Laser Power Scaling (DLPS), which further introduces a `STANDBY` state and an `INTERMEDIATE` state to the lasers in order to leverage a smaller switching delay. However, despite some improvement over the `ON-OFF` strategies observed for a few application benchmarks, the effectiveness of the DLPS strategy is found minimal for many other application benchmarks as they do not have favorable traffic patterns. This fundamental limitation of the DLPS strategy, to be further elaborated in Section 8.2.2, motivates us to explore adjustments to the traffic patterns so that laser power reconfiguration techniques can be maximally exploited.

In this chapter, we propose our Task Mapping–Assisted Laser Power Scaling (TMALPS) framework to jointly optimize the energy consumption and the application execution time of an ONoC. Task mapping is the process of assigning tasks to the processing units available on the MPSoC. Thus, different mapping schemes result in different traffic patterns during application execution. Though well-studied for electrical network-on-chips (NoCs) [275, 276, 277], task mapping has only started to draw attention from ONoC researchers in recent years. The authors of [278] propose an ONoC task mapping algorithm that optimizes the

worst case crosstalk noise and optical loss and based on which, further optimizes the average laser power budget [279]. However, a minimum laser power budget does not necessarily lead to a minimum energy consumption if the application execution time is lengthy under that particular mapping scheme. Their energy computation method in [279] based on hop counting is oversimplified. In this chapter, we employ validated simulators which resolve task dependencies under arbitrary mapping schemes, and compute the energy consumption and latency of the optical network during application execution. The simulation results indicate that our TMALPS framework helps eliminate unfavorable traffic patterns through task mapping exploration, and thus significantly improves the effectiveness of runtime laser power reconfiguration techniques. By jointly optimizing the energy consumption and the application execution time of the ONoC, an average of 66 % saving of the energy-delay product is observed for a wide range of application benchmarks, compared to a baseline scenario where the optimization techniques are not applied. Our TMALPS framework also demonstrates better scalability than previous techniques across the evaluated application benchmarks.

The rest of the chapter is organized as follows. In Section 8.2, we introduce our target ONoC architecture and the limitations of existing power reconfiguration strategies. In Section 8.3, we formulate the exploration of task mapping as an integer programming problem and present our optimization framework. In Section 8.4, we summarize the optical power models and the simulation setup employed in this study. In Section 8.5, we evaluate the effectiveness and scalability of our TMALPS framework. And Finally, in Section 8.6, we draw the conclusion of this chapter.

8.2 Background

8.2.1 Optical Network-on-Chip

An ONoC is a collection of optical interconnects that perform on-chip data communication between the processing units of an MPSoC. Various ONoC architectures have been proposed [258], many of which incorporate sophisticated topologies to address specific design considerations, such as higher link utilization, lower crosstalk noise, etc. In this chapter, we target an ONoC architecture based on a generic optical ring bus [257] to demonstrate our TMALPS framework. As illustrated in Fig. 8.1, the processing nodes of the MPSoC are connected to an optical ring bus with a multiple-reader-multiple-writer (MWMR) architecture. Each node has both write and read access to the optical bus, achieved by silicon photonic microring modulators and filters [25], respectively. Wavelength-division multiplexing (WDM) is enabled at each node. Therefore, for a network with n nodes, each with m multiplexed channels, the total number of microring modulators/filters required is $n \cdot m$.

In a typical MWMR configuration, all processing nodes are allowed to utilize all available wavelengths in order to increase the utilization. A wavelength allocation scheme is thus required to ensure mutual exclusion of the selected wavelengths. In this chapter, however, we assume that mutual exclusion is always satisfied, and focus on the joint impact of task

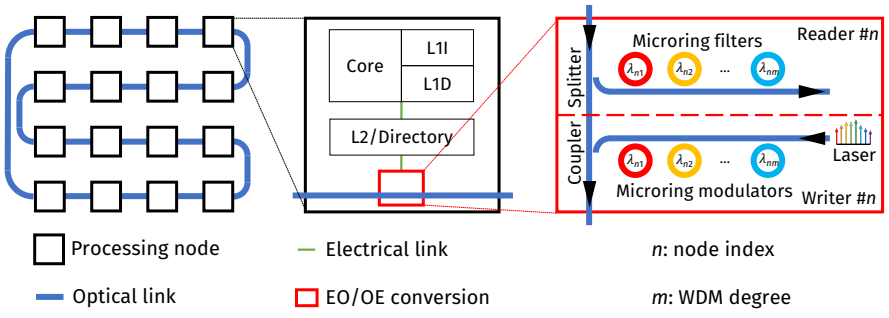


Figure 8.1: Illustration of an MWMR ONoC architecture.

mapping and runtime laser power reconfiguration. The models and assumptions for each network component are detailed in Section 8.4.

8.2.2 Limitations of Existing Strategies

An ON-OFF strategy such as [197] switches the lasers OFF when there is no pending transmission. However, a non-trivial turn-on delay will be introduced when the laser is switched back ON. As per [274], it can take as long as ~10 ns for the laser to stabilize before any data transmission should take place, deteriorating both the energy consumption and the application execution time of the ONoC. DLPS is proposed in [199] as an extension to the ON-OFF strategy, where the laser can operate in one of the four modes, namely OFF, STANDBY, INTERMEDIATE, and FULL-ON. In the STANDBY mode, the laser is biased slightly above the threshold current to reduce the energy consumption yet maintain the capability of a fast turn-on (assumed 1 ns in [199]). In the INTERMEDIATE mode, the laser provides just enough optical power (between zero and the maximum supported) to accommodate the requested data rate. The switching principle of the DLPS strategy is as follows:

- When traffic presents, the laser switches to either the INTERMEDIATE mode or the FULL-ON mode.
- When no traffic presents, the laser switches to the STANDBY mode and stays for at most t_{idle} clock cycles before switching OFF, where t_{idle} is a tunable parameter for different applications.

Fig. 8.2 shows an example of the DLPS switching operations with t_{idle} set to 5 clock cycles. Under the DLPS strategy, the laser will not be switched OFF if the interval between two transmissions is smaller than t_{idle} cycles, and thus the second transmission can start with a faster switching.

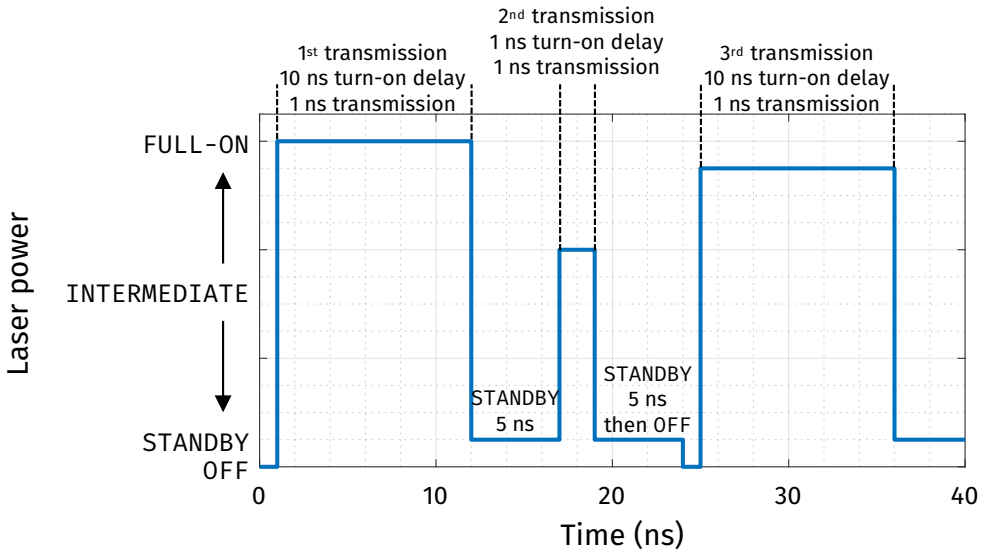


Figure 8.2: Example of DLPS [199] with $t_{\text{idle}} = 5$.

While a larger t_{idle} reduces the occurrences of `ON-OFF` switching, it also consumes energy during the `STANDBY` period. A straightforward upper bound for t_{idle} can be derived as

$$t_{\text{idle}} \leqslant t_{\text{max}} = \frac{\Delta_{\text{off}} \cdot P_{\text{on}} - \Delta_{\text{standby}} \cdot P_{\text{on}}}{P_{\text{standby}}}, \tag{8.1}$$

where Δ_{off} and Δ_{standby} are the laser turn-on delay from the `OFF` mode and the `STANDBY` mode, respectively; P_{on} and P_{standby} are the power consumption of the `FULL-ON` mode and the `STANDBY` mode, respectively. This inequality ensures that the energy consumption of the `STANDBY` period is smaller than that of an `OFF-ON` switching. The authors of [199] notice that the energy saving of the DLPS strategy is no better than that of the `ON-OFF` strategy for many application benchmarks, due to that most of the transmission intervals in these applications are greater than t_{max}. In other words, it is simply better to switch the laser `OFF` immediately after each transmission, rather than wait in the `STANDBY` mode for any period of time.

The lack of access to adjusting the traffic patterns limits the effectiveness of the DLPS

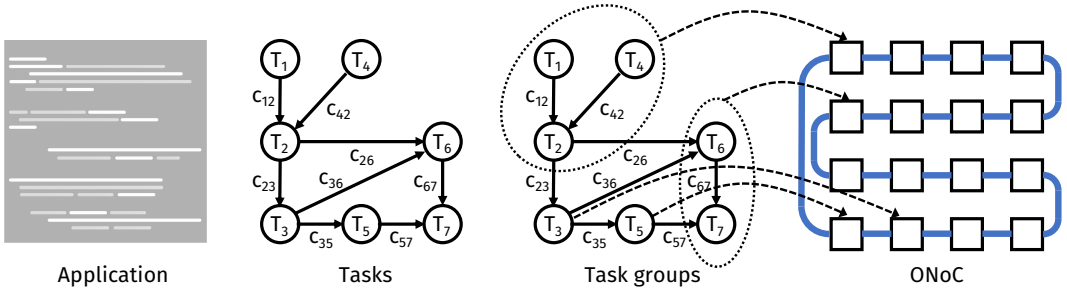


Figure 8.3: Illustration of task partition, grouping, and mapping.

strategy, which motivates our study to incorporate task mapping exploration to help resolve this fundamental limitation.

8.3 Problem Formulation

8.3.1 Mapping Scheme Encoding

The execution of an application on multiple processing units involves several preliminary steps, as illustrated in Fig. 8.3. *Task partition* divides the application into a finite number of tasks. An optional *task grouping* step rejoins certain tasks into task groups, followed by *task mapping* which assigns each task or task group to a processing unit. Each of these steps affects the communication among tasks/task groups, and thus results in different traffic patterns during application execution. Note that in this chapter, we treat task grouping as a special case of task mapping, where multiple tasks are mapped onto the same processing unit.

For an application with T tasks, we encode its mapping scheme into a vector:

$$\mathbf{M}_{1\times T} = \left[p_1, p_2, \ldots, p_T\right], \tag{8.2}$$

where the integer $p_i \in [1, P]$ is the processing unit to which task $\#i$ is assigned, and P is the

total number of processing units. Consecutive tasks are allowed to be assigned to the same processing unit. In such cases, the communication between the two tasks becomes internal to the processing unit and is not carried out by the optical network.

8.3.2 Optimization Framework

In order to explore the impact of task mapping on the effectiveness of runtime laser power reconfiguration strategies, we propose our TMALPS framework illustrated in Fig. 8.4. Our framework takes the application-specific task mapping schemes as the input population. The figure of merit for each mapping scheme is evaluated by two metrics, namely the system energy consumption and the application execution time. Here, instead of having closed-form objective functions, both metrics are computed by simulating the application under the mapping scheme being evaluated. Specifically, a network simulator integrated with laser power reconfiguration features computes the network energy consumption and the latency. The latency information is then fed into an architectural simulator which computes the energy consumption of the cores/cache hierarchies and the overall application execution time. Finally, the overall system energy consumption and the application execution time are the two objectives both seeking to be optimized. The choice of simulators and some implementation considerations are introduced in Section 8.4.3.

The fact that the input mapping schemes are vectors with all integer elements renders our TMALPS a multi-objective integer programming problem. In this chapter, we modify and employ a multi-objective particle swarm optimization (MOPSO) solver [271] to identify a Pareto front of the two objectives, where improvement in either one will require sacrificing the other. The solver also generates new mapping schemes for the next iteration of optimization, where the best schemes in both the current and the past iterations affect how far the new scheme deviates from the current one. Implementation details of the MOPSO

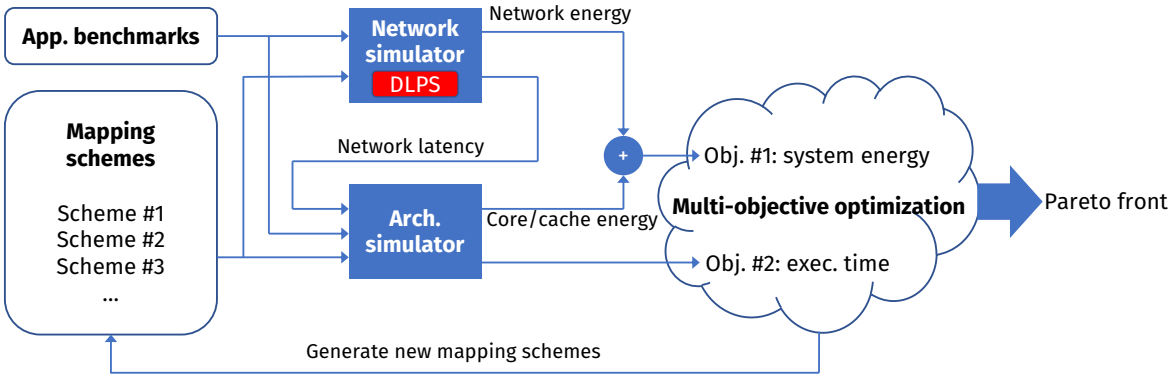


Figure 8.4: TMALPS optimization framework.

solver can be found in [271] and are not further elaborated here.

8.4 Simulation Setup

8.4.1 Laser Turn-On Delay

In order to model the laser turn-on delay for various initial and final power states, we build a SPICE-compatible [82] laser transient model based on coupled rate equations [22] and calibrated it against the measured result of a fabricated laser diode reported in [214]. Fig. 8.5a shows a good match of the simulated and the measured laser output. Fig. 8.5b illustrates the laser turn-on delay for different initial and final power levels. It can be observed that the turn-on delay from the `STANDBY` state to various final states can be limited within sub-nanosecond, while the turn-on delay from zero bias can take up to ~10 ns. Inverse proportionality is observed between the turn-on delay and the final power level, which is also consistent with what reported in [274]. It is noteworthy that the actual delay incurred on the data transmission also depends on the clock rate of the system. For example, the granularity of the control logics under a 1 GHz clock is 1 ns, which means that the data transmission must be delayed by a full nanosecond even if the laser turn-on delay can be smaller than

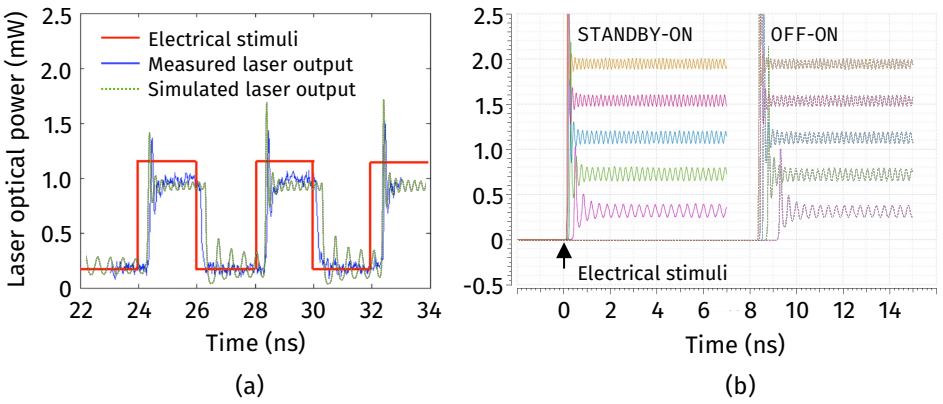


Figure 8.5: (a) Laser transient model, and (b) turn-on delay simulation where different colors correspond to different final states.

that.

8.4.2 Optical Link Power

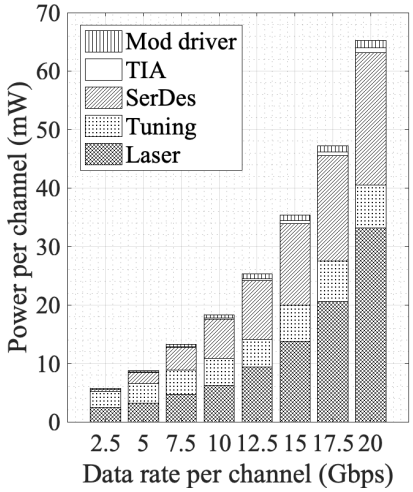
We consider an ONoC based on MWMR optical links introduced in Section 8.2.1. We further assume 5-channel WDM for each node, and a maximum data rate of 20 Gb/s per channel yielding a total of 100 Gb/s supported by each node. When a transmission is requested, a 5-channel optical link is activated between a writer and a reader, each channel with a power consumption of:

$$P_{\text{channel}} = P_{\text{laser}} + P_{\text{mod}} + 2P_{\text{tuning}} + P_{\text{TIA}} + P_{\text{SerDes}}. \tag{8.3}$$

Here, P_{mod}, P_{TIA}, and P_{SerDes} are the power consumptions of the modulator driver, the receiver transimpedance amplifier (TIA), and the serializer/deserializer (SerDes) circuitry, which largely depend on the data rate. A decent analysis of such dependency is given in [230]. P_{tuning} is required for two microrings (a modulator and a filter) to align their resonance wavelengths to the laser wavelength. Low-power tuning techniques are proposed in [105, 173, 176, 225, 226], while a recent method reported in [226], evaluated on measured

Data rate–dependent	
P_{mod}, P_{TIA}, P_{SerDes}	[230]
$P_{\text{sensitivity}}$	[243]
Microring	[226]
Passing loss	0.2 dB
Insertion loss	0.5 dB
Drop-port loss	1 dB
Tuning efficiency	0.15 nm/mW
Laser	
Wall-plug efficiency	20 % [214]
Other	
WG propagation loss	1 dB/cm [243]
Targeted BER	10^{-12}

(a)



(b)

Figure 8.6: (a) Power models and (b) power vs. data rate for an optical channel.

process variations, is adopted for the computation of P_{tuning} in this chapter.

P_{laser} is the electrical power consumed by the laser diode. It subjects to a wall-plug efficiency (WPE) to convert the electrical power into the optical power, which in turn must be high enough to overcome the power losses (PL) along the optical link, and eventually suffice the receiver sensitivity:

$$P_{\text{optical}} \cdot \prod_i \text{PL}_i \geqslant P_{\text{sensitivity}} \tag{8.4}$$

Here, $P_{\text{optical}} = P_{\text{laser}} \cdot \text{WPE}$. $P_{\text{sensitivity}}$ is another data rate–dependent term, for which a reasonable model is provided in [243] with respect to various bit error rate (BER) requirements. A summary of the power models featured in this chapter for optical components is given in Fig. 8.6a, and based on which, the computed power consumption per channel w.r.t. different data rates are plotted in Fig. 8.6b.

Table 8.1: Architectural configurations for evaluating TMALPS.

Instruction set architecture	ARM	alpha
CPU frequency	1 GHz	1 GHz
L1 I/D cache size	32 kB/core	64 kB/core
L2 cache size	512 kB/core	2 MB/core
Cache line size	32 b	64 b
Cache coherency protocol	Directory-based MSI_MOSI	
Technology node (proc. die/mem. die)	40 nm/40 nm	180 nm/90 nm

8.4.3 Simulator Choice

In our study featured in this chapter, the laser turn-on delay is variable throughout the application execution. To the best of our knowledge, such feature is not supported by any readily available network simulators. Therefore, we implemente our in-house network simulator to compute the energy and the latency of the ONoC. For a fair comparison, we validate our network simulator with parameters identical to those reported in [199], and are able to match their results. We then configure our network simulator with the parameters listed in Section 8.4.2 and interface the network latency result to the architectural simulator. We employ JADE [186] as our architectural simulator as it supports customized task mapping. Table 8.1 lists its configurations.

8.5 Evaluation

We evaluate our TMALPS framework on a 64-node ONoC based on the MWMR architecture introduced in Section 8.2.1. Twelve application benchmarks are obtained from the COSMIC Multiprocessor Benchmark Suite [201]. The applications are already partitioned into fine-grained tasks, as summarized in Table 8.2. Each application comes with a default task mapping scheme based on load balancing. Detailed descriptions of the applications can be found in [201].

Table 8.2: Application benchmarks used for evaluating TMALPS.

Synthetic application benchmarks		
Type	**# of tasks**	**# of communications**
Uniform	128	496
Hotspot	128	960
Realistic application benchmarks		
Name	**# of tasks**	**# of communications**
Cifar-10	33	50
FaceRecgonition	33	50
RS-encoder	141	140
RS-decoder	526	789
HPCG	2912	18323
Snap	2299	20844
FFT	15360	24064
Ultrasound	567	44746
RayTracing	25576	45468
MolecularDynamics	4334	194419

The solutions given by our TMALPS framework are compared to three other strategies. The baseline strategy does not include runtime laser power reconfiguration, and uses the default task mapping scheme. A simple `ON-OFF` strategy and the DLPS strategy [199] are also included for comparison, both using the default task mapping scheme as well.

8.5.1 Case Study on RS-encoder

We first show a case study on the RS-encoder application to demonstrate the interpretation of the simulation results. The application task graph is profiled under the ARM instruction set architecture.

The Pareto Front

The output from our TMALPS framework is a Pareto front of the system energy consumption and the application execution time. In Fig. 8.7, we normalize both metrics by the baseline strategy to show the improvement/overhead of various solutions. For RS-encoder, the

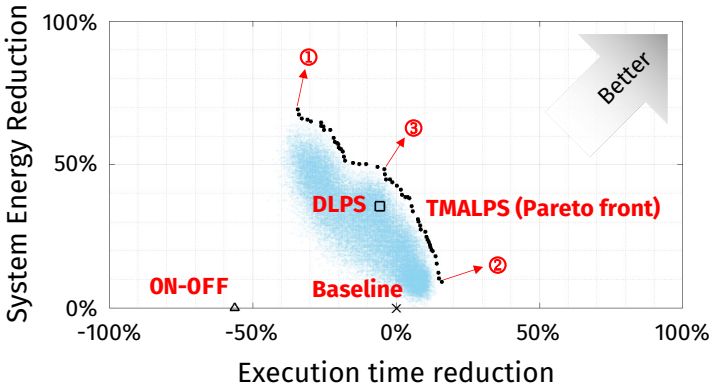


Figure 8.7: Pareto front of system energy reduction and execution time reduction for RS-encoder given by our TMALPS framework.

simple ON-OFF strategy results in a large execution time overhead without providing any energy saving, due to the frequent ON-OFF switching of the lasers. The DLPS strategy achieves considerable energy saving at the cost of a small execution time overhead by leveraging the STANDBY mode. Our TMALPS framework identifies a Pareto front (black dots) among all evaluated mapping schemes (blue shade). As the DLPS solution with the default mapping scheme is among the initial population fed into our optimizer, any new solution, if adopted by the Pareto front, is guaranteed to outperform the DLPS strategy in at least one of the two metrics.

Impact of Task Mapping Exploration

Among the TMALPS solutions in Fig. 8.7, three special cases are highlighted. Solution ① achieves the highest energy saving (69 %) at the cost of 34 % longer execution time. Solution ② reduces the execution time by 16 % while achieving less energy saving, nevertheless, still better than the baseline strategy in both metrics. Solution ③ is one of our TMALPS solutions that outperform the DLPS strategy in both metrics.

In Fig. 8.8a, the system energy consumption and the execution time of these solutions are plotted w.r.t. different t_{idle} values. It can be observed that the curves experience several

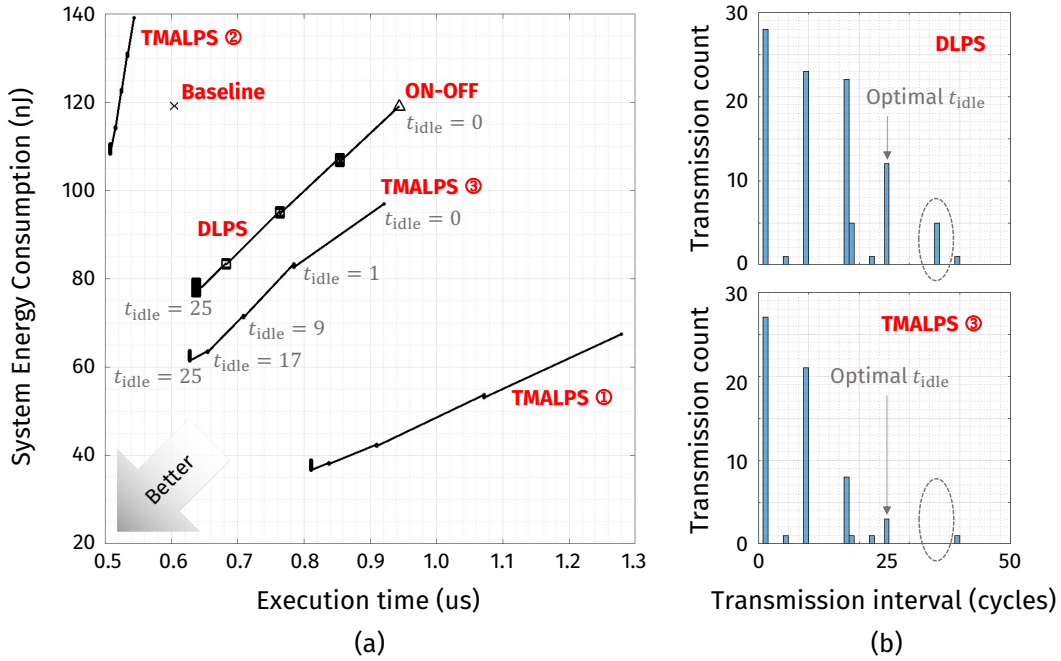


Figure 8.8: Task mapping exploration improves DLPS effectiveness by eliminating unfavorable transmissions.

abrupt drops at $t_{idle} = 1, 9, 17$, and 25, for both the DLPS and our TMALPS solutions. By referencing Fig. 8.8b, we notice that these values correspond to the major bars in the histogram of possible transmission intervals for the RS-encoder application. As soon as t_{idle} reaches one of these critical values, those transmissions with the corresponding interval begin to take advantage of the `STANDBY` mode without the need to switch `OFF` the laser.

The circled bar in Fig. 8.8b indicates that there are transmissions with an interval of 36 clock cycles in the DLPS solution. However, the DLPS energy curve does not drop at $t_{idle} = 36$. This means that the energy consumption of 36 `STANDBY` cycles already exceeds that of an `OFF-ON` switching. While the optimal t_{idle} is still 25 cycles for the RS-encoder, these transmissions with an interval of 36 cycles are always suffering from the `OFF-ON` switching penalty, and thus counteract the effectiveness of the DLPS strategy. By changing the task mapping scheme of the RS-encoder application, our TMALPS solution ③ is able to eliminate such unfavorable transmissions, and thus provide a better optimization result compared to

the DLPS strategy in both optimization targets. Our TMALPS framework also explores other task mapping schemes and provides a set of solutions with different emphasis. A solution with either lower energy consumption or shorter execution time can be identified based on the specific priority of a user.

8.5.2 Evaluation on Other Benchmarks

We evaluate our TMALPS framework across a wide range of application models that are profiled under two instruction set architectures (ISAs), namely ARM and alpha. For each application, we report the normalized energy-delay product (EDP) of different strategies, which are computed as the product of the system energy consumption and the application execution time normalized by that of the baseline strategy. Among the multiple Pareto solutions given by our TMALPS framework, the one with the smallest EDP is chosen to be the representative solution.

Overall EDP Comparison

Fig. 8.9 summarizes the EDP comparison between our TMALPS framework and other strategies. Despite slight differences in some numbers, identical patterns were observed for both ISAs, that our TMALPS solution yields the lowest EDP in 23 out of 24 configurations. Overall, an average of 66 % reduction of EDP can be achieved by our TMALPS framework for the twelve application benchmarks, compared to the baseline strategy where neither task mapping exploration nor laser power reconfiguration is involved.

Complementary Behavior of `ON-OFF` and DLPS

An interesting observation arises when analyzing the second and the third bar of each group. For illustration purpose, we reorder the application benchmarks in Fig. 8.9 in a way

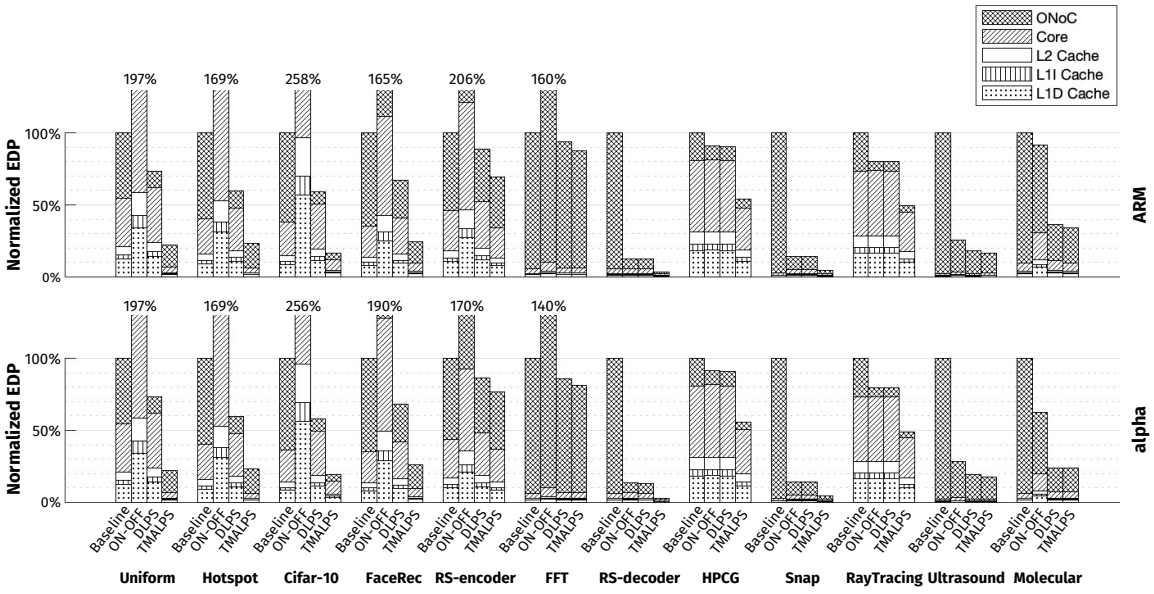


Figure 8.9: Evaluation of our TMALPS framework on 12 application benchmarks profiled under 2 ISAs.

that for the first half of the twelve benchmarks, the ON-OFF strategy performs worse than the baseline. As previously discussed, these applications are traffic intensive for which too frequently switching the lasers OFF could compromise the energy efficiency and the application execution time. The DLPS strategy, on the other hand, performs reasonably well on these applications, as bursty transmissions benefit from the STANDBY mode. For the second half of the benchmarks, the simple ON-OFF strategy works considerably well, which indicates longer intervals between transmissions. Consequently, the improvement of the DLPS strategy over the ON-OFF strategy is quite limited, especially for RS-decoder, HPCG, Snap, and RayTracing. That being said, our TMALPS framework is scalable enough to further optimize the EDP in both scenarios.

Curse of Dimensionality

For an ONoC with P processing units and an application with T tasks, Eq. (8.2) suggests an exponential searching space of $O(P^T)$. Although our TMALPS framework employs

a metaheuristic optimizer which does not require an exhaustive search, an adequate number of iterations should the optimizer run, otherwise, the optimality of the solutions can be impaired. However, the two objectives in our TMALPS framework are evaluated by the execution of multiple simulators, which renders the problem evaluation-expensive for applications with a large number of communications. This limitation starts to surface when evaluated on the MolecularDynamics application, where our TMALPS framework failed to bring improvement over the DLPS strategy under the alpha ISA. In future work, other optimizers should be explored to tackle high-dimensional, evaluation-expensive problems. A hybrid of multiple optimizers may help further enhance the scalability of TMALPS. Alternatively, an online task mapping algorithm for ONoCs is also worth investigation.

8.5.3 Hardware Overhead

The hardware overhead required to support our TMALPS framework mainly comes from idle time tracking. Fortunately, devices on modern MPSoCs are typically controlled by finite state machines, which means that a multiple-bit `STATE` register already exists and can be reused to indicate the device idle time. The `STATE` register itself only reflects the instantaneous state of the device. In order to extract time information, the authors of [280] propose a simple hardware modification to map the multi-bit `STATE` register into a single-bit busy/idle indicator. The indicator is periodically polled by the processor running a power management software. The indicator sets to *busy* whenever there are pending tasks; it resets to *idle* only when the `STATE` register shows idle AND a polling signal has been received. Under such logic, if polling of the indicator returns *idle*, the device must have stayed in the idle state for at least one polling period. The implementation of the busy/idle indicator only requires several tens of logic gates, as reported in [280]. By setting the polling period to t_{idle} in this study, we are able to keep track of the transmitters that have been idle for threshold

cycles and trigger the turn-off signal.

8.6 Concluding Remarks

In this chapter, we propose TMALPS, a Task Mapping–Assisted Laser Power Scaling framework for optical network-on-chips. Our TMALPS framework combines task mapping exploration and runtime laser power reconfiguration to optimize the energy consumption and the application execution time of the ONoC. Simulation results across various application benchmarks show that our TMALPS framework is able to achieve an average of 66 % saving of the system energy-delay product. With the assistance from task mapping exploration, our TMALPS framework scales well to many situations where previous techniques fail to bring improvement. The hardware overhead required to support our TMALPS framework can be well-controlled with intelligent use of software-level techniques and existing registers accessible to on-chip devices.

Chapter 9

POLESTAR: Power Level Scaling with Traffic-Adaptive Reconfiguration

Silicon microring–based optical interconnects offer great potential for high-bandwidth data communication in future data centers and high-performance computing (HPC) systems. However, a lack of effective runtime power management strategies for optical links, especially during idle or low-utilization periods, is devastating to the energy efficiency and the energy proportionality of the network. In this chapter, we propose POLESTAR, i.e., POwer LEvel Scaling with Traffic-Adaptive Reconfiguration, for microring-based optical interconnects. POLESTAR offers a collection of runtime reconfiguration strategies targeting the power states of the lasers and the microring tuning circuitry. The reconfiguration mechanism of the power states is traffic-adaptive for exploiting the trade-off between energy saving and application execution time. The evaluation of POLESTAR with production data-center traces demonstrates up to 87 % reduction in pJ/b consumption and significant improvements in energy proportionality metrics for optical links and networks, notably outperforming existing strategies.

9.1 Introduction

The recent explosive growth of data-driven artificial intelligence (AI) applications has triggered the convergence between data centers and HPC systems in terms of performance requirements [281]. As the computational capability continuously improves through hardware parallelism and specialization, the bottleneck of system performance is gradually shifting from computation to communication [3]. According to the latest technology projections, the bandwidth capacity provisioned for intra–data center/HPC interconnects has exceeded hundreds of Gb/s, a data rate at which traditional electrical interconnects become uneconomical [4, 31]. As a result, optical interconnects are expected to replace electrical ones in both data centers and HPC systems with a growing trend toward shorter reach [1].

Silicon photonics is considered a scalable and cost-effective technology for implementing optical interconnects with a CMOS-compatible fabrication process [13]. In particular, optical links based on quantum-dot (QD) comb lasers [21] and silicon microring resonators (MRRs) [25] have drawn increasing attention for achieving dense wavelength-division multiplexing (DWDM) within compact footprints [29]. Innovations at device, link, and system levels have been reported to advocate microring-based interconnect solutions for future data centers and HPC systems [282, 283, 284].

Unlike the bandwidth capacity for which technologies beyond 1 Tb/s are already under active investigation [6, 7], the energy issues of microring-based optical interconnects have long remained challenging [285]. Despite recent advances in device design [33, 34, 35] and link-level power mitigation techniques [105, 173, 174, 225, 226, 286, 287] that are pushing the best-case energy efficiency of an individual link toward ~1 pJ/b, the effective energy efficiency of an interconnected network is often far from this optimum due to traffic dynamics [4]. Moreover, failure to properly manage the link power during idle or low-utilization periods is devastating to the energy proportionality of the network [37]. Given these issues, the

following major contributors to the energy consumption of microring-based optical interconnects must be addressed by system-level reconfiguration strategies at application runtime:

Static power consumed by the continuous-wave (CW) lasers and the microring tuning circuitry can take up over 80 % of the link power consumption, according to a recent analysis of microring-based DWDM links [32]. The static power is inevitable as long as the link remains on, even if it is not transmitting data. Due to the burstiness of traffic in data centers and HPC systems, the interconnects can often stay idle for relatively long periods [38]. As a result, both the lasers and the microring tuning circuitry need runtime power reconfiguration to avoid excessive waste of energy during idle periods.

Bandwidth overprovisioning is a common practice in data centers and HPC systems [288]. This naïve strategy aims to avoid data-starved computation nodes by deploying optical links that can accommodate the peak bandwidth requirement at the cost of higher communication power [289]. However, as the laser power drastically increases at higher data rates, the optimal energy-per-bit consumption of an optical link often occurs at a data rate slower than the peak [243]. Meanwhile, the skewed (spatially non-uniform) traffic patterns in data centers and HPC systems [39] often result in underutilized links in certain parts of the network. Therefore, it is unwise to always use the maximum data rate for all network activities.

Inspired by the dynamic voltage and frequency scaling (DVFS) techniques that are commonly adopted by the computational hardware (e.g., CPUs and GPUs) [290], runtime reconfiguration of power states can also be applied to the lasers and the microring tuning circuitry to save energy [36]. However, special considerations must be taken regarding the reconfiguration delay of the optical components, such as the turn-on delay of the lasers [274] and

the stabilization time of microring tuning [291]. The reconfiguration delay harms the application execution time and incurs extra energy consumption that could offset the energy saving. Therefore, the reconfiguration strategies must be adaptive to the runtime traffic patterns to avoid unnecessary changes of the power states to the maximum extent.

Given such design considerations, we propose POLESTAR, i.e., POwer LEvel Scaling with Traffic-Adaptive Reconfiguration, for microring-based optical interconnects in data centers and HPC systems. To be elaborated in Section 9.3, POLESTAR offers a collection of runtime power reconfiguration strategies designed around the following objectives:

1) reducing the energy consumption of idle links by switching the lasers and the microring tuning circuitry to *off* or some *low-power* states;

2) improving the energy efficiency of active links by using an *intermediate* (as opposed to the maximum) data rate for select network activities; and

3) minimizing the overhead to the application execution time by making the reconfiguration mechanism of the power states *traffic-adaptive*.

We evaluate the effectiveness of POLESTAR for representative network topologies with an event-driven simulator modified from [194] and [292]. The network in simulation is driven by production data center traces from Alibaba, Inc., containing task execution details of ~4000 machines over eight days [293]. The simulation results demonstrate a 72 % to 87 % reduction in pJ/b consumption and significant improvements in energy proportionality metrics for optical links and networks, notably outperforming existing strategies.

The rest of this chapter is organized as follows. In Section 9.2, we provide an overview of microring-based optical interconnects and their energy issues, as well as a brief discussion of related work. In Section 9.3, we present the detailed design of POLESTAR and analyze the motivation for each power reconfiguration strategy featured in it. In Section 9.4, we

describe the simulation setup for evaluating the effectiveness of POLESTAR. In Section 9.5, we demonstrate the evaluation of POLESTAR with production data-center traces. And finally, in Section 9.6, we draw the conclusion of this chapter.

9.2 Background

9.2.1 Overview of Microring-Based Optical Interconnects

The communication links in data centers and HPC systems connect the computation nodes to their entry-point routers and the routers among themselves. As of today, optical links have dominated the interconnects above the rack-to-rack level and started to penetrate the intra-rack regime [4]. Silicon microring–based optical interconnects are made up of active links enabled by optical transceivers. As illustrated in Fig. 9.1, a microring-based optical transceiver (TRx) achieves DWDM communication by deploying cascaded microrings along a shared waveguide. At the transmitter (Tx) side, each microring modulator modulates a specific wavelength at its resonance. At the receiver (Rx) side, a corresponding microring filter couples the signal out for detection. Concurrent transmission of multiple channels has demonstrated an aggregated data rate of 400 Gb/s [5], and technologies toward 1 Tb/s are under active investigation [6].

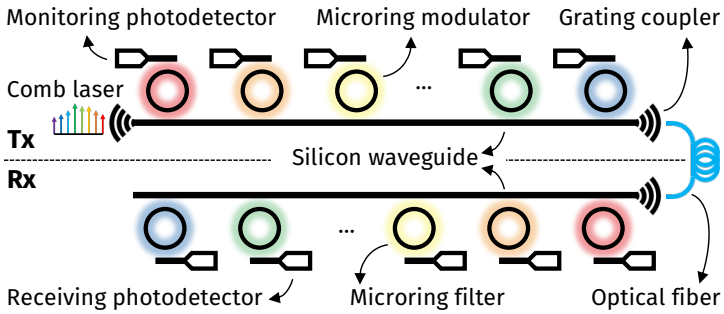


Figure 9.1: Illustration of a silicon microring–based optical transceiver.

9.2.2 Energy Issues of Microring-Based Optical Interconnects

Energy Efficiency

The energy efficiency of optical interconnects is usually measured in pJ/b. However, in most literature, this metric is computed as mW/ (Gb/s), reflecting the power required to attain a target data rate, which has a unit equivalent to pJ/b as Watt = J/s. In this chapter, we refer to the above power-oriented metric as the *nominal energy efficiency* to distinguish it from the *effective energy efficiency*, the latter measuring the actual energy consumption associated with data movement.

The *nominal energy efficiency* of a microring-based optical link heavily relies on the power consumption of three components, namely the laser, the microring tuning circuitry, and the electrical driver circuitry:

$$E_{\text{nom}} = \frac{P_{\text{laser}} + P_{\text{tuning}} + P_{\text{driver}}}{m \cdot \text{DR}}. \tag{9.1}$$

Here, m is the number of DWDM channels, and DR is the target data rate per channel. Due to the process variations that deviate the resonance wavelengths of the microrings from their design values [269], P_{tuning} is required to thermally tune the microrings and align the Tx/Rx channels to a mutual set of carrier wavelengths. As analyses show that the tuning power can take over half of the link power consumption [32], many link-level optimization techniques were proposed to improve the nominal energy efficiency of optical links by reducing the tuning power [105, 173, 174, 225, 226, 286, 287].

The *effective energy efficiency*, on the other hand, measures the actual energy consumed by the optical interconnects to transfer a total number of bits during the entire timespan of

application execution:

$$E_{\text{eff}} = \frac{\text{Energy consumption}}{\text{\# of bits transferred}}. \tag{9.2}$$

In the presence of traffic dynamics, an optical link without proper power management may still consume energy when it is idle. As a result, the effective energy efficiency of an optical network can be orders of magnitude worse than the nominal energy efficiency of individual links [4]. Several techniques were proposed to reconfigure the laser power at application runtime [197, 198, 199, 294]. However, these techniques target the optical network-on-chip (ONoC) [8], where the traffic patterns are significantly different from off-chip scenarios. For ONoCs, the inter-arrival time between two data transmission requests usually ranges from nanoseconds to hundreds of nanoseconds [200, 201], much smaller than the thermal time constants of microring tuning (~1 µs to ~1 ms) [202, 203, 204]. Thus, power reconfiguration for the microring tuning circuitry was deemed unnecessary for ONoCs. However, for data centers and HPC systems where the links can often stay idle for milliseconds to seconds [38], such reconfiguration capability becomes imperative for the effective energy efficiency of the optical interconnects.

Energy Proportionality

The energy proportionality of data-center/HPC interconnects is also becoming more critical as the computation nodes become more energy-proportional over the years [295]. As illustrated in Fig. 9.2, the energy proportionality can be measured by various metrics, such as the *idle-to-peak ratio* (IPR) [296]:

$$\text{IPR} = P_{\text{idle}} / P_{\text{peak}}, \tag{9.3}$$

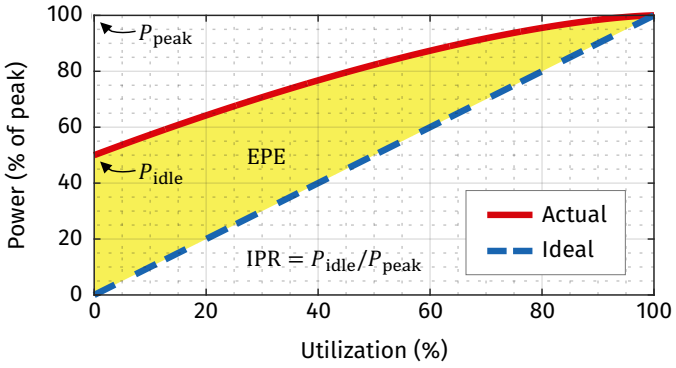


Figure 9.2: Illustration of energy proportionality metrics.

and the *energy proportionality error* (EPE) [297]:

$$\text{EPE} = \int_0^1 |P(u) - u| \, \mathrm{d}u, \tag{9.4}$$

where $u \in [0,1]$ is the utilization of the network, and $P(u) \in [0,1]$ is the normalized network power as a function of the utilization level. The energy proportionality of the optical interconnects also relies on proper management of the link power during idle or low-utilization periods.

9.2.3 Scope of This chapter

Our POLESTAR strategies aim to improve the *effective* energy efficiency and the energy proportionality of the optical interconnects for data-center/HPC applications by reconfiguring the power states of the lasers and the microring tuning circuitry on the fly. Note that another line of work on energy-efficient optical interconnects focuses on connectivity reconfiguration, which lets busy links borrow bandwidth from idle ones to reduce the need for bandwidth overprovisioning [205, 206]. POLESTAR is orthogonal to and can be applied on top of these techniques because power reconfiguration of optical devices is always applicable as long as traffic dynamics exist.

9.3 Strategy Design and Motivation Analysis

POLESTAR features a collection of power reconfiguration strategies for optical interconnects designed with the following questions in mind:

1) What are the power states that a link can switch to when it becomes idle?

2) What data rate should be assigned when a data transmission request is received?

3) How can the reconfiguration mechanism of the power states adapt to the traffic patterns that are spatially non-uniform and constantly changing?

We thus elaborate on the design of our POLESTAR strategies in three steps.

9.3.1 Power Reconfiguration for Idle Links

Due to the turn-on delay of the lasers and the microring tuning circuitry, it is unwise to immediately turn off all the components of an optical link as soon as it becomes idle, in case there is an upcoming transmission request shortly after. POLESTAR extends the laser power scaling concept proposed in [199] into a fine-grained power reconfiguration strategy that includes both the lasers and the microring tuning circuitry. As summarized in Table 9.1 and illustrated in Fig. 9.3, besides `ON` and `OFF`, two additional states are introduced to the optical link, namely `READY` and `STANDBY`. The switching between the power states depends on how long the link has remained idle, where two threshold values, t_1 and t_2, come into play:

- When the link becomes idle, it is first switched to the `READY` state from the `ON` state by reducing the laser bias current to its threshold. At this state, the laser consumes significantly less power with only spontaneous emission and maintains the capability of a fast turn-on. The reconfiguration delay from `READY` to `ON` is roughly proportional to the differential carrier lifetime of the laser and in the order of several nanoseconds [274].

Table 9.1: Available power states for idle links.

Idle time	Power state	Laser bias	MRR tuning	Turn-on delay
0	ON	$> I_{\text{th}}$	On	-
$(0, t_1]$	READY	$\sim I_{\text{th}}$	On	Small
$(t_1, t_2]$	STANDBY	0	On	Medium
$(t_2, +\infty)$	OFF	0	Off	Large

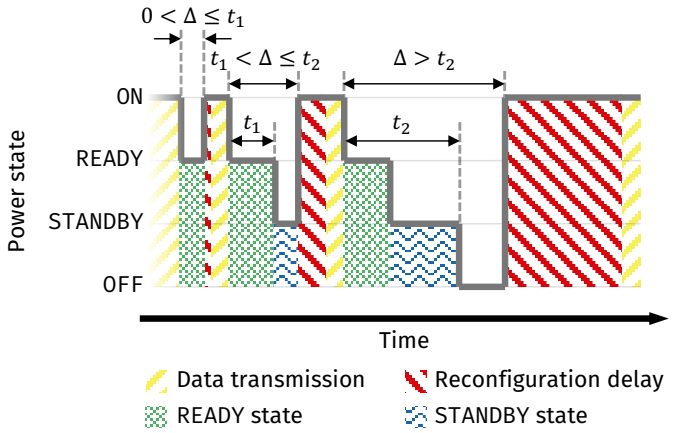


Figure 9.3: Power state reconfiguration for idle links (not drawn to scale for illustration purpose). Δ denotes the idle time since the last transmission.

- When the link has remained idle for longer than t_1, it is then switched to the STANDBY state by turning the laser bias off. The reconfiguration delay from STANDBY to ON is roughly proportional to the total carrier lifetime of the laser and up to ~100 ns [18].

- When the link has remained idle for longer than t_2 ($t_2 \geqslant t_1$), it is finally switched to the OFF state by suspending the microring tuning circuitry. The reconfiguration delay from OFF to ON is dominated by the thermal time constants of microring tuning (up to ~1 ms [202, 203, 204]).

It is further assumed that during the reconfiguration delay, the link consumes the same power as that of the final state but cannot transmit data. Note that in this chapter, the time required for clock recovery and frame synchronization is not counted toward the reconfiguration delay. In contrast to traditional electrical interconnects that maintain synchro-

nization between two connected ports by filling idle periods with repetitive patterns, optical interconnects for data centers and HPC systems usually employ burst-mode receivers to perform synchronization of clock and data amid bursty traffic [298]. Such receiver technologies prepend some carefully designed preamble bits to the data packets and can achieve fast synchronization of clock and data within several nanoseconds [299, 300]. The synchronization time is considered as a small overhead to the packet transfer time independent of the initial power state of the optical link.

As shown in Fig. 9.3, the state transition profile of a link during the idle period is determined by the relationship between Δ (i.e., the idle time since the last transmission) and the values of t_1 and t_2. If $t_1 = t_2 = 0$, the strategy reduces to simple `ON-OFF` reconfiguration that switches the link off as soon as it becomes idle. While positive t_1 and t_2 could benefit some transmissions with reduced turn-on delay, both thresholds cannot be infinitely large as the `READY` and `STANDBY` states themselves also consume energy. Therefore, t_1 and t_2 should be made adjustable at application runtime for the constantly changing Δ, in other words, traffic-adaptive. Section 9.3.3 will further elaborate on this design objective.

9.3.2 Power Reconfiguration for Active Links

Besides providing multiple power states for idle links, POLESTAR also features two strategies that take effect when a link receives a data transmission request and becomes active. The first strategy seeks to improve the link energy efficiency by using an *intermediate data rate* (as opposed to the maximum supported) for select network activities. As illustrated in Fig. 9.4, due to the highly nonlinear growth of laser power consumption at higher data rates, the optimal pJ/b of an optical link could occur at a data rate slower than the peak, which we denote by DR_{opt}. (The power models for the optical link will be revisited in Section 9.4.3.) In this chapter, we propose to use DR_{opt} for transmitting control messages (whose volume is

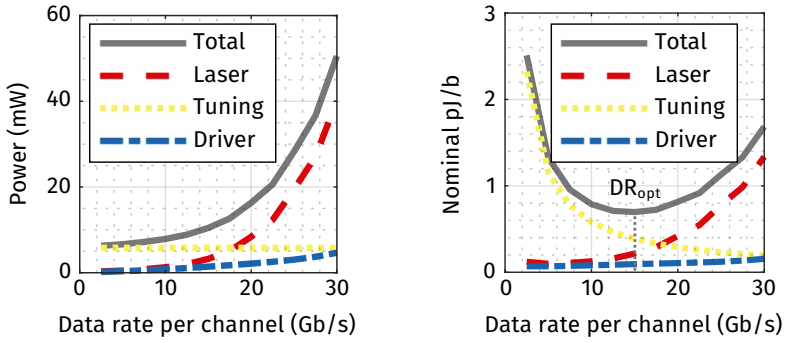


Figure 9.4: (a) Power consumption and (b) nominal energy efficiency of an optical link as functions of the data rate per channel.

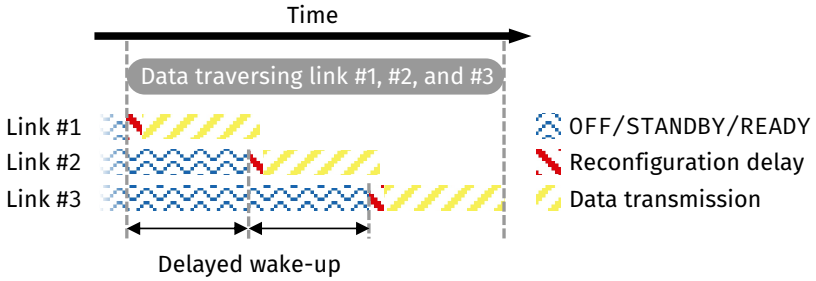


Figure 9.5: Illustration of delayed wake-up of downstream links for a network activity traversing multiple links.

usually orders of magnitude smaller than the actual data) and data that do not serve as the input to any pending tasks. According to our observation from the Alibaba cluster traces, these messages can take up over 25 % of all network activities, offering a reasonable opportunity for energy optimization.

Another strategy accounts for the scenario where a network activity traverses multiple links. This is common for data-center/HPC interconnects as the network topology is usually hierarchical rather than point-to-point. As illustrated in Fig. 9.5, when a data transmission request is received, instead of immediately waking up all the links en route, POLESTAR wakes

up a downstream link with a delay d computed as

$$d_i = \max\left(0, \sum_{j=1}^{i-1} s/\mathrm{DR}_j - \delta_i\right), \tag{9.5}$$

where i denotes the sequential position of the target link en route, s denotes the data size, DR_j denotes the data rate assigned to the data by link $\#j$, and δ_i denotes the reconfiguration delay corresponding to the current power state of link $\#i$. Eq. (9.5) ensures that only the reconfiguration delay of the first link will affect the overall communication time, and those of the downstream links can be already in the `ON` state by the time the data arrives.

9.3.3 Making the Reconfiguration Mechanism Traffic-Adaptive

As the traffic patterns vary from link to link and changes with time, POLESTAR features a mechanism that adjusts the values of the idle thresholds, t_1 and t_2, at application runtime. Referring to Fig. 9.3, an upper bound for t_1 can be calculated from

$$P_{\mathrm{READY}} \cdot t_1 + P_{\mathrm{ON}} \cdot \delta_{\mathrm{READY}} \leqslant P_{\mathrm{STANDBY}} \cdot t_1 + P_{\mathrm{ON}} \cdot \delta_{\mathrm{STANDBY}}, \tag{9.6}$$

which gives

$$t_1 \leqslant t_{1,\max} = \frac{P_{\mathrm{ON}}\,(\delta_{\mathrm{STANDBY}} - \delta_{\mathrm{READY}})}{P_{\mathrm{READY}} - P_{\mathrm{STANDBY}}}, \tag{9.7}$$

where P_* is the link power consumption of state $*$, and δ_* is the turn-on delay of the link from state $*$. In other words, for an idle time Δ greater than $t_{1,\max}$, it is rather less energy-consuming to skip the `READY` state and directly use the `STANDBY` state for the entire idle period, despite a larger turn-on delay. Similarly, an upper bound for t_2 can be calculated

from

$$P_{\text{READY}} \cdot t_1 + P_{\text{STANDBY}} (t_2 - t_1) + P_{\text{ON}} \cdot \delta_{\text{STANDBY}} \leqslant P_{\text{ON}} \cdot \delta_{\text{OFF}}, \tag{9.8}$$

which gives

$$t_2 \leqslant t_{2,\max} = \frac{P_{\text{ON}} (\delta_{\text{OFF}} - \delta_{\text{STANDBY}}) - (P_{\text{READY}} - P_{\text{STANDBY}}) t_1}{P_{\text{STANDBY}}}. \tag{9.9}$$

For an idle time Δ greater than $t_{2,\max}$, it becomes less energy-consuming if the link remains `OFF` for the entire idle period despite the even larger `OFF-ON` delay.

An ideal mechanism is expected to predict the next Δ and adjust the idle thresholds to ensure that

$$\begin{cases} \Delta \leqslant t_1 \leqslant t_{1,\max}, & \text{if } \Delta \in \left(0, t_{1,\max}\right], \quad (9.10a) \\ t_1 = 0, & \text{if } \Delta \in \left(t_{1,\max}, +\infty\right); \quad (9.10b) \end{cases}$$

and

$$\begin{cases} \Delta \leqslant t_2 \leqslant t_{2,\max}, & \text{if } \Delta \in \left(0, t_{2,\max}\right], \quad (9.11a) \\ t_2 = 0, & \text{if } \Delta \in \left(t_{2,\max}, +\infty\right). \quad (9.11b) \end{cases}$$

However, it is impossible to predict the exact length of an upcoming idle period. Moreover, as the runtime adjustment of t_1 and t_2 is local to each link, it is desirable that the implementation could be done at the router level with simple hardware logic and require no centralized management or sophisticated software support. To this end, we propose a simplified mechanism for the runtime adjustment of the idle thresholds. Taking the adjustment of t_1 as an example (the adjustment of t_2 follows the same principle), we define that an idle

period Δ is

$$\begin{cases} \textit{in-range}, & \text{if } \Delta \in (0, t_{1,\max}]; \quad (9.12a) \\ \textit{out-of-range}, & \text{if } \Delta \in (t_{1,\max}, +\infty). \quad (9.12b) \end{cases}$$

Then, by recording this piece of information for historic idle periods, the simplified mechanism predicts whether the next Δ will be in-range or not, instead of predicting its exact length. The mapping of Δ into binary states enables us to explore simple digital logic for adjusting the idle thresholds, inspired by the extensively-studied branch prediction strategies in the computer architecture domain [301]. Specifically, in this chapter, we compare two types of implementations for adjusting t_1 and t_2 based on the prediction of Δ range:

One-level idle threshold adjustment Inspired by the existing one-level branch prediction techniques [301], the one-level adjustment mechanism for idle thresholds maintains an n-bit saturating up-down counter. If the last recorded idle period (denoted by Δ_{last}) is in-range (Eq. (9.12a)), the counter increases by one (and saturates at $(2^n - 1)$); otherwise, the counter decreases by one (and saturates at 0). Then, the idle thresholds, t_1 and t_2, are updated based on the following criteria:

1) If the counter is greater than $(2^n - 1)/2$, an operation called *match* is performed:

$$\begin{cases} t_1 \leftarrow \min\left[\max\left(t_1, \Delta_{\text{last}}\right), t_{1,\max}\right], & (9.13a) \\ t_2 \leftarrow \min\left[\max\left(t_2, \Delta_{\text{last}}\right), t_{2,\max}\right]; & (9.13b) \end{cases}$$

2) otherwise, if the counter is smaller than $(2^n - 1)/2$, an operation called *reset* is performed:

$$t_1, t_2 \leftarrow 0. \quad (9.14)$$

The one-level idle threshold adjustment mechanism based on the n-bit counter can

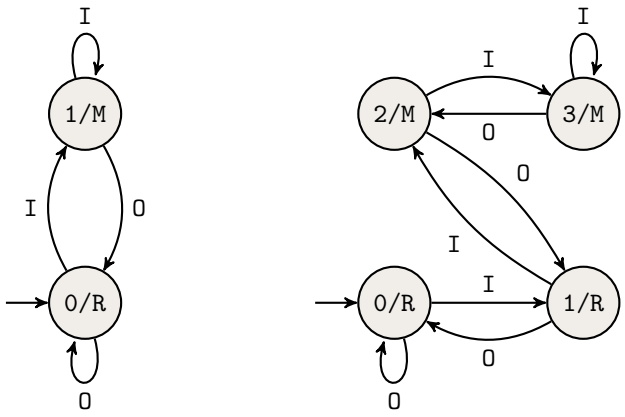


Figure 9.6: State diagrams of the one-level idle threshold adjustment mechanism using (a) a 1-bit saturating counter and (b) a 2-bit saturating counter, where I: Δ_{last} in-range (Eq. (9.12a)); O: Δ_{last} out-of-range (Eq. (9.12b)); M: match operation (Eqs. (9.13a) and (9.13b)); and R: reset operation (Eq. (9.14)).

be represented by finite-state machines. Fig. 9.6 shows the cases for $n = 1$ and $n = 2$, which are analogous to the 1-bit and 2-bit branch predictors described in [301].

Two-level idle threshold adjustment Inspired by the existing two-level branch prediction techniques [302], a two-level adjustment mechanism for idle thresholds is also implemented for comparison. As illustrated in Fig. 9.7, the first level of the adjustment mechanism is a k-bit *range history register* for historic Δ's. A "1" indicates that the recorded Δ was in-range (Eq. (9.12a)), while a "0" indicates out-of-range (Eq. (9.12b)). The range history of the past k idle periods forms an index to address the *pattern table*, the second level of the adjustment mechanism. In this chapter, each entry of the pattern table is an n-bit saturating up-down counter as described in Fig. 9.6. In total, the two-level adjustment mechanism maintains a k-bit shift register and 2^k n-bit counters. According to [302], as each counter is updated when a unique historic pattern occurs, the 2-level mechanism is beneficial if the patterns are temporally correlated.

Note that another source of hardware overhead is for keeping track of the length of each

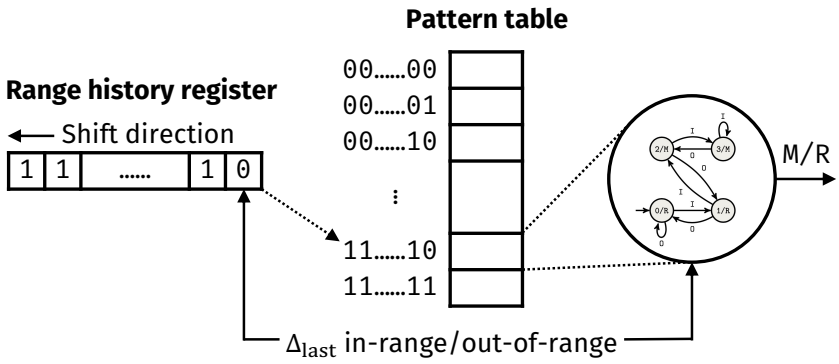


Figure 9.7: Illustration of the 2-level idle threshold adjustment mechanism, where `M` denotes the match operation (Eqs. (9.13a) and (9.13b)) and `R` denotes the reset operation (Eq. (9.14)).

idle period Δ. Hardware-assisted techniques, such as the 1-bit busy/idle register proposed in [280], is able to extract this information with several logic gates per link. A comparison of the runtime adjustment mechanisms for the idle thresholds is conducted in Section 9.5, together with the evaluation of our POLESTAR strategies for various network configurations.

9.4 Simulation Setup

9.4.1 Overview of the Simulation Environment

Dataset

Among various public datasets of data-center/HPC workloads [38], we opted for the traces recorded on a production cluster of Alibaba, Inc. [293] to evaluate our POLESTAR strategies. The Alibaba traces, published in 2018, contain execution details of ~1.3 billion tasks on ~4000 machines over eight days. Besides its recentness and large size, another reason for choosing the Alibaba traces is the inclusion of task dependency information. As illustrated in Fig. 9.8a, a group of dependent tasks (often referred to as a job or a workflow) can be characterized by a directed acyclic graph (DAG) describing the inter-task data de-

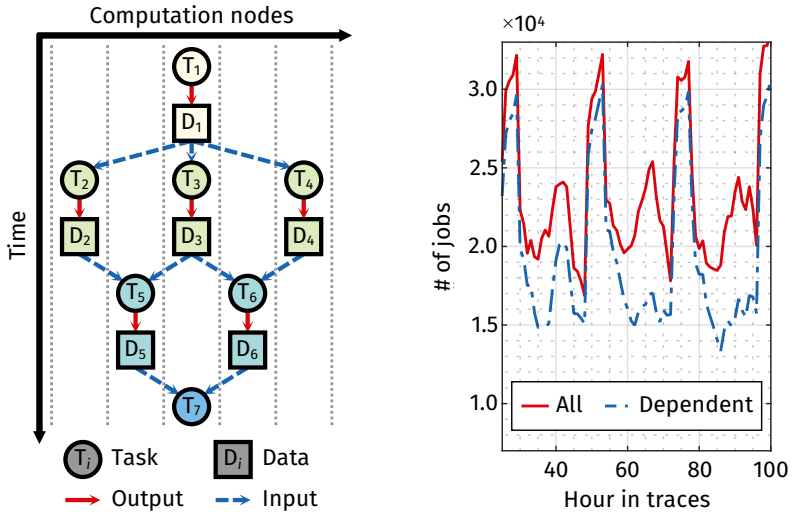


Figure 9.8: (a) DAG representation of a job containing multiple tasks and data dependencies and (b) temporal distribution of jobs in the Alibaba traces.

pendencies. A task is assumed to generate a single piece of output data, which may serve as the input data to multiple child tasks. A task can also depend on the output of multiple parent tasks. As the tasks are distributed to different computation nodes, the data dependencies among the tasks result in communication patterns across the network. As shown in Fig. 9.8b, an average of 80 % of the jobs in the Alibaba traces are dependent ones. The task dependency information enables us to investigate the impact of POLESTAR on application execution time because we can identify the tasks that are subsequently affected by changing the communication.

Simulator

For replaying the Alibaba traces and simulating our POLESTAR strategies in operation, we employed two open-source tools, namely WREHCN [292], a library for workflow management, and SimGrid [194], a matured simulation framework for distributed computing platforms. WREHCN features a DAG processing engine which we tweaked to parse the Alibaba

traces and generate simulation entities recognizable by SimGrid. Then, SimGrid, modified and implemented with our POLESTAR strategies, simulates the task execution and the network communication.

9.4.2 Trace Preprocessing and Simulator Calibration

Synthetic Data Size

The Alibaba traces do not include statistics on the data size or communication time. Instead, only the execution time of each task is recorded. To this end, we generate synthetic data sizes associated with the inter-task data dependencies. Specifically, we denote the execution time of task T_i (Fig. 9.8a) by $t_{\mathrm{exec},i}$ and assume that it consists of two parts: 1) the time spent waiting for data communication from its parent tasks, $t_{\mathrm{comm},i}$; and 2) the time spent on actual computation, $t_{\mathrm{comp},i}$. We then define a parameter ρ, which we refer to as the *communication-to-computation ratio*, and thus

$$t_{\mathrm{comm},i} = \frac{\rho}{1+\rho} t_{\mathrm{exec},i}, \tag{9.15}$$

$$t_{\mathrm{comp},i} = \frac{1}{1+\rho} t_{\mathrm{exec},i}. \tag{9.16}$$

The SimGrid simulator can strictly enforce $t_{\mathrm{comp},i}$ for each task in the simulation, but $t_{\mathrm{comm},i}$ has to be simulated based on the data size information. According to the observation of a linear relationship between the input data size and the computation time for various datacenter applications [303], we denote the size of data D_i (Fig. 9.8a) by s_i and further assume that

$$s_i = a \cdot \sum_{j \in \mathbb{C}_i} t_{\mathrm{exec},j} / \left|\mathbb{P}_j\right| + b, \tag{9.17}$$

where $\mathbb{C}_i$ is the set of child tasks of task T_i, and $\mathbb{P}_j$ is the set of parent tasks of task T_j. Finally, we find proper values for a and b by solving

$$\min_{a,b} \left\| \hat{\boldsymbol{t}}_{\text{comm}} - \boldsymbol{t}_{\text{comm}} \right\|^2, \tag{9.18}$$

where $\hat{\boldsymbol{t}}_{\text{comm}}$ is the simulated values of the communication time for all tasks by replaying the traces in SimGrid, and $\boldsymbol{t}_{\text{comm}}$ is the expected values computed from Eq. (9.15).

In this chapter, we assume $\rho = 0.5$ by default, as it was observed in [304] that the time spent on data communication is roughly half of that spent on computation. However, we also vary the value of ρ between 0.2 and 5 to account for broader scenarios.

Simulator Calibration

We calibrate the simulator by minimizing Eq. (9.18) assuming $\rho = 0.5$. As each evaluation of the function being minimized requires a simulation run, the optimization process is considered evaluation-expensive. Thus, we employed an implementation of the Bayesian optimization algorithm [305] based on Gaussian process models [306] to explore proper values for a and b in Eq. (9.18). As shown in Fig. 9.9a, the optimal values for a and b are found to be 209 MB/s and 56 KB, respectively. Fig. 9.9b shows the cumulative number of finished tasks in an hour by replaying the Alibaba traces in the SimGrid simulator. The simulation results with and without the data size information are both compared to the ground truth as recorded in the traces. The close match between the simulated curve and the recorded one indicates that our simulator is well-calibrated, and the method for generating synthetic data sizes for the Alibaba traces is justified. For $\rho = 0.5$, the synthetic data sizes generated for the packets in one trace hour range from 237 KB to 10.2 GB with a median of 216 MB. The idle time between transmissions simulated for all optical links in the 64-node Fat-Tree cluster has a long-tailed distribution with a median of 43 µs and a maximum of 302 s, indicating

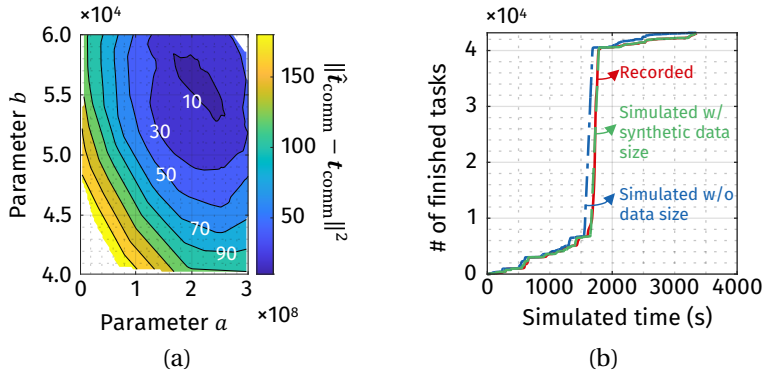


Figure 9.9: Data size generation and simulator calibration: (a) finding proper values for a and b in Eq. (9.17) by optimizing Eq. (9.18); (b) simulated vs. recorded task execution with and without data size information.

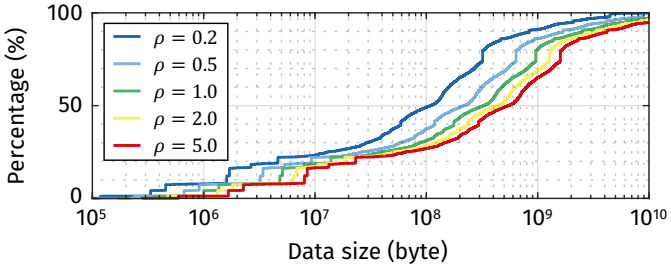


Figure 9.10: Cumulative distributions of data sizes w.r.t. different values of ρ.

traffic burstiness.

The data size information for other values of ρ is generated in the same manner. Fig. 9.10 shows the cumulative distributions of data sizes generated for the Alibaba traces w.r.t. different values of ρ. The overall range and shape of the distributions are comparable to other communication traces used in system-level studies of optical interconnects [206].

9.4.3 Network Configuration

Topology

The network topologies for data centers and HPC systems can be categorized as direct or indirect. In *direct* topologies, every router has computation nodes directly connected

to it, while in *indirect* topologies, some routers are not exposed to the computation nodes and only connect to other routers [4]. In this chapter, we evaluate our POLESTAR strategies for one representative topology from each category. For indirect topologies, we choose *Fat-Tree*, which has been widely adopted in many real-world clusters. For direct topologies, we choose *Dragonfly*, which is promising for future high-throughput data-center/HPC networks [307]. While both topologies have been implemented in the SimGrid simulator, as described in [308], we added a property to the link implementation specifying whether it is electrical or optical. We further configure the links above the first-level routers as optical, which are the reconfiguration targets of POLESTAR.

Power Models

We assume the pairing of a 24-channel comb laser (reported in [283]) and a 24-channel microring-based transceiver (reported in [269]) to form an optical link with a maximum data rate of 30 Gb/s per channel. As Eq. (9.1) in Section 9.2.2 has mentioned, the computation of link energy relies on models for the power of the laser (P_{laser}), the microring tuning circuitry (P_{tuning}), and the electrical driver circuitry (P_{driver}). The power models employed in this chapter are summarized in Table 9.2 and further explained as follows.

Table 9.2: Power models for optical links.

Laser efficiency			
Wall-plug efficiency	5 % [283]	Spectrum efficiency	-3.2 dB [18]
Data rate dependency			
$P_{\text{sensitivity}}$	[243]	P_{driver}	[230]
Microring			
Passing loss	0.2 dB [287]	Drop-port loss	1 dB [287]
Insertion loss	0.5 dB [287]	Tuning efficiency	0.15 nm/mW [242]
Waveguide			
Coupling loss	1 dB [243]	Propagation loss	1 dB/cm [243]

Laser We assume a Gaussian-shaped optical spectrum of the comb laser with a spectrum efficiency $\eta = P_{\text{usable}}/P_{\text{total}} \approx -3.2\,\text{dB}$ [18]. The optical power at the laser output must be high enough so that the following power budget equation holds for any channel $k \in \{1, 2, \cdots, m\}$:

$$P_{\text{comb},k} \cdot \alpha_k \geqslant P_{\text{sensitivity}}. \tag{9.19}$$

Here, m is the number of DWDM channels; $P_{\text{comb},k}$ is the optical power of the kth comb line; $\alpha_k \in (0,1)$ is the accumulated loss of optical power as the light travels along channel k, including the waveguide coupling loss, propagation loss, microring passing loss, insertion loss, and drop-port loss [243, 287]; and $P_{\text{sensitivity}}$ is the sensitivity requirement of the receiver, which is a function of the data rate [243]. The laser subjects to a wall-plug efficiency (WPE) when converting electrical power into optical power, and then the spectrum efficiency (η) that accounts for the usable portion of the comb lines:

$$P_{\text{laser}} \cdot \text{WPE} \cdot \eta = \sum_{k=1}^{m} P_{\text{comb},k}. \tag{9.20}$$

Based on Eqs. (9.19) and (9.20), the laser power consumption can be computed for various data rates, as seen in Fig. 9.4a.

Microring tuning The microring tuning power is estimated from the variation distribution of the resonance wavelengths measured from a wafer fabricated with 66 24-channel transceivers [269]. The transceivers have a channel spacing of ~0.35 nm (~61 GHz in the O-band) and are designed to support up to 30 Gb/s per channel. For lower data rates, we assume the same channel spacing (0.35 nm) despite that denser channels may be used. As a results, the microring tuning power is considered independent of the data rate in this chapter. Fig. 9.4a shows the modeled tuning power assuming a thermal tuning efficiency of 0.15 nm/mW.

Electrical driver The power models for the electrical driving circuitry, including the modulator drivers at the Tx side and the transimpedance amplifiers (TIAs) at the Rx side, are taken from [230], both depending on the target data rate. Note that in this chapter, the serializer/deserializer (SerDes) circuitry is considered part of the computation nodes rather than the link drivers. Thereforbe, its power consumption is not included in the link power. Similar assumptions are found in other literature, such as [32].

Reconfiguration Delay

The reconfiguration delay, described in Section 9.3.1, is a key parameter affecting the trade-off between energy saving and application execution time. In this chapter, we consider four corner cases corresponding to the fast/slow stabilization of the laser/microring tuning [18, 202, 203, 204, 274], as summarized in Table 9.3.

9.5 Evaluation

9.5.1 Comparison of Idle Threshold Adjustment Mechanisms

As reasoned in Section 9.3.3, the capability to adjust the idle thresholds, t_1 and t_2, adaptively to the runtime traffic patterns is vital to the effectiveness of the POLESTAR strategies. Therefore, we first compare the two adjustment mechanisms, namely the one-level and the two-level idle threshold adjustment, in terms of their traffic adaptability. As indicated by

Table 9.3: Corner cases for the reconfiguration delay.

	Corner	FF	FS	SF	SS
Delay	READY-ON	Assumed 1/10 of STANDBY-ON delay			
	STANDBY-ON	10 ns	10 ns	100 ns	100 ns
	OFF-ON	1 µs	1 ms	1 µs	1 ms

Eqs. (9.13a)–(9.14), the adjustment operations for the idle thresholds, namely to match or to reset, are determined based on the historic patterns of idle periods. In our experiments, we compare the operations predicted by both the one-level and the two-level mechanisms to the ground truth theoretically derived from Eqs. (9.10a)–(9.11b) based on the actual length of the upcoming idle period. The prediction accuracy of both mechanisms is calculated as we simulate the Alibaba traces in the SimGrid simulator. Also added for comparison are two static strategies that do not adjust the idle thresholds at application runtime: 1) the `ON-OFF` strategy that switches the link off as soon as it becomes idle, corresponding to $t_1 = t_2 = 0$; and 2) a static strategy that always has $t_1 = t_{1,\max}$ and $t_2 = t_{2,\max}$.

Fig. 9.11 shows the comparison of the prediction accuracy for adjusting t_1, calculated for a specific link in a simulated 64-node Fat-Tree cluster. For the n-bit saturating counter

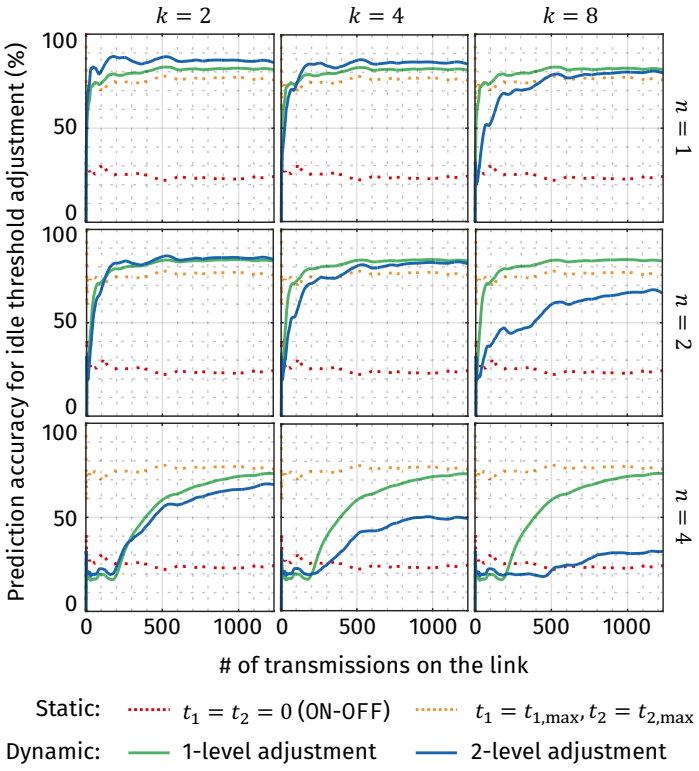


Figure 9.11: Prediction accuracy comparison for various idle threshold adjustment mechanisms and settings.

employed by both the one-level and the two-level idle threshold adjustment mechanisms, n is varied from 1 to 4; for the k-bit register employed by the two-level mechanism, k is varied from 2 to 8. The following observations are made:

- Overall, the two-level idle threshold adjustment mechanism with $k = 2$ and $n = 1$ achieves the best prediction accuracy of ~86 % and the steepest learning curve (fastest learning rate).

- The one-level idle threshold adjustment mechanism with $n = 2$ achieves a prediction accuracy that is comparable to the two-level mechanism in the long term and incurs less hardware overhead as described in Section 9.3.3. The trade-off is the slightly slower learning rate compared to the two-level mechanism with $k = 2$ and $n = 1$.

- The prediction accuracy of the two static strategies is complement to each other, reflecting the the average portion of the idle periods that are in-range/out-of-range (Eqs. (9.12a) and (9.12b)). Both the one-level and the two-level mechanisms with proper settings outperform the static strategies, justifying the necessity of runtime traffic-adaptive adjustment for the idle thresholds.

- Further increasing k beyond 4 and n beyond 2 is unnecessary due to the degraded prediction accuracy and learning rate, as well as the increased hardware overhead.

The evaluation of the adjustment mechanisms for t_2 reaches the same conclusion, and consistency is observed for various links in the simulated network. In the following experiments, we opt for the one-level idle threshold adjustment mechanism with $n = 2$ due to its simpler implementation with negligible performance loss.

9.5.2 Case Study for Strategy Effectiveness

We then conduct a case study of our POLESTAR strategies for a simulated Fat-Tree cluster with 64 nodes to evaluate its effectiveness in energy optimization. The SS corner in Table 9.3 is assumed for conservativeness. A one-hour segment of the Alibaba traces is used for stressing the network with the job arrival rate down-sampled to match the 64 nodes.

Improvement of Effective Energy Efficiency

The effective energy efficiency can be calculated for the overall network or for each individual link, using Eq. (9.2). In Fig. 9.12a, we first show the improvement of effective pJ/b of the network achieved by different strategies compared to a baseline scenario where the links are always kept on. Among existing strategies [197, 198, 199, 294] that only consider the laser power as a tuning knob, we include the Dynamic Laser Power Scaling (DLPS) strategy [199] for comparison. Note that DLPS also proposes to use an intermediate data rate for transmissions that can finish within a clock cycle. However, such transmissions are not observed in the Alibaba traces where the data size is significantly larger than the on-chip scenario discussed in [199]. As observed in Fig. 9.12a, POLESTAR is able to reduce the effective pJ/b of the network by ~85 % when all of its featured strategies are enabled (the rightmost bar), notably outperforming DLPS (the leftmost bar). This indicates the necessity of extending the power reconfiguration mechanism to include the microring tuning circuitry in data-center/HPC interconnects. Among the POLESTAR strategies, the power reconfiguration for idle links contributes the most energy saving (~57 %), followed by the delayed wake-up (~20 %), and finally the intermediate data rate (~8 %). The second bar corresponds to the `ON-OFF` strategy where $t_1 = t_2 = 0$. The fact that each data transmission must start with an `OFF-ON` delay results in considerably less energy saving. The third bar corresponds to the static case where $t_1 = t_{1,\max}$ and $t_2 = t_{2,\max}$. The lack of traffic adaptability for the idle thresholds also results

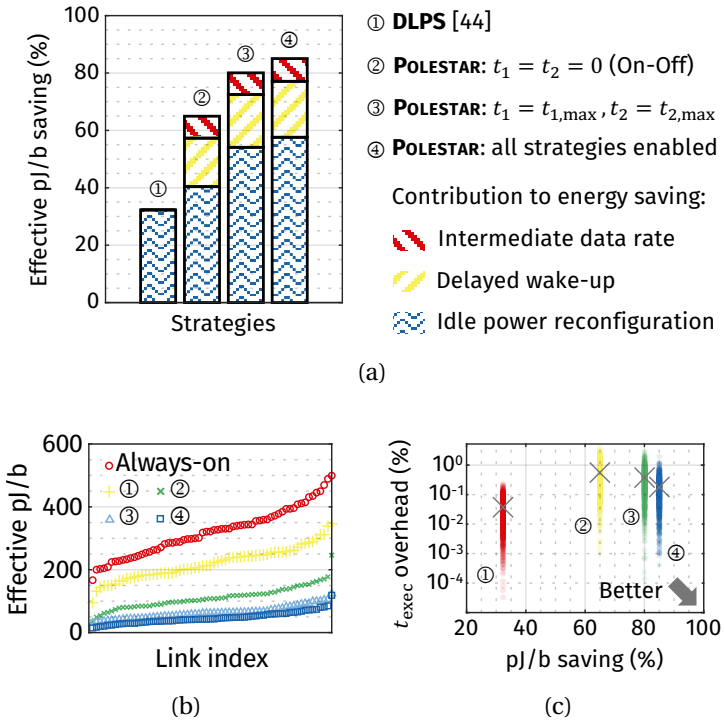


Figure 9.12: Case study of POLESTAR for a 64-node Fat-Tree topology: (a) improvement of effective energy efficiency for the network; (b) effective energy efficiency for individual links; and (c) trade-off between energy saving and application execution time.

in slightly less energy saving compared to POLESTAR at full play.

In Fig. 9.12b, we plot the effective pJ/b for each individual link in ascending order for different reconfiguration strategies. Even though the links in the network are identical and share the same nominal energy efficiency, their effective pJ/b could be vastly different due to the unbalanced traffic patterns across the network. For the baseline and the DLPS strategy where the links are never turned off, it is especially devastating to the effective pJ/b of the low-utilization links due to the small number of bits transmitted. The flattest curve in Fig. 9.12b belongs to POLESTAR with all features enabled, indicating that the POLESTAR strategies are particularly effective for managing the energy of idle and low-utilization links.

Overhead to Application Execution Time

Fig. 9.12c demonstrates the trade-off between energy saving and application execution time observed in this case study. The horizontal axis is again the saving of effective pJ/b of the network for each strategy compared. The vertical axis plots the overhead to the execution time of all jobs, where the cross signs indicate the mean values. Intuitively, the DLPS strategy incurs the smallest overhead as it only involves laser reconfiguration. However, the energy saving achieved by DLPS is far from ideal. Our POLESTAR strategies, on the other hand, can significantly improve the energy saving while still keeping the overhead to the application execution time manageable. Notably, POLESTAR with traffic adaptability enabled (④ in Fig. 9.12c) can limit the average overhead of execution time within 0.18 %, outperforming the two strategies with static t_1 and t_2 (② and ③) in both energy saving and application execution time.

Despite that the worst reconfiguration delay in this case study is 1 ms for restarting the microring tuning circuitry, the average overhead to the transfer time of individual data packets is simulated to be 0.11 ms with the introduction of the `READY` and `STANDBY` states, which translates to 0.46 % of the 24 ms average transfer time of data packets. The overhead to the execution time of a job in percentage terms is even smaller because the job execution time often contains the computation time of a few largest tasks. As a communication-to-computation ratio (ρ) of 0.5 is assumed in this case study, the simulated average overhead to job execution time for POLESTAR further reduces to 0.18 % when task computation time is taken into consideration.

Improvement of Energy Proportionality

Similar to the effective energy efficiency, the energy proportionality can also be computed for either the overall network or each individual link. For the network, its utiliza-

tion rate at a specific time is calculated as the sum of bandwidth capacity requested by the active links over the total bandwidth capacity supported by the network. The power consumption of the network comes from the active links, as well as the idle links that are in READY/STANDBY states or in state transition. Fig. 9.13 plots all of the utilization-power pairs observed during the simulation of the 64-node Fat-Tree network with our POLESTAR strategies. An averaging curve is also drawn as the energy proportionality curve for the overall network, which is significantly closer to an ideal energy-proportional curve compared to the baseline scenario that always keeps the links on.

As for the energy proportionality of an individual link, we calculate its average power consumption for a specific utilization rate as the accumulated energy divided by the total time spent at that utilization rate. For example, the average idle power of a link is computed as the energy consumed during its idle periods, including the energy consumption of the READY and STANDBY states, as well as those consumed during state transition. We compute the idle-to-peak power ratio, as defined in Section 9.2.2, for all links in the simulated network, which demonstrates an average of ~82 % improvement compared to the baseline scenario that always keeps the links on.

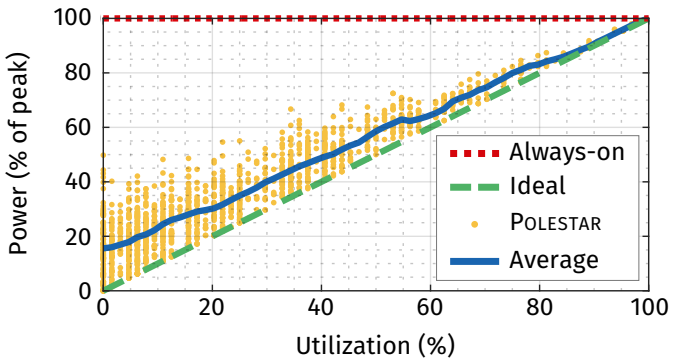


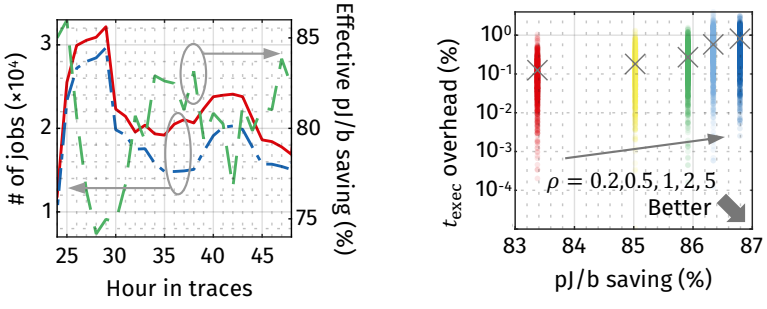
Figure 9.13: Energy proportionality curve for the overall network with POLESTAR, averaged from the simulated utilization-power pairs.

9.5.3 Scalability Analyses

We further evaluate the scalability of the POLESTAR strategies for a broader range of network configurations.

Network Loads

Different hours in the traces As shown in Fig. 9.8b, the workloads in data centers can drastically fluctuate with time. Therefore, we evaluate the POLESTAR strategies for 24 consecutive trace hours with other assumptions unchanged. As shown in Fig. 9.14a, the improvement of network energy efficiency achieved by POLESTAR also fluctuates with the workloads. The reduction in energy saving at higher workloads could be explained by the increased link utilization. As the load balancing mechanism of the computation infrastructure tends to schedule tasks uniformly across the nodes, the traffic patterns resulted from task execution also becomes more spatially uniform under higher workloads. In other words, there are less idle links in the network, which means less opportunities for idle power reconfiguration. Moreover, the increased link utilization also reduces the opportunities for using the intermediate data rate. Nevertheless, POLESTAR can still achieve a ~72 % reduction of the



(a) (b)

Figure 9.14: Evaluation of POLESTAR strategies for (a) different hours in the traces, showing fluctuations of attainable energy saving with workloads; and (b) different values for ρ, showing increased energy saving as well as execution time overhead with ρ.

network pJ/b for the worst case, demonstrating a considerable scalability for network loads.

Communication-to-computation ratio (ρ) Data centers and HPC systems running different applications could have different traffic characteristics. As some applications are computational-intensive while others are communication-bounded, we also vary parameter ρ, the communication-to-computation ratio defined in Section 9.4.2, to study its impact on the POLESTAR strategies. Fig. 9.14b plots the trade-off between the attainable energy saving and the execution time overhead w.r.t. different values for ρ. The horizontal axis is the saving of effective pJ/b of the network. The vertical axis plots the overhead to the execution time of all jobs, where the cross signs indicate the mean values. As can be observed, applying POLESTAR for applications with a larger ρ could lead to greater energy saving at the cost of larger overhead to application execution time. Overall, our POLESTAR strategies scale well across a wide range of ρ by achieving at least 83 % of energy saving with less than 0.8 % overhead to the application execution time.

Corner Cases for the Reconfiguration Delay

To account for potential advances in device design in the near future, we also evaluate POLESTAR for various technology corners mentioned in Table 9.3. As summarized and observed in Table 9.4, POLESTAR can achieve even greater energy saving compared to the SS corner evaluated in the previous sections by using microrings with faster thermal time constants. Further reducing the laser turn-on delay, on the other hand, has limited impact on the effectiveness of POLESTAR strategies, as the laser turn-on delay is already small enough compared to most communication transactions in the traces. This motivates future effort on device design for data-center/HPC interconnects to focus on reducing the stabilization time required for microring tuning.

Table 9.4: Energy improvement of POLESTAR at four technology corners.

Corner	FF	FS	SF	SS
Effective pJ/b **improvement**	87.45 %	84.78 %	87.86 %	85.02 %

Aggregated Data Rate

We further investigate the impact of the aggregated data rate on the effectiveness of POLESTAR strategies. Without changing the workloads, we vary the aggregated data rate of each optical link in the simulated network from 180 Gb/s to 1.44 Tb/s. As the power models for the optical links are based on the assumption of a 24-channel comb laser and a 24-channel microring-based transceiver with a per-channel data rate up to 30 Gb/s (refer to Section 9.4.3), for an aggregated data rate r lower than 720 Gb/s, we assume that each link employs a single of pair transceivers with a per-channel data rate of $r/24$. For an aggregated data rate beyond 720 Gb/s, multiple pairs of transceivers are used for composing a single link. We simulate a one-hour segment of the Alibaba traces with the above settings. Fig. 9.15 summarizes the simulated execution time of the traces and the effective energy efficiency of the network with and without POLESTAR strategies. The following observations are made:

- The prolonged execution time of the one-hour traces at lower aggregated data rates indicate that the overprovisioning of bandwidth capacity for the peak requirement is necessary to some extent; otherwise the task execution time may be bounded by the communication due to congested links. However, further increasing the bandwidth capacity beyond a certain point is unworthy as the task execution time becomes computation-bounded.

- Despite that the microring-based DWDM links can incorporate additional channels to increase the aggregated data rate without changing the nominal energy efficiency

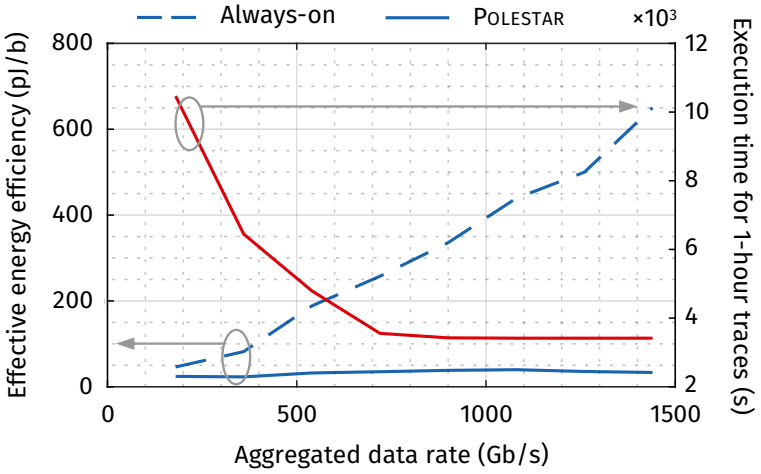


Figure 9.15: Impact of the aggregated data rate on network energy efficiency and task execution time.

(Eq. 9.1), the effective energy efficiency of the network quickly deteriorates with the aggregated data rate, if without power reconfiguration strategies, due to the poorly-managed idle power that is more prominent at higher data rates. Our POLESTAR strategies, on the other hand, can keep the effective energy efficiency of the network relatively constant, thus providing extra room for bandwidth provisioning.

Topology

Finally, we evaluate the scalability of POLESTAR for both the Fat-Tree and the Dragonfly topologies in a simulated network with up to 256 nodes. Further increasing the number of nodes requires excessive memory space during simulation, which is beyond the capability of our server. Table 9.5 summarizes the simulated results for energy saving. It is first observed that the energy saving achieved by POLESTAR for the Dragonfly topology is slightly smaller than that for the Fat-Tree topology. A possible reason is that the Dragonfly topology strongly relies on a grouped structure where intra-group links are assumed to be electrical. Therefore, for the same network size, the Dragonfly topology contains less optical links, thus offering less opportunities for reconfiguration. Another observation is the increase of

Table 9.5: Energy improvement of POLESTAR for different topologies and network sizes.

# of nodes	64	128	256
Fat-Tree	85.02 %	86.18 %	87.24 %
Dragonfly	82.77 %	83.01 %	83.89 %

achievable energy saving with the network size, possibly due to the lower utilization of links in larger networks. Overall, the effectiveness of POLESTAR strategies on various network topologies and sizes remains significant.

9.6 Concluding Remarks

In this chapter, we proposed POLESTAR, i.e., POwer LEvel Scaling with Traffic-Adaptive Reconfiguration, for microring-based optical interconnects. Featuring a collection of runtime reconfiguration strategies that target the power states of the lasers and the microring tuning circuitry, POLESTAR demonstrates remarkable effectiveness for improving the energy efficiency and energy proportionality of underutilized data-center/HPC interconnects. Through traffic-adaptive adjustment of the reconfiguration mechanism, POLESTAR achieves a reasonable balance between energy saving and application execution time. Good scalability across topologies, network loads, and potential advances in optical device design is also observed. POLESTAR is extensible by incorporating more reconfiguration strategies and improving existing ones. With future work targeting better traffic prediction techniques and the possible inclusion of runtime traffic scheduling, POLESTAR paves a promising way to the energy-efficient and energy-proportional optical interconnects for future data-center/HPC applications.

Part V

Discussion

Chapter 10

Conclusion and Future Work

10.1 Dissertation Conclusions

Despite mass deployment in the long-haul telecom regime for decades, optical interconnects for short-reach datacom applications are still being actively explored. Issues that accompany the fast-growing complexity of system integration must be effectively addressed before any broad adoption can take place. Energy efficiency is the main driving force for optical interconnects to replace electrical ones with ever-decreasing distances, thus drawing particular research attention to its optimization. The solutions proposed in this dissertation address the challenges that manifest in three levels, namely the device, link, and system, for the design and optimization of energy-efficient optical interconnects. The target challenges include 1) inadequate support from electronic-photonic design automation (EPDA) methodologies that are indispensable for the development of design optimization techniques, 2) oversimplified characterization of process variations, resulting in variation alleviation techniques with limited effectiveness, and 3) the lack of runtime reconfiguration strategies for the optical interconnects under traffic dynamics, leading to unoptimized energy efficiency. To this end, this dissertation makes the following contributions.

At the device level, compact models are developed for key components of the optical in-

terconnects, including lasers and modulators of multiple types, and extensively validated by measurement data of fabricated devices and circuits. A novel hierarchical model is also proposed for characterizing the spatial variation patterns of microring resonators (MRRs), developed based on the measurement data of a record number of fabricated microrings. The enriched library of device-level models enables accurate circuit-level simulation of optical links and variation-aware estimation of the link power budget, serving as the fundamentals of the optimization techniques proposed at the link and system levels.

At the link level, three techniques are proposed for improving the energy efficiency of the optical interconnects under wafer-scale process variations. The techniques explore, respectively, 1) sub-channel redundancy of laser comb lines, 2) a hybrid mechanism combining electrical and thermal tuning, and 3) optimal grouping of a batch of fabricated transceivers, achieving significant reductions in the nominal energy efficiency (NEE) of optical links.

At the system level, traffic dynamics are addressed by two strategies that reconfigure the power states of the optical interconnects at application runtime. The two strategies incorporate assistance from 1) traffic pattern adjustment achieved by task mapping exploration, and 2) traffic adaptability enabled by idle period prediction, respectively, and demonstrate substantial improvements in the effective energy efficiency (EEE) with minimal overhead to application execution time, notably outperforming existing strategies.

In summary, this dissertation presents a feasible framework for cross-level optimization of optical interconnects, paving the way to the quality design of variation-aware runtime-reconfigurable optical interconnects with optimized energy efficiency.

10.2 Possible Future Directions

Despite what have been achieved, there are still missing pieces in all three levels, namely the device, link, and system, that need to be incorporated for a holistic cross-level optimiza-

tion framework for optical interconnects.

At the device level, it is desirable to develop more models for photonic devices featuring the latest technology advances, preferably to include more photonic-specific information to the model description (e.g., phase, noise, and chirp) and capture more physical-level effects. It is also worth investigating the source of the spatial variation patterns of photonic manufacturing for process optimization, as well as the temporal variation patterns if more fabrication data become available. Moreover, thermal variations should be characterized and modeled for enabling runtime thermal management and optimization of photonic integrated systems.

At the link level, an interesting direction is the co-optimization of the photonic devices with their electronic driving circuitry, such as the CMOS drivers for directly-modulated lasers (DMLs), external modulators, and thermal tuners, to account for both process and thermal variations. It is also desirable to have hardware prototypes of the optimization techniques proposed at the link level for evaluating their implementation overhead. In terms of the algorithms for solving the link-level optimization problems, it is worth investigating customized heuristics with greater efficiency, as many of the problems are formulated as NP-complete.

At the system level, a possible future direction is to investigate the impact of runtime traffic scheduling on the effectiveness of the power reconfiguration strategies. It is also worth looking into other runtime reconfiguration targets besides the power states, such as the link bandwidth and the wavelength allocation scheme. Moreover, as the increasingly-popular distributed machine learning framework deploying more training tasks to edge devices, moving the performance bottleneck of central servers from computation to communication, it becomes an interesting topic to investigate the usage of optical interconnects for eliminating the communication bottleneck of emerging machine learning applications in future heterogeneous computation infrastructures.

Additionally, with the developed methodologies for the modeling and simulation of photonic devices, circuits, and systems, it is worth looking into novel applications of integrated photonics for other functionalities besides communication, such as optical processors and storage, as well as optical accelerators for computational-intensive machine learning tasks.

Bibliography

[1] R. G. Beausoleil, M. McLaren, and N. P. Jouppi, *Photonic architectures for high-performance data centers*, *IEEE Journal of Selected Topics in Quantum Electronics* **19** (Mar., 2013) 3700109.

[2] M. J. Heck, H. W. Chen, A. W. Fang, B. R. Koch, D. Liang, H. Park, M. N. Sysak, and J. E. Bowers, *Hybrid silicon photonics for optical interconnects*, *IEEE Journal on Selected Topics in Quantum Electronics* **17** (2011), no. 2 333–346.

[3] R. Lucas, J. Ang, K. Bergman, S. Borkar, W. Carlson, L. Carrington, G. Chiu, R. Colwell, W. Dally, J. Dongarra, A. Geist, G. L. A. N. L. Grider, R. Haring, J. Hittinger, A. Hoisie, D. M. Klein, P. Kogge, R. Lethin, V. Sarkar, R. Schreiber, J. Shalf, T. Sterling, R. Stevens, J. Bashor, R. Brightwell, P. Coteus, E. Debenedictus, J. Hiller, K. H. Kim, H. Langston, J. S. N. L. Laros III, S. A. N. L. Leyffer, R. M. Murphy, R. A. N. L. Ross, C. Webster, and S. Wild, *DOE advanced scientific computing advisory subcommittee (ASCAC) report: Top ten exascale research challenges*, tech. rep., USDOE Office of Science, SC, United States, Feb., 2014.

[4] S. Rumley, K. Bergman, M. A. Seyedi, and M. Fiorentino, *Evolving requirements and trends of HPC*, in Mukherjee *et. al.* [309], pp. 725–755.

[5] S. Pitris, M. Moralis-Pegios, T. Alexoudi, K. Fotiadis, Y. Ban, P. de Heyn, J. van Campenhout, and N. Pleros, *A 400 Gb/s O-band WDM (8×50 Gb/s) silicon photonic ring modulator-based transceiver*, in *Optical Fiber Communication Conference (OFC)*, p. M4H.3, 2020.

[6] Y. London, T. Van Vaerenbergh, L. Ramini, A. J. Rizzo, P. Sun, G. Kurczveil, A. Seyedi, J. Rhim, M. Fiorentino, and K. Bergman, *Performance requirements for terabit-class silicon photonic links based on cascaded microring resonators*, *Journal of Lightwave Technology* **38** (July, 2020) 3469–3477.

[7] F. Douglis, S. Robertson, E. van den Berg, J. Micallef, M. Pucci, A. Aiken, K. Bergman, M. Hattink, and M. Seok, *FLEET—fast lanes for expedited execution at 10 terabits: Program overview*, *IEEE Internet Computing* **25** (May, 2021) 79–87.

[8] S. Pasricha and M. Nikdast, *A survey of silicon photonics for energy-efficient manycore computing*, *IEEE Design & Test* **37** (Aug., 2020) 60–81.

[9] J. E. Bowers, *Heterogeneous photonic integration on silicon*, in *Cadence Photonics Summit and Workshop*, 2016.

[10] M. Smit, J. van der Tol, and M. Hill, *Moore's law in photonics*, *Laser & Photonics Reviews* **6** (Jan., 2012) 1–13.

[11] B. Jalali and S. Fathpour, *Silicon photonics*, *Journal of Lightwave Technology* **24** (Dec., 2006) 4600–4615.

[12] R. Soref, *The past, present, and future of silicon photonics*, *IEEE Journal of Selected Topics in Quantum Electronics* **12** (Nov., 2006) 1678–1687.

[13] R. G. Beausoleil, *Large-scale integrated photonics for high-performance interconnects*, *ACM Journal on Emerging Technologies in Computing Systems* **7** (June, 2011) 1–54.

[14] V. R. Almeida, Q. Xu, and M. Lipson, *Ultrafast integrated semiconductor optical modulator based on the plasma-dispersion effect*, *Optics Letters* **30** (Sept., 2005) 2403.

[15] A. Y. Liu, S. Srinivasan, J. Norman, A. C. Gossard, and J. E. Bowers, *Quantum dot lasers for silicon photonics [Invited]*, *Photonics Research* **3** (Oct., 2015) B1.

[16] T. Komljenovic, M. Davenport, J. Hulme, A. Y. Liu, C. T. Santis, A. Spott, S. Srinivasan, E. J. Stanton, C. Zhang, and J. E. Bowers, *Heterogeneous silicon photonic integrated circuits*, *Journal of Lightwave Technology* **34** (2016), no. 1 20–35.

[17] M. A. Seyedi, C.-H. Chen, M. Fiorentino, and R. Beausoleil, *Error-free DWDM transmission and crosstalk analysis for a silicon photonics transmitter*, *Optics Express* **23** (Dec., 2015) 32968.

[18] M. J. R. Heck and J. E. Bowers, *Energy efficient and energy proportional optical interconnects for multi-core processors: Driving the need for on-chip sources*, *IEEE Journal of Selected Topics in Quantum Electronics* **20** (July, 2014) 332–343.

[19] X. Zheng, E. Chang, I. Shubin, G. Li, Y. Luo, J. Yao, H. Thacker, J.-H. Lee, J. Lexau, F. Liu, P. Amberg, K. Raj, R. Ho, J. E. Cunningham, and A. V. Krishnamoorthy, *A 33mW 100Gbps CMOS silicon photonic WDM transmitter using off-chip laser sources*, in *Optical Fiber Communication Conference (OFC)/National Fiber Optic Engineers Conference (NFOEC)*, p. PDP5C.9, OSA, 2013.

[20] G. L. Wojcik, D. Yin, A. R. Kovsh, A. E. Gubenko, I. L. Krestnikov, S. S. Mikhrin, D. A. Livshits, D. A. Fattal, M. Fiorentino, and R. G. Beausoleil, *A single comb laser source for short reach WDM interconnects*, in *Proc. SPIE* (A. A. Belyanin and P. M. Smowton, eds.), vol. 7230, p. 72300M, Feb., 2009.

[21] D. Livshits, A. Gubenko, S. Mikhrin, V. Mikhrin, C.-H. Chen, M. Fiorentino, and R. Beausoleil, *High efficiency diode comb-laser for DWDM optical interconnects*, in *2014 Optical Interconnects Conference (OI)*, vol. 6, pp. 83–84, IEEE, May, 2014.

[22] L. A. Coldren, S. W. Corzine, and M. L. Mašanović, *Diode Lasers and Photonic Integrated Circuits*. John Wiley & Sons, Inc., Hoboken, NJ, USA, Mar., 2012.

[23] Y. Tang, J. D. Peters, and J. E. Bowers, *Energy-efficient hybrid silicon electroabsorption modulator for 40-Gb/s 1-V uncooled operation*, *IEEE Photonics Technology Letters* **24** (Oct., 2012) 1689–1692.

[24] W. M. Green, M. J. Rooks, L. Sekaric, and Y. A. Vlasov, *Ultra-compact, low RF power, 10 Gb/s silicon Mach-Zehnder modulator*, *Optics Express* **15** (2007), no. 25 17106.

[25] Q. Xu, B. Schmidt, S. Pradhan, and M. Lipson, *Micrometre-scale silicon electro-optic modulator*, *Nature* **435** (May, 2005) 325–327.

[26] Q. Xu, B. Schmidt, J. Shakya, and M. Lipson, *Cascaded silicon micro-ring modulators for WDM optical interconnection*, *Optics Express* **14** (Feb., 2006) 9431.

[27] S. Manipatruni, L. Chen, and M. Lipson, *Ultra high bandwidth WDM using silicon microring modulators*, *Optics Express* **18** (Aug., 2010) 16858.

[28] Y. Liu, R. Ding, Q. Li, Z. Xuan, Y. Li, Y. Yang, A. E.-j. Lim, P. G.-Q. Lo, K. Bergman, T. Baehr-Jones, and M. Hochberg, *Ultra-compact 320 Gb/s and 160 Gb/s WDM transmitters based on silicon microrings*, in *Optical Fiber Communication Conference*, p. Th4G.6, OSA, 2014.

[29] M. Ashkan Seyedi, J. Hulme, P. Sun, T. Van Vaerenbergh, X. Zheng, G. Kurczveil, Z. Huang, D. Liang, M. Fiorentino, and R. Beausoleil, *Overview of silicon photonics components for commercial DWDM applications*, *OSA Advanced Photonics Congress (IPR, Networks, NOMA, PVLED, SPPCom)* (2019) ITh1A.3.

[30] C. Zhang, S. Zhang, J. D. Peters, and J. E. Bowers, $8 \times 8 \times 40$ *Gbps fully integrated silicon photonic network on chip*, *Optica* **3** (July, 2016) 785.

[31] "2020 IPSR-I integrated photonic systems roadmap." `https://photonicsmanufacturing.org/2020_iprs-i_roadmap_chapters`. Accessed: 2021-03-02.

[32] Y. London, T. Van Vaerenbergh, A. J. Rizzo, P. Sun, J. Hulme, G. Kurczveil, A. Seyedi, B. Wang, X. Zeng, Z. Huang, J. Rhim, M. Fiorentino, and K. Bergman, *Energy efficiency analysis of comb source carrier-injection ring-based silicon photonic link*, *IEEE Journal of Selected Topics in Quantum Electronics* **26** (Mar., 2020) 1–13.

[33] C. A. Thraskias, E. N. Lallas, N. Neumann, L. Schares, B. J. Offrein, R. Henker, D. Plettemeier, F. Ellinger, J. Leuthold, and I. Tomkos, *Survey of photonic and plasmonic interconnect technologies for intra-datacenter and high-performance computing communications*, *IEEE Communications Surveys & Tutorials* **20** (2018), no. 4 2758–2783.

[34] Q. Cheng, M. Bahadori, M. Glick, S. Rumley, and K. Bergman, *Recent advances in optical technologies for data centers: a review*, *Optica* **5** (Nov., 2018) 1354.

[35] J. C. Norman, D. Jung, Z. Zhang, Y. Wan, S. Liu, C. Shang, R. W. Herrick, W. W. Chow, A. C. Gossard, and J. E. Bowers, *A review of high-performance quantum dot lasers on silicon*, *IEEE Journal of Quantum Electronics* **55** (Apr., 2019) 1–11.

[36] A. Hammadi and L. Mhamdi, *A survey on architectures and energy efficiency in data center networks*, *Computer Communications* **40** (Mar., 2014) 1–21.

[37] L. A. Barroso and U. Hölzle, *The case for energy-proportional computing*, *IEEE Computer* **40** (Dec., 2007) 33–37.

[38] L. Versluis, R. Matha, S. Talluri, T. Hegeman, R. Prodan, E. Deelman, and A. Iosup, *The workflow trace archive: Open-access data from public and private computing infrastructures*, *IEEE Transactions on Parallel and Distributed Systems* **31** (Sept., 2020) 2170–2184.

[39] C. Avin, M. Ghobadi, C. Griner, and S. Schmid, *On the complexity of traffic traces and implications*, in *International Conference on Measurement and Modeling of Computer Systems*, vol. 48, pp. 47–48, ACM, June, 2020.

[40] A. Dwivedi and R. Wagner, *Traffic model for usa long-distance optical network*, in *Optical Fiber Communication Conference (OFC). Technical Digest Postconference Edition. Trends in Optics and Photonics Vol.37 (IEEE Cat. No. 00CH37079)*, pp. 156–158, OSA, 2000.

[41] "Accurately Simulate Photonic Components and Circuits - Lumerical." `https://www.lumerical.com/products/`. Accessed: 2021-03-04.

[42] "COMSOL - Software for Multiphysics Simulation." `https://www.comsol.com/`. Accessed: 2021-03-09.

[43] R. Scarmozzino, A. Gopinath, R. Pregla, and S. Helfert, *Numerical techniques for modeling guided-wave photonic devices*, *IEEE Journal of Selected Topics in Quantum Electronics* **6** (Jan., 2000) 150–162.

[44] L. Zhang, S. Yu, M. Nowell, D. Marcenac, J. Carroll, and R. Plumb, *Dynamic analysis of radiation and side-mode suppression in a second-order DFB laser using time-domain large-signal traveling wave model*, *IEEE Journal of Quantum Electronics* **30** (jun, 1994) 1389–1395.

[45] H. Bahrami, H. Sepehrian, C. S. Park, L. A. Rusch, and W. Shi, *Time-domain large-signal modeling of traveling-wave modulators on SOI*, *Journal of Lightwave Technology* **34** (jun, 2016) 2812–2823.

[46] S. Balac and F. Mahé, *An embedded split-step method for solving the nonlinear Schrödinger equation in optics*, *Journal of Computational Physics* **280** (jan, 2015) 295–305.

[47] M. Fiers, T. Van Vaerenbergh, K. Caluwaerts, D. Vande Ginste, B. Schrauwen, J. Dambre, and P. Bienstman, *Time-domain and frequency-domain modeling of nonlinear optical components at the circuit-level using a node-based approach*, *Journal of the Optical Society of America B* **29** (May, 2012) 896.

[48] B. Gallinet, J. Butet, and O. J. F. Martin, *Numerical methods for nanophotonics: standard problems and future challenges*, *Laser & Photonics Reviews* **9** (Nov., 2015) 577–603.

[49] Ning-Ning Feng and Wei-Ping Huang, *Modeling and simulation of photonic devices by generalized space mapping technique*, *Journal of Lightwave Technology* **21** (jun, 2003) 1562–1567.

[50] Y. Ye, D. Spina, Y. Xing, W. Bogaerts, and T. Dhaene, *Numerical modeling of a linear photonic system for accurate and efficient time-domain simulations*, *Photonics Research* **6** (jun, 2018) 560.

[51] S. Matsuo, T. Fujii, K. Hasebe, K. Takeda, T. Sato, and T. Kakitsuka, *Directly modulated buried heterostructure DFB laser on SiO_2/Si substrate fabricated by regrowth of InP using bonded active layer*, *Optics Express* **22** (May, 2014) 12139.

[52] C. Zhang, D. Liang, G. Kurczveil, A. Descos, and R. G. Beausoleil, *Hybrid quantum-dot microring laser on silicon*, *Optica* **6** (Sept., 2019) 1145.

[53] Z. Zhang, D. Jung, J. C. Norman, W. W. Chow, and J. E. Bowers, *Linewidth enhancement factor in InAs/GaAs quantum dot lasers and its implication in isolator-free and narrow linewidth applications*, *IEEE Journal of Selected Topics in Quantum Electronics* **25** (Nov., 2019) 1–9.

[54] Z. Zhang, *Monolithic passive-active integration of epitaxially grown quantum dot lasers*. PhD thesis, University of California, Santa Barbara, 2021.

[55] M. Bahadori, *Physical Layer Modeling and Optimization of Silicon Photonic Interconnection Networks*. PhD thesis, Columbia University, 2018.

[56] M. Bahadori, K. Bergman, M. Nikdast, S. Rumley, L. Y. Dai, N. Janosik, T. Van Vaerenbergh, A. Gazman, Q. Cheng, and R. Polster, *Design space exploration of microring resonators in silicon photonic interconnects: Impact of the ring curvature*, *Journal of Lightwave Technology* **36** (July, 2018) 2767–2782.

[57] X. Chen, Z. Wang, Y.-S. Chang, J. Xu, J. Feng, P. Yang, Z. Wang, and L. H. K. Duong, *Modeling and analysis of optical modulators based on free-carrier plasma dispersion effect*, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **39** (May, 2020) 977–990.

[58] R. Palmer, W. Freude, J. Leuthold, C. Koos, S. Koeber, D. L. Elder, M. Woessner, W. Heni, D. Korn, M. Lauermann, W. Bogaerts, and L. Dalton, *High-speed, low drive-voltage silicon-organic hybrid modulator based on a binary-chromophore electro-optic material*, *Journal of Lightwave Technology* **32** (Aug., 2014) 2726–2734.

[59] J. Sun, R. Kumar, M. Sakib, J. B. Driscoll, H. Jayatilleka, and H. Rong, *A 128 Gb/s PAM4 silicon microring modulator with integrated thermo-optic resonance tuning*, *Journal of Lightwave Technology* **37** (Jan., 2019) 110–115.

[60] D. Benedikovic, L. Virot, G. Aubin, F. Amar, B. Szelag, B. Karakus, J.-M. Hartmann, C. Alonso-Ramos, X. L. Roux, P. Crozat, E. Cassan, D. Marris-Morini, C. Baudot, F. Boeuf, J.-M. Fédéli, C. Kopp, and L. Vivien, *25 Gbps low-voltage hetero-structured silicon-germanium waveguide pin photodetectors for monolithic on-chip nanophotonic architectures*, *Photonics Research* **7** (Apr., 2019) 437.

[61] J. Jensen and O. Sigmund, *Topology optimization for nano-photonics*, *Laser & Photonics Reviews* **5** (Mar., 2011) 308–321.

[62] S. Molesky, Z. Lin, A. Y. Piggott, W. Jin, J. Vucković, and A. W. Rodriguez, *Inverse design in nanophotonics*, *Nature Photonics* **12** (Nov., 2018) 659–670.

[63] Y. Xu, J. Huang, L. Yang, H. Ma, H. Yuan, T. Xie, J. Yang, and Z. Zhang, *Inverse-designed ultra-compact high efficiency and low crosstalk optical interconnect based on waveguide crossing and wavelength demultiplexer*, *Scientific Reports* **11** (Dec., 2021) 12842.

[64] S. D. Campbell, D. Sell, R. P. Jenkins, E. B. Whiting, J. A. Fan, and D. H. Werner, *Review of numerical optimization techniques for meta-device design [invited]*, *Optical Materials Express* **9** (Apr., 2019) 1842.

[65] W. J. Kim and J. D. O'Brien, *Optimization of a two-dimensional photonic-crystal waveguide branch by simulated annealing and the finite-element method*, *Journal of the Optical Society of America B* **21** (Feb., 2004) 289.

[66] Y. Zhao, X. Cao, J. Gao, Y. Sun, H. Yang, X. Liu, Y. Zhou, T. Han, and W. Chen, *Broadband diffusion metasurface based on a single anisotropic element and optimized by the simulated annealing algorithm*, *Scientific Reports* **6** (July, 2016) 23896.

[67] S. Preble, M. Lipson, and H. Lipson, *Two-dimensional photonic crystals designed by evolutionary algorithms*, *Applied Physics Letters* **86** (Feb., 2005) 061111.

[68] Y. Shi, W. Li, A. Raman, and S. Fan, *Optimization of multilayer optical films with a memetic algorithm and mixed integer programming*, *ACS Photonics* **5** (Mar., 2018) 684–691.

[69] A. Y. Piggott, J. Lu, K. G. Lagoudakis, J. Petykiewicz, T. M. Babinec, and J. Vučković, *Inverse design and demonstration of a compact and broadband on-chip wavelength demultiplexer*, *Nature Photonics* **9** (June, 2015) 374–377.

[70] C. M. Lalau-Keraly, S. Bhargava, O. D. Miller, and E. Yablonovitch, *Adjoint shape optimization applied to electromagnetic design*, *Optics Express* **21** (Sept., 2013) 21693.

[71] A. M. Hammond and R. M. Camacho, *Designing integrated photonic devices using artificial neural networks*, *Optics Express* **27** (Oct., 2019) 29620.

[72] J. Jiang, M. Chen, and J. A. Fan, *Deep neural networks for the evaluation and design of photonic devices*, *Nature Reviews Materials* (Dec., 2020).

[73] Z. A. Kudyshev, A. V. Kildishev, V. M. Shalaev, and A. Boltasseva, *Machine learning-assisted global optimization of photonic devices*, *Nanophotonics* **10** (2020), no. 1 371–383.

[74] W. Ma, Z. Liu, Z. A. Kudyshev, A. Boltasseva, W. Cai, and Y. Liu, *Deep learning for the design of photonic structures*, *Nature Photonics* **15** (Feb., 2021) 77–90.

[75] M. S. Wartak, *Computational Photonics: An Introduction with MATLAB*. Cambridge University Press, 2013.

[76] "openEPDA Overview — openEPDA™ documentation." https://openepda.org/. Accessed: 2021-06-22.

[77] "PIC Design and Simulation Software - Lumerical INTERCONNECT." https://www.lumerical.com/products/interconnect/. Accessed: 2021-06-22.

[78] "OptSim Circuit | Synopsys." https://www.synopsys.com/photonic-solutions/pic-design-suite/optsim-circuit.html. Accessed: 2021-06-22.

[79] "IPKISS | Luceda Photonics." https://www.lucedaphotonics.com/en/product/ipkiss. Accessed: 2021-06-22.

[80] "VPIcomponentMaker™ Photonic Circuits – Overview." https://www.vpiphotonics.com/Tools/PhotonicCircuits/. Accessed: 2021-06-22.

[81] "Introduction to Verilog-A — Documentation." https://verilogams.com/tutorials/vloga-intro.html. Accessed: 2021-06-22.

[82] L. W. Nagel and D. Pederson, *SPICE (simulation program with integrated circuit emphasis)*, Tech. Rep. UCB/ERL M382, EECS Department, University of California, Berkeley, Apr., 1973.

[83] "Spectre Simulation Platform." https://www.cadence.com/en_US/home/tools/custom-ic-analog-rf-design/circuit-simulation/spectre-simulation-platform.html. Accessed: 2021-06-22.

[84] T. Christen, S. Odermatt, M. Jungo, D. Erni, and W. Baechtold, *Verilog-A implementation of a 2D spatiotemporal VCSEL model for system-oriented simulations of optical links*, *Microwave and Optical Technology Letters* **38** (Aug., 2003) 304–308.

[85] B. Wang, W. V. Sorin, S. Palermo, and M. R. T. Tan, *Comprehensive vertical-cavity surface-emitting laser model for optical interconnect transceiver circuit design*, *Optical Engineering* **55** (dec, 2016) 126103.

[86] K. Zhu, V. Saxena, and W. Kuang, *Compact Verilog-A modeling of silicon traveling-wave modulator for hybrid CMOS photonic circuit design*, in *2014 IEEE 57th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pp. 615–618, IEEE, Aug., 2014.

[87] J. Rhim, Y. Ban, B.-M. Yu, J.-M. Lee, and W.-Y. Choi, *Verilog-A behavioral model for resonance-modulated silicon micro-ring modulator*, *Optics Express* **23** (Apr., 2015) 8762.

[88] R. Wu, C.-H. Chen, J.-M. Fedeli, M. Fournier, K.-T. Cheng, and R. G. Beausoleil, *Compact models for carrier-injection silicon microring modulators*, *Optics Express* **23** (June, 2015) 15545.

[89] B. Wang, C. Li, C.-H. Chen, K. Yu, M. Fiorentino, R. G. Beausoleil, and S. Palermo, *A compact Verilog-A model of silicon carrier-injection ring modulators for optical interconnect transceiver circuit design*, *Journal of Lightwave Technology* **34** (June, 2016) 2996–3005.

[90] E. Kononov, *Modeling photonic links in Verilog-A*, Master's thesis, Massachusetts Institute of Technology, 2013.

[91] C. Sorace-Agaskar, J. Leu, M. R. Watts, and V. Stojanovic, *Electro-optical co-simulation for integrated CMOS photonic circuits with VerilogA*, *Optics Express* **23** (Oct., 2015) 27180.

[92] R. Wu, *Variation-Aware Modeling and Design of Nanophotonic Interconnects*. PhD thesis, University of California, Santa Barbara, 2017.

[93] M. J. Shawon and V. Saxena, *Rapid simulation of photonic integrated circuits using Verilog-A compact models*, *IEEE Transactions on Circuits and Systems I: Regular Papers* **67** (Oct., 2020) 3331–3341.

[94] M. De Wilde, O. Rits, R. Bockstaele, J. M. Van Campenhout, and R. G. Baets, *Circuit-level simulation approach to analyze system-level behavior of VCSEL-based optical interconnects*, in *VCSELs and Optical Interconnects* (H. Thienpont and J. Danckaert, eds.), vol. 4942, p. 247, Apr., 2003.

[95] P. Martin, F. Gays, E. Grellier, A. Myko, and S. Menezo, *Modeling of silicon photonics devices with Verilog-A*, in *2014 29th International Conference on Microelectronics Proceedings (MIEL)*, pp. 209–212, 2014.

[96] T. Pinguet, S. Gloeckner, G. Masini, and A. Mekis, *CMOS photonics: A platform for optoelectronics integration*, in *Topics in Applied Physics* (D. J. Lockwood and L. Pavesi, eds.), vol. 119 of *Topics in Applied Physics*, pp. 187–216. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[97] K. Qian, B. Nikolić, and C. J. Spanos, *Hierarchical modeling of spatial variability with a 45nm example*, in *Design for Manufacturability through Design-Process Integration III* (V. K. Singh and M. L. Rieger, eds.), vol. 7275, p. 727505, Mar., 2009.

[98] R. Wu, C.-H. Chen, T.-C. Huang, R. Beausoleil, and K.-T. Cheng, *Spatial pattern analysis of process variations in silicon microring modulators*, in *2016 IEEE Optical Interconnects Conference (OI)*, vol. 2, pp. 116–117, IEEE, May, 2016.

[99] N. Boynton, A. Pomerene, A. Starbuck, A. Lentine, and C. T. DeRose, *Characterization of systematic process variation in a silicon photonic platform*, in *2017 IEEE Optical Interconnects Conference (OI)*, pp. 11–12, IEEE, June, 2017.

[100] P. Sun, R. G. Beausoleil, J. Hulme, T. Van Vaerenbergh, J. Rhim, C. Baudot, F. Boeuf, N. Vulliet, A. Seyedi, and M. Fiorentino, *Statistical behavioral models of silicon ring resonators at a commercial CMOS foundry*, *IEEE Journal of Selected Topics in Quantum Electronics* **26** (Mar., 2020) 1–10.

[101] X. Chen, M. Mohamed, Z. Li, L. Shang, and A. R. Mickelson, *Process variation in silicon photonic devices*, *Applied Optics* **52** (Nov., 2013) 7638.

[102] L. Chrostowski, X. Wang, J. Flueckiger, Y. Wu, Y. Wang, and S. Talebi Fard, *Impact of fabrication non-uniformity on chip-scale silicon photonic integrated circuits*, *Optical Fiber Communication Conference (OFC)* (2014) 2–4.

[103] Z. Lu, J. Jhoja, J. Klein, X. Wang, A. Liu, J. Flueckiger, J. Pond, and L. Chrostowski, *Performance prediction for silicon photonics integrated circuits with layout-dependent correlated manufacturing variability*, *Optics Express* **25** (May, 2017) 9712.

[104] S. K. Selvaraja, W. Bogaerts, P. Dumon, D. Van Thourhout, and R. Baets, *Subnanometer linewidth uniformity in silicon nanophotonic waveguide devices using CMOS fabrication technology*, *IEEE Journal of Selected Topics in Quantum Electronics* **16** (Jan., 2010) 316–324.

[105] R. Wu, M. A. Seyedi, Y. Wang, J. Hulme, M. Fiorentino, R. G. Beausoleil, and K.-T. Cheng, *Pairing of microring-based silicon photonic transceivers for tuning power optimization*, in *2018 23rd Asia and South Pacific Design Automation Conference (ASP-DAC)*, pp. 135–140, IEEE, Jan., 2018.

[106] Y. Xing, J. Dong, U. Khan, and W. Bogaerts, *Hierarchical model for spatial variations of integrated photonics*, in *2018 IEEE 15th International Conference on Group IV Photonics (GFP)*, pp. 1–2, IEEE, Aug., 2018.

[107] W. A. Zortman, D. C. Trotter, and M. R. Watts, *Silicon photonics manufacturing*, *Optics Express* **18** (Nov., 2010) 23598.

[108] Crid Yu, Tinaung Maung, C. Spanos, D. Boning, J. Chung, Hua-Yu Liu, Keh-Jeng Chang, and D. Bartelink, *Use of short-loop electrical measurements for yield improvement*, *IEEE Transactions on Semiconductor Manufacturing* **8** (May, 1995) 150–159.

[109] B. Stine, D. S. Boning, J. E. Chung, D. A. Bell, and E. Equi, *Inter- and intra-die polysilicon critical dimension variation*, in *Proceedings of SPIE - The International Society for Optical Engineering* (A. Keshavarzi, S. Prasad, and H.-D. Hartmann, eds.), vol. 2874, pp. 27–35, Sept., 1996.

[110] B. Stine, D. Boning, and J. Chung, *Analysis and decomposition of spatial variation in integrated circuit processes and devices*, *IEEE Transactions on Semiconductor Manufacturing* **10** (1997), no. 1 24–41.

[111] S. J. Bae, J. Y. Hwang, and W. Kuo, *Yield prediction via spatial modeling of clustered defect counts across a wafer map*, *IIE Transactions* **39** (Oct., 2007) 1073–1083.

[112] W. Zhang, K. Balakrishnan, X. Li, D. Boning, and R. Rutenbar, *Toward efficient spatial variation decomposition via sparse regression*, in *2011 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pp. 162–169, IEEE, Nov., 2011.

[113] Wangyang Zhang, K. Balakrishnan, Xin Li, D. S. Boning, S. Saxena, A. Strojwas, and R. A. Rutenbar, *Efficient spatial pattern analysis for variation decomposition via robust sparse regression*, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **32** (July, 2013) 1072–1085.

[114] K. Huang, N. Kupp, J. M. Carulli, and Y. Makris, *Process monitoring through wafer-level spatial variation decomposition*, in *2013 IEEE International Test Conference (ITC)*, pp. 1–10, IEEE, Sept., 2013.

[115] W. Zhang, X. Li, and R. A. Rutenbar, *Bayesian virtual probe: Minimizing variation characterization cost for nanoscale IC technologies via Bayesian inference*, in *Proceedings of the 47th Design Automation Conference (DAC)*, p. 262, ACM Press, 2010.

[116] W. Zhang, X. Li, F. Liu, E. Acar, R. A. Rutenbar, and R. D. Blanton, *Virtual probe: A statistical framework for low-cost silicon characterization of nanoscale integrated circuits*, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **30** (Dec., 2011) 1814–1827.

[117] N. Kupp, K. Huang, J. Carulli, and Y. Makris, *Spatial estimation of wafer measurement parameters using Gaussian process models*, in *2012 IEEE International Test Conference (ITC)*, pp. 1–8, IEEE, Nov., 2012.

[118] C.-K. Hsu, F. Lin, K.-T. Cheng, W. Zhang, X. Li, J. M. Carulli, and K. M. Butler, *Test data analytics — exploring spatial and test-item correlations in production test data*, in *2013 IEEE International Test Conference (ITC)*, pp. 1–10, IEEE, Sept., 2013.

[119] S. Zhang, F. Lin, C.-K. Hsu, K.-T. Cheng, and H. Wang, *Joint virtual probe: Joint exploration of multiple test items' spatial patterns for efficient silicon characterization and test prediction*, in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, (New Jersey), pp. 1–6, IEEE Conference Publications, 2014.

[120] J. Luan and Z. Zhang, *Prediction of multidimensional spatial variation data via bayesian tensor completion*, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **39** (Feb., 2020) 547–551.

[121] Cadence Design Systems, Inc., "Integrated Design Flows for Photonics Circuits." https://www.cadence.com/content/dam/cadence-www/global/en_US/documents/solutions/photonics-brochure.pdf, 2020. Accessed: 2021-08-01.

[122] "Process Design Kit (PDK) — AIM Photonics." https://www.aimphotonics.com/process-design-kit. Accessed: 2021-08-01.

[123] "Silicon photonic ICs for prototyping - imec | imec." https://www.imec-int.com/en/silicon-photonic-ICs-prototyping. Accessed: 2021-08-01.

[124] "Silicon Photonics (SiPho) - Tower Semiconductor." https://towersemi.com/technology/rf-and-hpa/silicon-photonics-rf/. Accessed: 2021-08-01.

[125] "Technology - Advanced Micro Foundry." http://www.advmf.com/technology/. Accessed: 2021-08-01.

[126] "Silicon Photonics Solutions - CompoundTek Pte Ltd." https://compoundtek.com/our-solutions/. Accessed: 2021-08-01.

[127] "AN Technology - LIGENTEC." https://www.ligentec.com/ligentec-an-technology/. Accessed: 2021-08-01.

[128] "LioniX International – Photonic IC modules based on Silicon-Nitride." https://photonics.lionix-international.com/. Accessed: 2021-08-01.

[129] "Photonic InP Foundry – Fraunhofer Heinrich Hertz Institute." https://www.hhi.fraunhofer.de/en/departments/pc/research-groups/photonic-inp-foundry.html. Accessed: 2021-08-01.

[130] "Home - SMART Photonics." https://smartphotonics.nl/. Accessed: 2021-08-01.

[131] "Ecosystem Partners - Lumerical." https://www.lumerical.com/partners/{#}foundry. Accessed: 2021-08-01.

[132] "Silicon Photonics Process Design Kits | Synopsys." https://www.synopsys.com/photonic-solutions/pic-design-suite/process-design-kits.html. Accessed: 2021-08-01.

[133] "IPKISS PDKs | Luceda Photonics Foundry Support." https://www.lucedaphotonics.com/en/ipkiss-pdks. Accessed: 2021-08-01.

[134] "VPItoolkit™ PDK." https://www.vpiphotonics.com/Tools/PDK/. Accessed: 2021-08-01.

[135] "Photonics." https://www.cadence.com/en_US/home/solutions/photonics.html. Accessed: 2021-06-22.

[136] "Synopsys Photonic Solutions." https://www.synopsys.com/photonic-solutions.html. Accessed: 2021-06-22.

[137] "Photonic Design | Siemens Digital Industries Software." https://eda.sw.siemens.com/en-US/ic/ic-custom/photonic/. Accessed: 2021-06-22.

[138] "High-Performance Photonic Simulation Software - Lumerical." https://www.lumerical.com/. Accessed: 2021-06-22.

[139] "Luceda | Software and services for integrated photonics designers." https://www.lucedaphotonics.com/en. Accessed: 2021-06-22.

[140] "VPIphotonics: Simulation Software and Design Services." https://www.vpiphotonics.com/index.php. Accessed: 2021-06-22.

[141] J. Pond, G. S. Lamant, and R. Goldman, *Software tools for integrated photonics*, in *Optical Fiber Telecommunications VII*, pp. 195–231. Elsevier, 2020.

[142] A. Farsaei, J. Klein, J. Pond, J. Flueckiger, X. Wang, G. Lamant, L. Chrostowski, and S. Mirabbasi, *A novel and scalable design methodology for the simulation of photonic integrated circuits*, in *Advanced Photonics 2016 (IPR, NOMA, Sensors, Networks, SPPCom, SOF)*, vol. 2016, (Washington, D.C.), p. JTu4A.2, OSA, 2016.

[143] L. Alloatti, M. Wade, V. Stojanovic, M. Popovic, and R. J. Ram, *Photonics design tool for advanced CMOS nodes*, *IET Optoelectronics* **9** (Aug., 2015) 163–167.

[144] J. S. Orcutt and R. J. Ram, *Photonic device layout within the foundry CMOS design environment*, *IEEE Photonics Technology Letters* **22** (Apr., 2010) 544–546.

[145] W. Bogaerts and L. Chrostowski, *Silicon photonics circuit design: Methods, tools and challenges*, *Laser & Photonics Reviews* **12** (Apr., 2018) 1700237.

[146] "SKILL Language Programming." https://www.cadence.com/en_US/home/training/all-courses/83018.html. Accessed: 2021-08-02.

[147] "IC Design Flow With Pyxis - Siemens Digital Industries Software." https://eda.learn.sw.siemens.com/training/courses/ic-design-flow-with-pyxis. Accessed: 2021-08-02.

[148] "OptoDesigner | Synopsys." https://www.synopsys.com/photonic-solutions/pic-design-suite/photonic-chip-mask-layout/advanced-connectors-module.html. Accessed: 2021-08-02.

[149] "KLayout Layout Viewer And Editor." https://klayout.de/. Accessed: 2021-08-02.

[150] A. N. McCaughan, A. M. Tait, S. M. Buckley, D. M. Oh, J. T. Chiles, J. M. Shainline, and S. W. Nam, *PHIDL: Python CAD layout and geometry creation for nanolithography*, arXiv:2103.01152.

[151] G. Hendry, J. Chan, L. P. Carloni, and K. Bergman, *VANDAL: A tool for the design specification of nanophotonic networks*, in *2011 Design, Automation & Test in Europe*, pp. 1–6, IEEE, Mar., 2011.

[152] S. Mingaleev, A. Richter, E. Sokolov, C. Arellano, and I. Koltchanov, *Towards an automated design framework for large-scale photonic integrated circuits*, in *Integrated Optics: Physics and Simulations II* (P. Cheben, J. Čtyroký, and I. Molina-Fernández, eds.), p. 951602, May, 2015.

[153] L. Chrostowski, Z. Lu, J. Flueckiger, X. Wang, J. Klein, A. Liu, J. Jhoja, and J. Pond, *Design and simulation of silicon photonic schematics and layouts*, in *Silicon Photonics and Photonic Integrated Circuits V* (L. Vivien, L. Pavesi, and S. Pelli, eds.), p. 989114, May, 2016.

[154] L. Chrostowski, J. Flueckiger, C. Lin, M. Hochberg, J. Pond, J. Klein, J. Ferguson, and C. Cone, *Design methodologies for silicon photonic integrated circuits*, in *Smart Photonic and Optoelectronic Integrated Circuits XVI* (L. A. Eldada, E.-H. Lee, and S. He, eds.), p. 89890G, Mar., 2014.

[155] R. Cao, J. Ferguson, F. Gays, Y. Drissi, A. Arriordaz, and I. O'Connor, *Silicon photonics design rule checking: Application of a programmable modeling engine for non-manhattan geometry verification*, in *2014 22nd International Conference on Very Large Scale Integration (VLSI-SoC)*, pp. 1–6, IEEE, Oct., 2014.

[156] R. Cao, J. Ferguson, A. Bakker, T. Korthorst, A. Arriordaz, and I. O'Connor, *Photonic designs with EDA approach: A novel methodology for curve design validation*, *2015 IEEE Summer Topicals Meeting Series, SUM 2015* **3** (2015) 29–30.

[157] R. Cao, J. Billoudet, J. Ferguson, L. Couder, J. Cayo, A. Arriordaz, and I. O'Connor, *Lvs check for photonic integrated circuits – curvilinear feature extraction and validation*, in *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2015*, pp. 1253–1256, 2015.

[158] M. Ismail, R. S. El Shamy, K. Madkour, S. Hammouda, and M. A. Swillam, *Toward automated parasitic extraction of silicon photonics using layout physical verifications*, *Journal of Optics* **18** (Aug., 2016) 085801.

[159] J. Pond, J. Klein, J. Flückiger, X. Wang, Z. Lu, J. Jhoja, and L. Chrostowski, *Predicting the yield of photonic integrated circuits using statistical compact modeling*, in *Proceedings of SPIE 10242, Integrated Optics: Physics and Simulations III* (P. Cheben, J. Čtyroký, and I. Molina-Fernández, eds.), p. 102420S, May, 2017.

[160] Y. Xing, D. Spina, A. Li, T. Dhaene, and W. Bogaerts, *Stochastic collocation for device-level variability analysis in integrated photonics*, *Photonics Research* **4** (Apr., 2016) 93.

[161] M. Eldred, *Recent advances in non-intrusive polynomial chaos and stochastic collocation methods for uncertainty analysis and design*, in *50th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, American Institute of Aeronautics and Astronautics, May, 2009.

[162] D. Spina, F. Ferranti, G. Antonini, T. Dhaene, and L. Knockaert, *Efficient variability analysis of electromagnetic systems via polynomial chaos and model order reduction*, *IEEE Transactions on Components, Packaging and Manufacturing Technology* **4** (June, 2014) 1038–1051.

[163] D. Cassano, F. Morichetti, and A. Melloni, *Statistical Analysis of Photonic Integrated Circuits Via Polynomial-Chaos Expansion*, in *Advanced Photonics 2013*, (Washington, D.C.), p. JT3A.8, OSA, 2013.

[164] T.-W. Weng, Z. Zhang, Z. Su, Y. Marzouk, A. Melloni, and L. Daniel, *Uncertainty quantification of silicon photonic devices with correlated and non-Gaussian random parameters*, *Optics Express* **23** (Feb., 2015) 4242.

[165] A. Waqas, D. Melati, and A. Melloni, *Sensitivity analysis and uncertainty mitigation of photonic integrated circuits*, *Journal of Lightwave Technology* **35** (Sept., 2017) 3713–3721.

[166] T.-W. Weng, D. Melati, A. Melloni, and L. Daniel, *Stochastic simulation and robust design optimization of integrated photonic filters*, *Nanophotonics* **6** (Jan., 2017) 299–308.

[167] Z. Zhang, K. Batselier, H. Liu, L. Daniel, and N. Wong, *Tensor computation: A new framework for high-dimensional problems in EDA*, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **36** (Apr., 2017) 521–536.

[168] Z. Zhang, T.-W. Weng, and L. Daniel, *Big-data tensor recovery for high-dimensional uncertainty quantification of process variations*, *IEEE Transactions on Components, Packaging and Manufacturing Technology* **7** (May, 2017) 687–697.

[169] C. Cui and Z. Zhang, *High-dimensional uncertainty quantification of electronic and photonic IC with non-Gaussian correlated process variations*, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **39** (Aug., 2020) 1649–1661.

[170] C. Cui, K. Liu, and Z. Zhang, *Chance-constrained and yield-aware optimization of photonic ICs with non-Gaussian correlated process variations*, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **39** (Dec., 2020) 4958–4970.

[171] Z. He and Z. Zhang, *High-dimensional uncertainty quantification via tensor regression with rank determination and adaptive sampling*, arXiv:2103.17236.

[172] D. Liang, G. Kurczveil, M. Fiorentino, S. Srinivasan, J. E. Bowers, and R. G. Beausoleil, *A tunable hybrid III-V-on-Si MOS microring resonator with negligible tuning power consumption*, in *Optical Fiber Communication Conference (OFC)*, (Washington, D.C.), p. Th1K.4, OSA, 2016.

[173] A. V. Krishnamoorthy, Xuezhe Zheng, Guoliang Li, Jin Yao, T. Pinguet, A. Mekis, H. Thacker, I. Shubin, Ying Luo, K. Raj, and J. E. Cunningham, *Exploiting CMOS manufacturing to reduce tuning requirements for resonant optical devices*, *IEEE Photonics Journal* **3** (June, 2011) 567–579.

[174] M. Georgas, J. Leu, B. Moss, C. Sun, and V. Stojanovic, *Addressing link-level design tradeoffs for integrated photonic interconnects*, in *2011 IEEE Custom Integrated Circuits Conference (CICC)*, pp. 1–8, IEEE, Sept., 2011.

[175] Yan Zheng, P. Lisherness, S. Shamshiri, A. Ghofrani, Shiyuan Yang, and K.-T. T. Cheng, *Post-fabrication reconfiguration for power-optimized tuning of optically connected multi-core systems*, in *17th Asia and South Pacific Design Automation Conference (ASP-DAC)*, pp. 615–620, IEEE, Jan., 2012.

[176] Y. Zheng, P. Lisherness, M. Gao, J. Bovington, K.-T. Cheng, H. Wang, and S. Yang, *Power-efficient calibration and reconfiguration for optical network-on-chip*, *Journal of Optical Communications and Networking* **4** (Dec., 2012) 955.

[177] C. Sun, C.-H. O. Chen, G. Kurian, L. Wei, J. Miller, A. Agarwal, L.-S. Peh, and V. Stojanovic, *DSENT - a tool connecting emerging photonics with electronics for opto-electronic networks-on-chip modeling*, in *2012 IEEE/ACM Sixth International Symposium on Networks-on-Chip*, pp. 201–210, IEEE, May, 2012.

[178] Z. Wang, J. Xu, P. Yang, X. Wang, Z. Wang, L. H. K. Duong, Z. Wang, R. K. V. Maeda, and H. Li, *Improve chip pin performance using optical interconnects*, *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* **24** (Apr., 2016) 1574–1587.

[179] L. H. K. Duong, Z. Wang, M. Nikdast, J. Xu, P. Yang, Z. Wang, Z. Wang, R. K. V. Maeda, H. Li, X. Wang, S. Le Beux, and Y. Thonnart, *Coherent and incoherent crosstalk noise analyses in interchip/intrachip optical interconnection networks*, *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* **24** (July, 2016) 2475–2487.

[180] Y. Ye, J. Xu, X. Wu, W. Zhang, X. Wang, M. Nikdast, Z. Wang, and W. Liu, *System-level modeling and analysis of thermal effects in optical networks-on-chip*, *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* **21** (Feb., 2013) 292–305.

[181] M. S. Kim, Y. W. Kim, and T. H. Han, *System-level signal analysis methodology for optical network-on-chip using linear model-based characterization*, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **39** (Oct., 2020) 2761–2771.

[182] S. Rumley, M. Bahadori, K. Wen, D. Nikolova, and K. Bergman, *PhoenixSim: Crosslayer design and modeling of silicon photonic interconnects*, in *Proceedings of the 1st International Workshop on Advanced Interconnect Solutions and Technologies for Emerging Computing Systems*, pp. 1–6, ACM, Jan., 2016.

[183] A. Varga and R. Hornig, *An overview of the OMNeT++ simulation environment*, in *Proceedings of the First International ICST Conference on Simulation Tools and Techniques for Communications Networks and Systems*, ICST, 2008.

[184] X. Ma, J. Yu, X. Hua, C. Wei, Y. Huang, L. Yang, D. Li, Q. Hao, P. Liu, X. Jiang, and J. Yang, *LioeSim: A network simulator for hybrid opto-electronic networks-on-chip analysis*, *Journal of Lightwave Technology* **32** (Nov., 2014) 4301–4310.

[185] A. Kahng, Bin Li, Li-Shiuan Peh, and K. Samadi, *ORION 2.0: A fast and accurate NoC power and area model for early-stage design space exploration*, in *2009 Design, Automation & Test in Europe Conference & Exhibition*, pp. 423–428, IEEE, Apr., 2009.

[186] R. K. V. Maeda, P. Yang, X. Wu, Z. Wang, J. Xu, Z. Wang, H. Li, L. H. K. Duong, and Z. Wang, *JADE: a heterogeneous multiprocessor system simulation platform using recorded and statistical application models*, in *Proceedings of the 1st International Workshop on Advanced Interconnect Solutions and Technologies for Emerging Computing Systems - AISTECS '16*, pp. 1–6, ACM Press, 2016.

[187] M. M. K. Martin, D. J. Sorin, B. M. Beckmann, M. R. Marty, M. Xu, A. R. Alameldeen, K. E. Moore, M. D. Hill, and D. A. Wood, *Multifacet's general execution-driven multiprocessor simulator (GEMS) toolset*, *ACM SIGARCH Computer Architecture News* **33** (Nov., 2005) 92–99.

[188] M. Zhang, L. He, and D. Fan, *Self-correction trace model: A full-system simulator for optical network-on-chip*, in *2012 IEEE 26th International Parallel and Distributed Processing Symposium Workshops & PhD Forum*, pp. 242–247, IEEE, May, 2012.

[189] D. Wang, M. Sivakumar, A. Kumar, J. McNair, and D. Richards, *Multi-layer simulation design and validation for a two-tier fault-tolerant WDM LAN*, *Journal of Optical Communications and Networking* **4** (Feb., 2012) 142.

[190] C. Tan, Y. Ou, S. Jiang, P. Pan, C. Torng, S. Agwa, and C. Batten, *PyOCN: A unified framework for modeling, testing, and evaluating on-chip networks*, in *2019 IEEE 37th International Conference on Computer Design (ICCD)*, pp. 437–445, IEEE, Nov., 2019.

[191] N. Binkert, B. Beckmann, G. Black, S. K. Reinhardt, A. Saidi, A. Basu, J. Hestness, D. R. Hower, T. Krishna, S. Sardashti, R. Sen, K. Sewell, M. Shoaib, N. Vaish, M. D. Hill, and D. A. Wood, *The gem5 simulator*, *ACM SIGARCH Computer Architecture News* **39** (May, 2011) 1–7.

[192] J. E. Miller, H. Kasture, G. Kurian, C. Gruenwald, N. Beckmann, C. Celio, J. Eastep, and A. Agarwal, *Graphite: A distributed parallel simulator for multicores*, in *HPCA - 16 2010 The Sixteenth International Symposium on High-Performance Computer Architecture*, pp. 1–12, IEEE, Jan., 2010.

[193] R. Ubal, B. Jang, P. Mistry, D. Schaa, and D. Kaeli, *Multi2Sim: A simulation framework for CPU-GPU computing*, in *2012 21st International Conference on Parallel Architectures and Compilation Techniques (PACT)*, pp. 335–344, 2012.

[194] H. Casanova, A. Giersch, A. Legrand, M. Quinson, and F. Suter, *Versatile, scalable, and accurate simulation of distributed applications and platforms*, *Journal of Parallel and Distributed Computing* **74** (Oct., 2014) 2899–2917.

[195] R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. F. De Rose, and R. Buyya, *CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms*, *Software: Practice and Experience* **41** (Jan., 2011) 23–50.

[196] M. Nance Hall, K.-T. Foerster, S. Schmid, and R. Durairajan, *A survey of reconfigurable optical networks*, *Optical Switching and Networking* **41** (Sept., 2021) 100621.

[197] Chao Chen and A. Joshi, *Runtime management of laser power in silicon-photonic multibus NoC architecture*, *IEEE Journal of Selected Topics in Quantum Electronics* **19** (Mar., 2013) 3700713–3700713.

[198] Y. Demir and N. Hardavellas, *EcoLaser: An adaptive laser control for energy-efficient on-chip photonic interconnects*, in *Proceedings of the 2014 International Symposium on Low Power Electronics and Design (ISLPED)*, pp. 3–8, ACM, Aug., 2014.

[199] F. Lan, R. Wu, C. Zhang, Y. Pan, and K.-T. Cheng, *DLPS: Dynamic laser power scaling for optical network-on-chip*, in *2017 22nd Asia and South Pacific Design Automation Conference (ASP-DAC)*, pp. 726–731, IEEE, Jan., 2017.

[200] W. Liu, J. Xu, X. Wu, Y. Ye, X. Wang, W. Zhang, M. Nikdast, and Z. Wang, *A NoC traffic suite based on real applications*, in *2011 IEEE Computer Society Annual Symposium on VLSI*, pp. 66–71, IEEE, July, 2011.

[201] Z. Wang, W. Liu, J. Xu, B. Li, R. Iyer, R. Illikkal, X. Wu, W. H. Mow, and W. Ye, *A case study on the communication and computation behaviors of real applications in NoC-based MPSoCs*, in *2014 IEEE Computer Society Annual Symposium on VLSI*, pp. 480–485, IEEE, July, 2014.

[202] P. Dong, W. Qian, H. Liang, R. Shafiiha, D. Feng, G. Li, J. E. Cunningham, A. V. Krishnamoorthy, and M. Asghari, *Thermally tunable silicon racetrack resonators with ultralow tuning power*, *Optics Express* **18** (Sept., 2010) 20298.

[203] W. Bogaerts, P. De Heyn, T. Van Vaerenbergh, K. De Vos, S. Kumar Selvaraja, T. Claes, P. Dumon, P. Bienstman, D. Van Thourhout, and R. Baets, *Silicon microring resonators*, *Laser & Photonics Reviews* **6** (Jan., 2012) 47–73.

[204] K. Padmaraju, D. F. Logan, X. Zhu, J. J. Ackert, A. P. Knights, and K. Bergman, *Integrated thermal stabilization of a microring modulator*, *Optics Express* **21** (June, 2013) 14342.

[205] M. Kennedy and A. K. Kodi, *Laser pooling: Static and dynamic laser power allocation for on-chip optical interconnects*, *Journal of Lightwave Technology* **35** (Aug., 2017) 3159–3167.

[206] M. Y. Teh, Z. Wu, and K. Bergman, *Flexspander: augmenting expander networks in high-performance systems with optical bandwidth steering*, *Journal of Optical Communications and Networking* **12** (Apr., 2020) B44.

[207] R. Wu, C.-h. Chen, C. Li, T.-C. Huang, F. Lan, C. Zhang, Y. Pan, J. E. Bowers, R. G. Beausoleil, and K.-t. Cheng, *Variation-aware adaptive tuning for nanophotonic interconnects*, in *2015 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pp. 487–493, IEEE, nov, 2015.

[208] L. H. K. Duong, M. Nikdast, J. Xu, Z. Wang, Y. Thonnart, S. L. Beux, P. Yang, X. Wu, and Z. Wang, *Coherent crosstalk noise analyses in ring-based optical interconnects*, in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 501–506, 2015.

[209] H. Li, A. Fourmigue, S. L. Beux, X. Letartre, I. O'Connor, and G. Nicolescu, *Thermal aware design method for VCSEL-based on-chip optical interconnect*, in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 1120–1125, 2015.

[210] P. Pintus, F. Gambini, S. Faralli, F. Di Pasquale, I. Cerutti, and N. Andriolli, *Ring versus bus: A theoretical and experimental comparison of photonic integrated NoC*, *Journal of Lightwave Technology* **33** (Dec., 2015) 4870–4877.

[211] "Virtuoso Analog Design Environment." `https://www.cadence.com/en_US/home/tools/custom-ic-analog-rf-design/circuit-design/virtuoso-analog-design-environment.html`. Accessed: 2021-06-22.

[212] A. Fang, M. Sysak, B. Koch, R. Jones, E. Lively, Ying-Hao Kuo, Di Liang, O. Raday, and J. Bowers, *Single-wavelength silicon evanescent lasers*, *IEEE Journal of Selected Topics in Quantum Electronics* **15** (2009), no. 3 535–544.

[213] S. R. Jain, Yongbo Tang, Hui-Wen Chen, M. N. Sysak, and J. E. Bowers, *Integrated hybrid silicon transmitters*, *Journal of Lightwave Technology* **30** (Mar., 2012) 671–678.

[214] C. Zhang, S. Srinivasan, Y. Tang, M. J. R. Heck, M. L. Davenport, and J. E. Bowers, *Low threshold and high speed short cavity distributed feedback hybrid silicon lasers*, *Optics Express* **22** (May, 2014) 10202.

[215] G. P. Agrawal and N. K. Dutta, *Semiconductor Lasers*. Springer US, Boston, MA, 1995.

[216] A. Fang, B. Koch, R. Jones, E. Lively, Di Liang, Ying-Hao Kuo, and J. Bowers, *A distributed Bragg reflector silicon evanescent laser*, *IEEE Photonics Technology Letters* **20** (Oct., 2008) 1667–1669.

[217] C. Zhang, D. Liang, Cheng Li, G. Kurczveil, J. E. Bowers, and R. G. Beausoleil, *High-speed hybrid silicon microring lasers*, in *2015 IEEE 58th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pp. 1–4, IEEE, Aug., 2015.

[218] Y.-h. Kuo, H.-W. Chen, and J. E. Bowers, *High speed hybrid silicon evanescent electroabsorption modulator*, *Optics Express* **16** (June, 2008) 9936.

[219] G. Li, C. Sun, S. Pappert, W. Chen, and P. Yu, *Ultrahigh-speed traveling-wave electroabsorption modulator-design and analysis*, *IEEE Transactions on Microwave Theory and Techniques* **47** (July, 1999) 1177–1183.

[220] Y. Tang, *Study on electroabsorption modulators and grating couplers for optical interconnects*. PhD thesis, KTH, Microelectronics and Applied Physics, 2010.

[221] Z. Zhang, R. Wu, Y. Wang, C. Zhang, E. J. Stanton, C. L. Schow, K.-T. Cheng, and J. E. Bowers, *Compact modeling for silicon photonic heterogeneously integrated circuits*, *Journal of Lightwave Technology* **35** (July, 2017) 2973–2980.

[222] D. A. B. Miller, *Device requirements for optical interconnects to CMOS silicon chips*, in *Integrated Photonics Research, Silicon and Nanophotonics and Photonics in Switching*, p. PMB3, OSA, 2010.

[223] A. Y. Liu, C. Zhang, J. Norman, A. Snyder, D. Lubyshev, J. M. Fastenau, A. W. K. Liu, A. C. Gossard, and J. E. Bowers, *High performance continuous wave 1.3 μm quantum dot lasers on silicon*, *Applied Physics Letters* **104** (Jan., 2014) 041104.

[224] Cheng Li, Rui Bai, A. Shafik, E. Z. Tabasy, Geng Tang, Chao Ma, Chin-Hui Chen, Zhen Peng, M. Fiorentino, P. Chiang, and S. Palermo, *A ring-resonator-based silicon photonics transceiver with bias-based wavelength stabilization and adaptive-power-sensitivity receiver*, in *2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers*, vol. 56, pp. 124–125, IEEE, Feb., 2013.

[225] Y. Wang, M. A. Seyedi, R. Wu, J. Hulme, M. Fiorentino, R. G. Beausoleil, and K.-T. Cheng, *Energy-efficient channel alignment of DWDM silicon photonic transceivers*, in *2018 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 601–604, IEEE, Mar., 2018.

[226] Y. Wang, M. A. Seyedi, J. Hulme, M. Fiorentino, R. G. Beausoleil, and K.-T. Cheng, *Bidirectional tuning of microring-based silicon photonic transceivers for optimal energy efficiency*, in *Proceedings of the 24th Asia and South Pacific Design Automation Conference*, pp. 370–375, ACM, Jan., 2019.

[227] Y. Wang, L. Shao, M. A. Lastras-Montaño, and K.-T. Cheng, *Taming emerging devices' variation and reliability challenges with architectural and system solutions [invited]*, in *2019 IEEE 32nd International Conference on Microelectronic Test Structures (ICMTS)*, pp. 90–95, IEEE, Mar., 2019.

[228] C.-H. Chen, M. A. Seyedi, M. Fiorentino, R. G. Beausoleil, D. Livshits, A. Gubenko, S. Mikhrin, and V. Mikhrin, *Concurrent multi-channel transmission of a DWDM silicon photonic transmitter based on a comb laser and microring modulators*, in *2015 International Conference on Photonics in Switching (PS)*, pp. 175–177, IEEE, Sept., 2015.

[229] M. A. Seyedi, C. H. Chen, M. Fiorentino, D. Livshits, A. Gubenko, S. Mikhrin, V. Mikhrin, and R. G. Beausoleil, *Concurrent DWDM transmission with ring modulators driven by a comb laser with 50GHz channel spacing*, in *2016 21st OptoElectronics and Communications Conference (OECC) - Held Jointly with 2016 International Conference on Photonics in Switching (PS)*, pp. 9–11, 2016.

[230] R. Polster, Y. Thonnart, G. Waltener, J.-L. Gonzalez, and E. Cassan, *Efficiency optimization of silicon photonic links in 65-nm CMOS and 28-nm FDSOI technology nodes*, *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* **24** (Dec., 2016) 3450–3459.

[231] "Home - STMicroelectronics." `https://www.st.com/content/st_com/en.html`. Accessed: 2021-06-29.

[232] B. E. Little, J.-P. Laine, and S. T. Chu, *Surface-roughness-induced contradirectional coupling in ring and disk resonators*, *Optics Letters* **22** (Jan., 1997) 4.

[233] A. Li, T. Van Vaerenbergh, P. De Heyn, P. Bienstman, and W. Bogaerts, *Backscattering in silicon microring resonators: a quantitative analysis*, *Laser & Photonics Reviews* **10** (May, 2016) 420–431.

[234] T. F. Coleman and Y. Li, *An interior trust region approach for nonlinear minimization subject to bounds*, *SIAM Journal on Optimization* **6** (May, 1996) 418–445.

[235] P. J. Huber, *Robust statistics*, in *International Encyclopedia of Statistical Science* (M. Lovric, ed.), pp. 1248–1251. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[236] G. Upton and I. Cook, *Understanding statistics*. Oxford University Press, 1996.

[237] S. Rabanser, O. Shchur, and S. Günnemann, *Introduction to tensor decompositions and their applications in machine learning*, arXiv:1711.10781.

[238] E. Acar, D. M. Dunlavy, T. G. Kolda, and M. Mørup, *Scalable tensor factorizations for incomplete data*, *Chemometrics and Intelligent Laboratory Systems* **106** (Mar., 2011) 41–56.

[239] Q. Zhao, G. Zhou, L. Zhang, A. Cichocki, and S.-I. Amari, *Bayesian robust tensor factorization for incomplete multiway data*, *IEEE Transactions on Neural Networks and Learning Systems* **27** (Apr., 2016) 736–748.

[240] T. Yokota, Q. Zhao, and A. Cichocki, *Smooth PARAFAC decomposition for tensor completion*, *IEEE Transactions on Signal Processing* **64** (Oct., 2016) 5423–5436.

[241] L. Yuan, C. Li, D. Mandic, J. Cao, and Q. Zhao, *Tensor ring decomposition with rank minimization on latent space: An efficient approach for tensor completion*, *Proceedings of the AAAI Conference on Artificial Intelligence* **33** (July, 2019) 9151–9158.

[242] K. Yu, C. Li, H. Li, A. Titriku, A. Shafik, B. Wang, Z. Wang, R. Bai, C.-H. Chen, M. Fiorentino, P. Y. Chiang, and S. Palermo, *A 25 Gb/s hybrid-integrated silicon photonic source-synchronous receiver with microring wavelength stabilization*, *IEEE Journal of Solid-State Circuits* **51** (Sept., 2016) 2129–2141.

[243] M. Bahadori, S. Rumley, R. Polster, A. Gazman, M. Traverso, M. Webster, K. Patel, and K. Bergman, *Energy-performance optimized design of silicon photonic interconnection networks for high-performance computing*, in *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2017*, pp. 326–331, IEEE, Mar., 2017.

[244] Q. Li, N. Ophir, L. Xu, K. Padmaraju, L. Chen, M. Lipson, and K. Bergman, *Experimental characterization of the optical-power upper bound in a silicon microring modulator*, in *2012 Optical Interconnects Conference (OI)*, vol. 5, pp. 38–39, IEEE, May, 2012.

[245] C.-H. Chen, M. Ashkan Seyedi, M. Fiorentino, D. Livshits, A. Gubenko, S. Mikhrin, V. Mikhrin, and R. G. Beausoleil, *A comb laser-driven DWDM silicon photonic transmitter based on microring modulators*, *Optics Express* **23** (Aug., 2015) 21541.

[246] M. Bahadori, S. Rumley, D. Nikolova, and K. Bergman, *Comprehensive design space exploration of silicon photonic interconnects*, *Journal of Lightwave Technology* **34** (June, 2016) 2975–2987.

[247] K. Bergman, *Nanophotonic interconnection networks for energy-performance optimized computing*, in *Salishan Conference on High Speed Computing*, 2012.

[248] P. Dong, W. Qian, H. Liang, R. Shafiiha, N.-N. Feng, D. Feng, X. Zheng, A. V. Krishnamoorthy, and M. Asghari, *Low power and compact reconfigurable multiplexing devices based on silicon microring resonators*, *Optics Express* **18** (May, 2010) 9852.

[249] B. Szelag, M. A. Seyedi, A. Myko, B. Blampey, A. Descos, C.-h. Chen, S. Brision, F. Gays, M. Fiorentino, R. Beausoleil, and C. Kopp, *Integration and modeling of photonic devices suitable for high performance computing and data center applications*, in *PROCEEDINGS OF SPIE* (G. T. Reed and A. P. Knights, eds.), vol. 10108, p. 1010819, Feb., 2017.

[250] X. Zheng, F. Liu, J. Lexau, D. Patil, G. Li, Y. Luo, H. D. Thacker, I. Shubin, J. Yao, K. Raj, R. Ho, J. E. Cunningham, and A. V. Krishnamoorthy, *Ultralow power 80 Gb/s arrayed CMOS silicon photonic transceivers for WDM optical links*, *Journal of Lightwave Technology* **30** (Feb., 2012) 641–650.

[251] H. Jayatilleka, K. Murray, M. Á. Guillén-Torres, M. Caverley, R. Hu, N. A. F. Jaeger, L. Chrostowski, and S. Shekhar, *Wavelength tuning and stabilization of microring-based filters using silicon in-resonator photoconductive heaters*, *Optics Express* **23** (Sept., 2015) 25084.

[252] J. F. Buckwalter, X. Zheng, G. Li, K. Raj, and A. V. Krishnamoorthy, *A monolithic 25-Gb/s transceiver with photonic ring modulators and Ge detectors in a 130-nm CMOS SOI process*, *IEEE Journal of Solid-State Circuits* **47** (June, 2012) 1309–1322.

[253] C. Sun, M. Wade, M. Georgas, S. Lin, L. Alloatti, B. Moss, R. Kumar, A. H. Atabaki, F. Pavanello, J. M. Shainline, J. S. Orcutt, R. J. Ram, M. Popovic, and V. Stojanovic, *A 45 nm CMOS-SOI monolithic photonics platform with bit-statistics-based resonant microring thermal tuning*, *IEEE Journal of Solid-State Circuits* **51** (Apr., 2016) 893–907.

[254] "Leti (English) - About Leti." `https://www.leti-cea.com/cea-tech/leti/english/Pages/Leti/About-Leti/about-leti.aspx`. Accessed: 2021-06-30.

[255] R. M. Karp, *Reducibility among combinatorial problems*, in *Complexity of Computer Computations* (R. E. Miller, J. W. Thatcher, and J. D. Bohlinger, eds.), pp. 85–103. Springer US, Boston, MA, 1972.

[256] "Genetic Algorithm - MATLAB & Simulink." `https://www.mathworks.com/help/gads/genetic-algorithm.html`. Accessed: 2021-07-01.

[257] P. Grani and S. Bartolini, *Design options for optical ring interconnect in future client devices*, *ACM Journal on Emerging Technologies in Computing Systems* **10** (May, 2014) 1–25.

[258] S. Werner, J. Navaridas, and M. Luján, *A survey on optical network-on-chip architectures*, *ACM Computing Surveys* **50** (Jan., 2018) 1–37.

[259] J. Bashir, E. Peter, and S. R. Sarangi, *A survey of on-chip optical interconnects*, *ACM Computing Surveys* **51** (Feb., 2019) 1–34.

[260] G. Hendry, E. Robinson, V. Gleyzer, J. Chan, L. P. Carloni, N. Bliss, and K. Bergman, *Time-division-multiplexed arbitration in silicon nanophotonic networks-on-chip for high-performance chip multiprocessors*, *Journal of Parallel and Distributed Computing* **71** (May, 2011) 641–650.

[261] C. W. Commander, *Maximum cut problem, MAX-CUT*, in *Encyclopedia of Optimization* (C. A. Floudas and P. M. Pardalos, eds.), pp. 1991–1999. Springer US, Boston, MA, 2009.

[262] K. Andreev and H. Räcke, *Balanced graph partitioning*, in *Proceedings of the sixteenth annual ACM symposium on Parallelism in algorithms and architectures - SPAA '04*, vol. 16, p. 120, ACM Press, 2004.

[263] M. R. Garey and D. S. Johnson, *Computers and Intractability; A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., USA, 1990.

[264] B. L. Chamberlain, *Graph partitioning algorithms for distributing workloads of parallel computations*, tech. rep., Univ. of Washington, 1998.

[265] R. Baños, C. Gil, M. G. Montoya, and J. Ortega, *A new Pareto-based algorithm for multi-objective graph partitioning*, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (C. Aykanat, T. Dayar, and İ. Körpeoğlu, eds.), vol. 3280, pp. 779–788. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[266] K. Aydin, M. Bateni, and V. Mirrokni, *Distributed balanced partitioning via linear embedding*, *Algorithms* **12** (Aug., 2019) 162.

[267] D. Bertsimas and J. Tsitsiklis, *Simulated annealing*, *Statistical Science* **8** (Feb., 1993) 10–15.

[268] P. Czyżak and A. Jaszkiewicz, *Pareto simulated annealing*, in *Multiple Criteria Decision Making* (G. Fandel and T. Gal, eds.), vol. 448 of *Lecture Notes in Economics and Mathematical Systems*, pp. 297–307. Springer Berlin Heidelberg, Berlin, Heidelberg, 1997.

[269] Y. Wang, J. Hulme, P. Sun, M. Jain, M. A. Seyedi, M. Fiorentino, R. G. Beausoleil, and K.-T. Cheng, *Characterization and applications of spatial variation models for silicon microring-based optical transceivers*, in *2020 57th ACM/IEEE Design Automation Conference (DAC)*, pp. 1–6, IEEE, July, 2020.

[270] T. G. Kolda and B. W. Bader, *Tensor decompositions and applications*, *SIAM Review* **51** (Aug., 2009) 455–500.

[271] C. Coello Coello and M. Lechuga, *MOPSO: a proposal for multiple objective particle swarm optimization*, in *Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No.02TH8600)*, vol. 2, pp. 1051–1056, IEEE, 2002.

[272] W. Wolf, A. Jerraya, and G. Martin, *Multiprocessor system-on-chip (MPSoC) technology*, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **27** (Oct., 2008) 1701–1713.

[273] Chao Chen, J. L. Abellan, and A. Joshi, *Managing laser power in silicon-photonic NoC through cache and NoC reconfiguration*, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **34** (June, 2015) 972–985.

[274] L. A. Coldren, S. W. Corzine, and M. L. Mašanović, *Dynamic effects*, in *Diode Lasers and Photonic Integrated Circuits* [22], ch. 5, pp. 247–333.

[275] S. Murali and G. De Micheli, *SUNMAP: a tool for automatic topology selection and generation for NoCs*, in *Proceedings of the 41st annual conference on Design automation (DAC)*, p. 914, ACM Press, 2004.

[276] S. Saeidi, A. Khademzadeh, and A. Mehran, *SMAP: An intelligent mapping tool for network on chip*, in *2007 International Symposium on Signals, Circuits and Systems (ISSCS)*, vol. 1, pp. 1–4, IEEE, July, 2007.

[277] L. Bononi, N. Concer, M. Grammatikakis, M. Coppola, and R. Locatelli, *NoC topologies exploration based on mapping and simulation models*, in *10th Euromicro Conference on Digital System Design Architectures, Methods and Tools (DSD)*, pp. 543–546, IEEE, Aug., 2007.

[278] E. Fusella and A. Cilardo, *PhoNoCMap: An application mapping tool for photonic networks-on-chip*, in *2016 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, p. 289–292, 2016.

[279] E. Fusella and A. Cilardo, *Reducing power consumption of lasers in photonic NoCs through application-specific mapping*, *ACM Journal on Emerging Technologies in Computing Systems* **14** (July, 2018) 1–11.

[280] C. Xu, F. X. Lin, Y. Wang, and L. Zhong, *Automated os-level device runtime power management*, in *Proceedings of the Twentieth International Conference on Architectural Support for Programming Languages and Operating Systems*, pp. 239–252, ACM, Mar., 2015.

[281] D. A. Reed and J. Dongarra, *Exascale computing and big data*, *Communications of the ACM* **58** (June, 2015) 56–68.

[282] G. Kurczveil, A. Descos, D. Liang, M. Fiorentino, and R. Beausoleil, *Hybrid silicon quantum dot comb laser with record wide comb width*, in *Frontiers in Optics / Laser Science*, vol. 1, p. FTu6E.6, OSA, 2020.

[283] D. Liang, G. Kurczveil, Z. Huang, B. Wang, A. Descos, S. Srinivasan, Y. Hu, X. Zeng, W. V. Sorin, S. Cheung, S. Liu, P. Sun, T. Van Vaerenbergh, M. Fiorentino, J. E. Bowers, and R. G. Beausoleil, *Integrated green DWDM photonics for next-gen high-performance computing*, in *Optical Fiber Communication Conference (OFC)*, p. Th1E.2, OSA, 2020.

[284] G. Michelogiannakis, Y. Shen, M. Y. Teh, X. Meng, B. Aivazi, T. Groves, J. Shalf, M. Glick, M. Ghobadi, L. Dennison, and K. Bergman, *Bandwidth steering in HPC using silicon nanophotonics*, in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–25, ACM, Nov., 2019.

[285] A. H. Ahmed, A. Sharkia, B. Casper, S. Mirabbasi, and S. Shekhar, *Silicon-photonics microring links for datacenters—challenges and opportunities*, *IEEE Journal of Selected Topics in Quantum Electronics* **22** (Nov., 2016) 194–203.

[286] H. Li, A. Fourmigue, S. Le Beux, I. O'Connor, and G. Nicolescu, *Towards maximum energy efficiency in nanophotonic interconnects with thermal-aware on-chip laser tuning*, *IEEE Transactions on Emerging Topics in Computing* **6** (July, 2018) 343–356.

[287] Y. Wang, P. Sun, J. Hulme, M. A. Seyedi, M. Fiorentino, R. G. Beausoleil, and K.-T. Cheng, *Energy efficiency and yield optimization for optical interconnects via transceiver grouping*, *Journal of Lightwave Technology* **39** (Mar., 2021) 1567–1578.

[288] A. K. Kodi, B. Neel, and W. C. Brantley, *Photonic interconnects for exascale and datacenter architectures*, *IEEE Micro* **34** (Sept., 2014) 18–30.

[289] Y. Shen, X. Meng, Q. Cheng, S. Rumley, N. Abrams, A. Gazman, E. Manzhosov, M. S. Glick, and K. Bergman, *Silicon photonics for extreme scale systems*, *Journal of Lightwave Technology* **37** (Jan., 2019) 245–259.

[290] Qiang Wu, Philo Juang, M. Martonosi, L.-s. Peh, and D. Clark, *Formal control techniques for power-performance management*, *IEEE Micro* **25** (Sept., 2005) 52–62.

[291] M. Borghi, D. Bazzanella, M. Mancinelli, and L. Pavesi, *On the modeling of thermal and free carrier nonlinearities in silicon-on-insulator microring resonators*, *Optics Express* **29** (Feb., 2021) 4363.

[292] H. Casanova, R. Ferreira da Silva, R. Tanaka, S. Pandey, G. Jethwani, W. Koch, S. Albrecht, J. Oeth, and F. Suter, *Developing accurate and scalable simulators of production workflow management systems with WRENCH*, *Future Generation Computer Systems* **112** (Nov., 2020) 162–175.

[293] Alibaba Group, "Alibaba Cluster Trace Program." `https://github.com/alibaba/clusterdata/blob/master/cluster-trace-v2018/trace_2018.md`, 2019. Accessed: 2021-05-10.

[294] Y. Wang and K.-T. Cheng, *Task mapping-assisted laser power scaling for optical network-on-chips*, in *2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pp. 1–6, IEEE, Nov., 2019.

[295] D. Abts, M. R. Marty, P. M. Wells, P. Klausler, and H. Liu, *Energy proportional datacenter networks*, in *Proceedings of the 37th annual international symposium on Computer architecture - ISCA '10*, p. 338, ACM Press, 2010.

[296] G. Varsamopoulos and S. K. S. Gup, *Energy proportionality and the future: Metrics and directions*, in *2010 39th International Conference on Parallel Processing Workshops*, pp. 461–467, IEEE, Sept., 2010.

[297] P. Ruiu, C. Fiandrino, P. Giaccone, A. Bianco, D. Kliazovich, and P. Bouvry, *On the energy-proportionality of data center networks*, *IEEE Transactions on Sustainable Computing* **2** (Apr., 2017) 197–210.

[298] N. Parsons and N. Calabretta, *Optical switching for data center networks*, in Mukherjee *et. al.* [309], pp. 795–825.

[299] Chao Su, Lian-Kuan Chen, and Kwok-Wai Cheung, *Theory of burst-mode receiver and its applications in optical multiaccess networks*, *Journal of Lightwave Technology* **15** (Apr., 1997) 590–606.

[300] X.-Z. Qiu, *[OFC 2013 Tutorial OW3G.4] burst-mode receiver technology for short synchronization*, in *Optical Fiber Communication Conference (OFC)/National Fiber Optic Engineers Conference 2013*, p. OW3G.4, OSA, 2013.

[301] J. E. Smith, *A study of branch prediction strategies*, in *25 years of the international symposia on Computer architecture (selected papers) - ISCA '98*, pp. 202–215, ACM Press, 1998.

[302] T.-Y. Yeh and Y. N. Patt, *Two-level adaptive training branch prediction*, in *Proceedings of the 24th annual international symposium on Microarchitecture - MICRO 24*, pp. 51–61, ACM Press, 1991.

[303] N. Ahmed, A. L. C. Barczak, T. Susnjak, and M. A. Rashid, *A comprehensive performance analysis of Apache Hadoop and Apache Spark for large scale data sets using HiBench*, *Journal of Big Data* **7** (Dec., 2020) 110.

[304] M. Chowdhury, M. Zaharia, J. Ma, M. I. Jordan, and I. Stoica, *Managing data transfers in computer clusters with orchestra*, *ACM SIGCOMM Computer Communication Review* **41** (Oct., 2011) 98–109.

[305] "Bayesian Optimization Algorithm - MATLAB & Simulink." `https://www.mathworks.com/help/stats/bayesian-optimization-algorithm.html`. Accessed: 2021-07-10.

[306] M. Frean and P. Boyle, *Using Gaussian processes to optimize expensive functions*, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (W. Wobcke and M. Zhang, eds.),

vol. 5360 LNAI of *Lecture Notes in Computer Science*, pp. 258–267. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[307] J. Kim, W. Dally, S. Scott, and D. Abts, *Cost-efficient Dragonfly topology for large-scale systems*, *IEEE Micro* **29** (Jan., 2009) 33–40.

[308] "Platform Examples — SimGrid documentation." https://simgrid.org/doc/latest/Platform_examples.html. Accessed: 2021-05-25.

[309] B. Mukherjee, I. Tomkos, M. Tornatore, P. Winzer, and Y. Zhao, eds., *Springer Handbook of Optical Networks*. Springer Handbooks. Springer International Publishing, Cham, 2020.