# TOWARDS RELIABLE NANOPHOTONIC INTERCONNECTION NETWORK DESIGNS

by

**Yi Xu**

B.S., Nanjing University, 2004

M.S., Nanjing University, 2007

Submitted to the Graduate Faculty of

the Swanson School of Engineering in partial fulfillment

of the requirements for the degree of

**Doctor of Philosophy**

University of Pittsburgh

2012

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This dissertation was presented

by

Yi Xu

It was defended on

November 19th 2012

and approved by

Jun Yang, Ph.D., Associate Professor, Department of Electrical and Computer Engineering

Youtao Zhang, Ph.D., Associate Professor, Department of Computer Science

Yiran Chen, Ph.D., Assistant Professor, Department of Electrical and Computer Engineering

Steven P. Levitan, Ph.D., Professor, Department of Electrical and Computer Engineering

Guangyong Li, Ph.D., Assistant Professor, Department of Electrical and Computer Engineering

Rami Melhem, Ph.D, Professor, Department of Computer Science

Dissertation Co-directors: Jun Yang, Ph.D., Associate Professor, Department of Electrical and Computer Engineering,

Youtao Zhang, Ph.D., Associate Professor, Department of Computer Science

**TOWARDS RELIABLE NANOPHOTONIC INTERCONNECTION NETWORK DESIGNS**

Yi Xu, PhD

University of Pittsburgh, 2012

As technology scales into deep submicron domains, electrical wires start to face critical challenges in latency and power since they do not scale well as compared to transistors. Many recent researches have shifted focus to optical on-chip interconnection because of its promises of high bandwidth density, low propagation delay, distance-independent power consumption (compared to metal), and natural support for multicast and broadcast.

Unfortunately, while optical interconnect provides many attractive features, there are also fundamental challenges in fabrication of those devices to providing robust and reliable on-chip communication. Microrings resonators, the basic components of nanophotonic interconnect, may not resonate at the designated wavelength under fabrication errors (a.k.a. process variations PV) or thermal fluctuation (TF), leading to communication errors and bandwidth loss. In addition, the power overhead required to correct the drift can overturn the benefits promised by this new technology. Hence, the objective of the thesis is to maximize network bandwidth through proper arrangement among microrings and wavelengths with minimum tuning power requirement. I propose the following techniques to achieve my goals. First, I will present a series of solutions, called "MinTrim", to address the wavelength drifting problem of microrings and subsequent bandwidth loss problem of an optical network, due to the PV. Next, to mitigate bandwidth loss and performance degradation caused by PV and TF, I will propose an architecture-level approach, "BandArb", which allocates the bandwidth at runtime according to network demands and temperature with low computation overhead. Finally, I will conclude the thesis and discuss the future works in this field.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

## PREFACE

Foremost, I am in deep gratitude for my thesis advisors, Dr.Jun Yang and Dr.Youtao Zhang for their continuous supports for my research work, study and life in U.S. They have guided me into the exciting research area on computer architecture since I came here. Their enthusiasm and serious attitude on research have set a great example for me to follow. I benefit a lot from their guidance on thesis writing, idea formulation, presentations, etc. everything required for a Ph.D. and constructive suggestions on life and career path. I could not have imagined having a better advisor and mentor for my Ph.D study.

I would like to thank Dr.Rami Melhem for his invaluable guidance, inspiration and stimuli that enable me to finish the research work presented in this thesis. I would like to thank the rest of my thesis committees: Dr.Yiran Chen, Dr.Steven P. Levitan and Dr.Guangyong Li for their encouragement, insightful comments and guidance.

I would like to thank my fellow labmates: Lei Jiang, Bo Zhao, Lin Li, Xiuyi Zhou, Ping Zhou, Yu Du, Weijia Li and so on for the tremendous helps, exciting discussions in the reading group and joyful time we had together in Pittsburgh.

Last but not the least, I would also like to acknowledge my parents, my family and friends for supporting and accompanying me through all these years and shaping me to be who I am now.

# 1.0 INTRODUCTION

Recent technology scaling has enabled the integration of billions of transistors on-chip. Due to increasing design complexity and diminishing return of utilizing on-chip transistors in uniprocessor design, chip multiprocessor (CMP) has emerged as a promising microarchitecture for keeping up performance with integration density [21, 47]. With the proliferation of CMPs, on-chip interconnection networks start to play a more and more important role in determining the performance and power of the entire chip [35].

However, electrical on-chip networks are hitting great challenges in power, latency and bandwidth density with technology scaling [22, 23]. Figure 1 shows that even with optimized design, the delay of electrical wires per unit length is still increasing while logic gates are becoming faster. The performance of electrical interconnects is lagging behind transistor performance.

Such challenges are especially pronounced in the era of multi-core computing where high bandwidth, low power, and low-latency global transmission are required. Additionally, it is difficult to improve the memory bandwidth substantially with traditional interconnection technology due to the limited number of I/O pins and tight power constraint on data transmission.

Fortunately, breakthroughs in nanophotonic technology has provided computer architects with an alternative for both on-chip and off-chip communication since optical networks have the advantages of bandwidth density (larger by up to 2 orders of magnitude [6]), energy-efficiency and propagation delay over the electrical counterparts, as summarized in Figure 2, which show the comparisons on relative latency, power and bandwidth density for electrical and optical links, respectively. It indicates that optical interconnects outperform the conventional electrical link in all the three aspects.

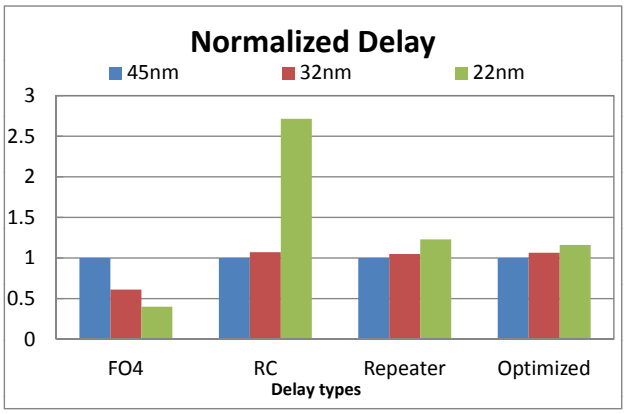



Figure 1: Gate, global wire(RC), global wire (repeater) and global wire (optimized+repeatered) delay [17]

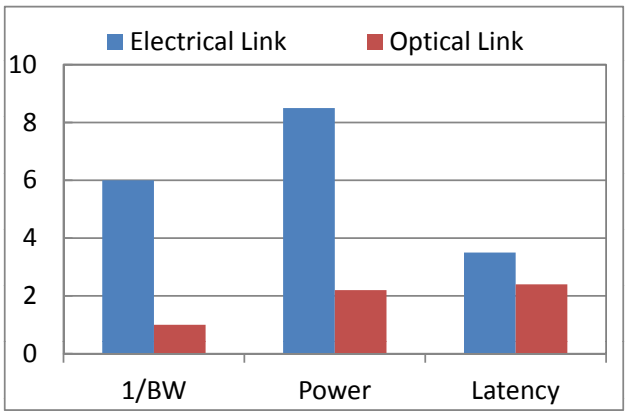



Figure 2: Relative latency, power and spatial bandwidth comparison chart for electrical and idealized optical link at 32nm technology node (in relative scale) [17]

These promising advantages attracted researchers to investigate designs that leverage nanophotonic technology for on-chip networks [16, 25, 29, 36, 50, 51, 67, 68, 69, 76, 78], as well as chip-to-chip communication, CPU-to-memory communication and high radix switches [26, 6, 10, 33, 32, 64].

## 1.1 CHALLENGES OF OPTICAL NETWORK

While optical interconnect provides many promising features, there are also fundamental challenges in integration and fabrication of those devices to providing robust and reliable on-chip communication. Among many challenges, the *thermal sensitivity* and *process variations* (PV) of silicon photonic devices are the key difficulties.

Thermal sensitivity refers to the changes in refractive index of optical components, e.g. photonic microring ($\mu$ring) resonator, due to temperature fluctuations, such that those components fail to resonate designated wavelengths in the waveguide. Studies have reported that $\mu$ring's resonance wavelength typically drifts by $\sim$0.1nm/°C [55, 54, 80], while chip temperature could fluctuate well beyond 30°C.

PV refers to variations of critical physical dimensions, e.g. thickness of silicon, width of waveguide, caused by lithography imperfection and etch non-uniformity of devices [59]. Those variations will directly affect the resonant wavelengths of a $\mu$ring [26, 48, 58, 73], a critical optical component used as a modulator, a filter or a switching element. Although there has not been clear characterization of wavelength drifts of $\mu$rings due to PV (termed PV-drift for short), several recent laboratory measurements have reported that they are indeed quite significant. For example, as much as $\sim$4.79nm of PV-drift within a wafer has been observed in a demonstration of a photonic platform leveraging the state-of-the-art CMOS foundry infrastructure [48]. A recent work [59] has also reported a standard deviation of 0.55nm for two $\mu$rings that are only 1.7mm apart. In a wavelength division multiplexing (WDM) enabled optical interconnect, the spacing between adjacent wavelengths, denoted as $\Delta\lambda$, is $\sim$0.8nm [61] or lower [17, 45]. A previous study shows that when PV-drift is over 1/3 of $\Delta\lambda$, the bit-error-rate of optical transmission would increase from $10^{-12}$ to $10^{-6}$ [39].

Larger PV-drifts and thermal variations would bring the $\mu$ring to resonate at a completely different wavelength that is several channels away. As a result, drifted $\mu$rings cannot be used for communication since they will create erroneous signals. Hence, network nodes that do not have all working $\mu$rings would lose bandwidth in communication.

## 1.2 CURRENT TECHNIQUES AND LIMITATIONS

At present, there are two types of techniques that can restore the resonance frequency of $\mu$rings. The first type is post-fabrication physical trimming, where high-energy particles such as UV light or electron beam is used to adjust the refractive index of $\mu$rings [20, 34, 44, 65] or effective refractive index of the waveguide [58] to achieve resonance correction. However, such techniques require trimming to be carefully tuned for individual $\mu$ring. Given that the number of $\mu$rings on-chip is on the order of thousands to millions [69, 51, 26, 2, 30], it is unclear if such physical trimming is practical for volume production. In addition, physical trimming may create degradation of the quality factor, "Q", of a $\mu$ring, bouncing of corrected wavelength, and faster aging of the trimmed devices [58].

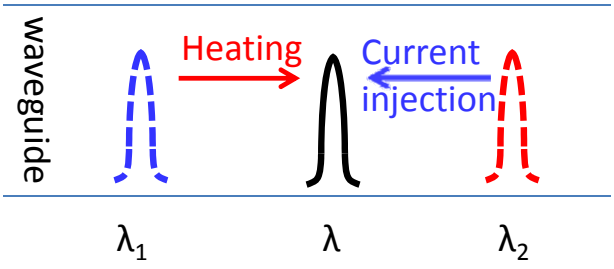



Figure 3: Power trimming method. $\lambda$ indicates the nominal wavelength of $\mu$ring, $\lambda_1$ and $\lambda_2$ stands for the drifted resonant wavelength caused by PV or TF.

The second type of techniques for restoring the resonance frequency is power trimming, in which heating or current injection into a $\mu$ring is used to correct its resonance wavelength. The former causes the wavelength to shift towards the red end and the latter towards the

blue end of the resonance spectrum, as illustrated in Figure 3. Although power trimming could address the drifts introduced by both PV and thermal variations, it can result in significant power consumption so as to nullify the power advantage that ideal on-chip optical interconnects are projected to have [17, 45, 39, 15]. In addition, current injection has very limited correction range, as it would generate thermal runaway beyond the trimming range [17, 45, 39]. Nevertheless, power trimming has been considered necessary for tackling thermal sensitivity, as demonstrated in the "Sliding Ring Window" technique [45]. Hence, I will assume that power trimming is already in place for thermal sensitivity, and propose techniques to minimize the total tuning power required for correcting PV and thermal drifts in this thesis.

## 1.3 THESIS OVERVIEW

In my thesis, I plan to investigate the impacts of PV and TF on network performance and power consumption and propose an architectural methodology to salvage network bandwidth loss, both statically and dynamically. First, I will introduce the background knowledge and recent works on nanophotonics.

Next, I will propose a serial of approaches, named "MinTrim" to address PV-drifts by maximizing the number of usable wavelengths for all nodes, each wavelength being resonant with one $\mu$ring while minimizing the power required in trimming. The first step of "MinTrim" tackles the limitation of current injection, and trims a $\mu$ring to a nearby wavelength rather than the nominal one. Integer linear programming (ILP) is used to maximize the likelihood of successful trimming with minimum trimming power. The next step further mitigates PV-drifts by provisioning additional $\mu$rings in the ILP framework, which brings more opportunities to finding a nearby $\mu$ring that can be trimmed to a desirable wavelength. The last step allows flexible wavelength assignment for each network node, as long as each one can be allocated with enough wavelengths, to give more freedom to trimming. MinTrim can salvage most of the lost bandwidth in the two baseline designs and reduce significant trimming power.

Third, I will present a two-level design, called "BandArb" to handle the bandwidth loss caused by PV and TF. The goal is to find a balance between achievable bandwidth provisioning and computation latency such that it is justifiable to pay the calculation and $\mu$ring tuning overhead for the bandwidth improvement. Given that the ILP algorithms used in MinTrim [79] are not affordable at run-time, I propose to use a heuristic algorithm to approximate the effect of MinTrim locally within each node and a coarse grained arbitration algorithm that uses the results of local alignment algorithms to find a wavelength mapping that maximizes the utilization of the available $\mu$rings. Next, the fine granularity of BandArb applies a wavelength allocation approach to further improve the bandwidth. Since not all nodes are communicating all the time, communicating nodes have the opportunities to borrow reachable wavelengths that are assigned to other nodes via a distributed arbitration scheme. Thus, the utilization of $\mu$rings is improved and an active network node may utilize 100% of the bandwidth or even more when thermal $\mu$rings [45] are also used for transmission.

## 1.4 CONTRIBUTIONS

In summary, the contributions of this thesis are as follows:

- An overview of optical technology.
- A serial of approaches to maximize the static bandwidth via supplementary $\mu$rings with minimum power requirement.
- Modeling PV of $\mu$rings.
- Two architectural techniques to maximizing bandwidth utilization at runtime.

## 1.5 ROADMAP

The remainder of this thesis is organized as follows. Chapter 2 presents background. The proposed mechanisms are explained in Chapter 3 and Chapter 4. Chapter 5 concludes and describes future work.

# 2.0 OPTICAL TECHNOLOGY OVERVIEW

In the past few years, advances in nanophotonics [41, 5, 57] have enabled optical interconnect technologies with greater integration, smaller and CMOS-compatible optical devices and higher bandwidths. The latest ITRS predicts that on-chip optical link could be a potential replacement for global wires. In this chapter, I will introduce the background knowledge of optical interconnects and recent research works in this field.

## 2.1 OPTICAL INTERCONNECTS

A typical optical network includes off-chip laser source that provides on-chip light, waveguides that route optical signal, ring modulators that convert electrical signals to optical ones, and ring filters to detect lights and translate it into electrical signals. Figure 4 illustrates a dense wavelength division multiplexing (DWDM) nanophotonic link. Since light of different wavelengths can be transmitted and modulated in the single waveguide, DWDM technology enables multiple data channels per waveguide, providing high network bandwidth density. At the sender side, electrical signals are imprinted to laser light by wavelength-selective silicon modulators that absorb and pass the light for signal '0' and '1' respectively. For modulation, $\mu$ring resonators are typically preferred over other modulators due to their high modulation speed(10$\sim$20Gbps), low power(47 fJ/bit) and small footprint($\mu m^2$) [40, 53, 75]. The same ring structure can be used as a wavelength selective detector to extract light out of the waveguide, if the $\mu$ring is doped with a photo-detecting material such as CMOS-compatible germanium. The resonant light will be absorbed by the germanium and converted into electrical signal.

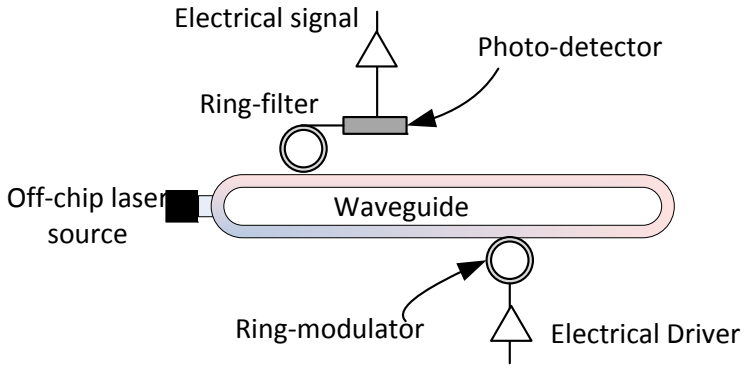



Figure 4: DWDM nanophotonic link.

The delays of various optical components at different technology nodes is summarized in Figure 5. Compared to the wire latency listed in Figure 1, the performance of optical link scales well with the technology.

Table 6 shows the energy/power data of optical and electrical interconnects. One of the benefits of optical link is that it only consume the power at the source and destination node, while the dynamic power of electrical wire increases with the length of transmission path.

On-chip laser source is also available. In a transmission system based on on-chip laser, VCSELs (Vertical Cavity Surface Emitting Laser) [63], where the modulator is not required. The light is emitted vertically, and then micro-mirrors transfer the light to the horizontal chip surface, which requires sophisticated lithographic technologies. The off-chip laser source is usually adopted in optical network design because of its saves on-chip power, area, and cost. The power of off-chip laser should be large enough to sustain all types of light loss such that the detector can receive sufficient optical power. The light losses of different optical modules are listed in table 1 Assuming $P_{PD}$ is the required power at the photo detector, and A is the attenuation of signal path, the minimum laser power per wavelength $P = P_{PD}10^{\frac{A}{10}}$ [45]. The link loss calculation starts at a photo detector and add all the attenuation losses along the way including the photo detector, waveguide, waveguide bends and intersection, coupling, on-resonance rings and off-resonance rings.

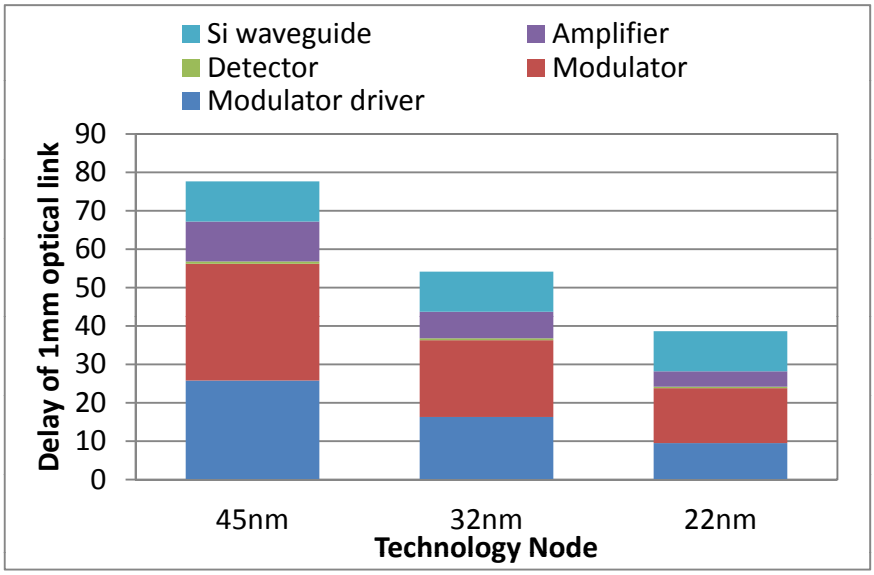



Figure 5: Delay breakdown for 1 mm optical link at different technology nodes [14]

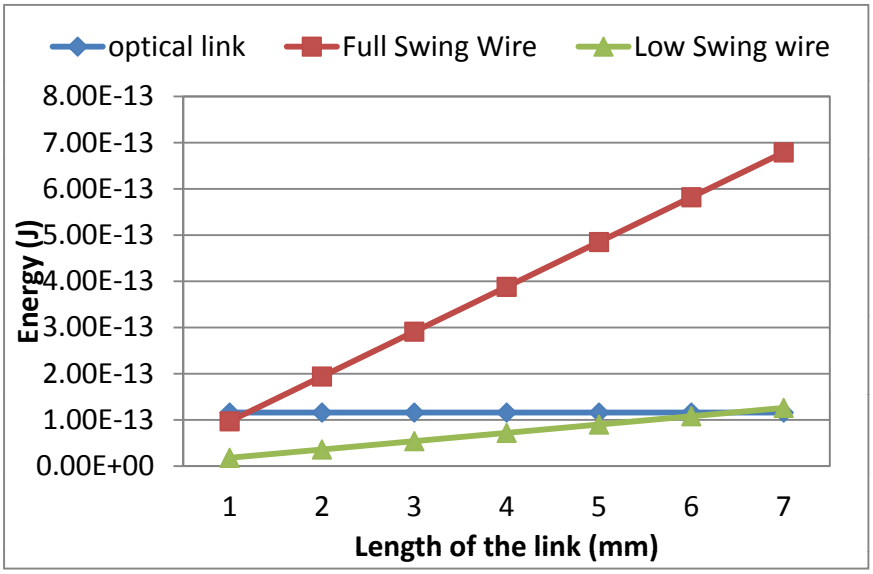



Figure 6: Comparisons on energy/power consumption of optical and electrical interconnects of different lengths [64, 71]

Table 1: Optical losses of different optical components [25, 30, 36].

| Photonic device | Loss (dB) | Photonic device | Loss (dB) |
|---|---|---|---|
| Waveguide loss | 0.3∼0.5/cm | Waveguide bend | 0.005 |
| Splitter | 0.2 | Coupler | 1 |
| Modulator insertion | 0.1∼1 | Detector insertion | 0.1 |
| Filter drop | 1.5∼3 | Ring through | 0.01∼0.001 |
| Laser efficiency | 30% | Detector sensitivity ($\mu w$) | 10 |

## 2.2 OPTICAL NETWORK ARCHITECTURE

### 2.2.1 Optical Crossbar Designs

There have also been studies exploiting the nanophotonic network topologies [16, 25, 29, 51, 50, 67] as well as nanophotonics interconnection for chip-to-chip communication [32]. In many cases, an optical crossbar is favored as the network backbone for cache coherence management [36, 68, 76] and data transmission due to its high bandwidth, natural support for broadcast, as well as short and uniform latency that simplifies protocol design.

**2.2.1.1 Network Category** I classify previous crossbar designs into two main categories: static and dynamical channel allocation. The channel defined here is a set of wavelengths used to transfer one flit (flit is the smallest unit of the transmission). The number of wavelengths per channel depends on the flit size and modulation speed of $\mu$rings. Crossbars using static channel allocation include single-write-multiple-read(SWMR) and multiple-write-single-read(MWSR). The microarchitectural designs of the crossbars are shown in Fig. 7(a) and (b), using a radix-4 crossbar as a simple example. $I_n$ and $O_n$ represent the sending and receiving interface of the optical router at node $n$. The different indices of rings in Fig. 7 indicate different optical channels. There are a total of 4 channels for the 4 network

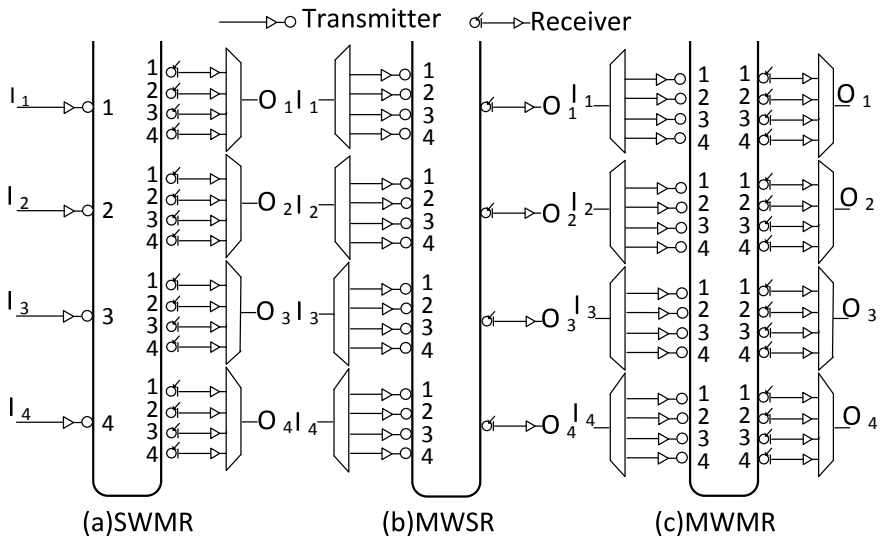



Figure 7: Crossbar microarchitectural design [4].

nodes in this example. Each node in SWMR has one dedicated channel to send data and can receive data from all channels. On the contrary, MWSR provides each node with a dedicated channel to read data and allows any node to write to the given channel. With exclusive sending channels, SWMR avoids starvation and does not need global arbitration to handle contention, which reduces design complexity and network latency. When traffic loads on the channels are evenly distributed, SWMR [29, 51] and MWSR [69] can perform well and provide high channel utilization. However, upon unbalanced traffic distribution, their dedicated channels will have low utilization and contribute little to the network throughput. Increasing throughput would require over provisioning of channels and causes proportional static power increase. Therefore the low channel utilization of SWMR and MWSR results in low energy efficiency.

Dynamic channel allocation design, e.g. multiple-write-multiple-read(MWMR) [50] shown in Fig. 7(c), can improve channel utilization and network throughput with channel sharing. In the figure, we can see that each network node can write to or read from any channel via more transmitters/receivers and MUXes than in SWMR and MWSR. Thus, under uneven traffic distribution, the nodes with high injection rate can utilize multiple channels to

Table 2: Power breakdowns of laser source and $\mu$ring trimming.

| Crossbar Designs | Laser Power | Trimming Power |
|---|---|---|
| Radix-32 SWMR | 35% [50] | 38% [50] |
| Radix-64 MWSR (Corona) | 5.4% [2] | 54% [2] |
| Radix-16 MWMR (FlexiShare) | 23% [50] | 42% [50] |

improve channel usage. However, as we can observe in Fig. 7(c), full channel sharing also requires more $\mu$ring resonators than SWMR or MWSR as in SWMR because every node is able to modulate light on all channels. Most of the time, the majority of $\mu$ring modulators are idle as only $M$ out of $NM$ ($N$ is the crossbar radix, $M$ is the number of channels) transmitters are used simultaneously. But idle $\mu$rings still consume significant trimming power which is proportional to $NM$ and cause more light losses in the waveguide. Hence, full sharing architectures also have low energy efficiency because of a large number of $\mu$rings.

**2.2.1.2 Power Consumption** For conventional electrical networks, dynamic power, which depends on the activities of routers and channels, typically dominates the total network power, whereas for optical networks, static power surpasses dynamic power and becomes dominant in the total network power. The static power of an optical network is mainly comprised of laser source power and $\mu$ring trimming power. The laser power is determined by the total number of wavelengths, the conversion efficiency from electrons to photons of the laser, and all types of transmission losses including both on-resonance $\mu$rings and scattered losses from off-resonance $\mu$rings [2]. Hence, laser power increases with the total number of $\mu$rings. The resonance wavelength of a $\mu$ring drifts with temperature variation. Such drift can be corrected, or *trimmed*, via either heating or carrier injection. Both methods consume power. Hence, the total power spent in trimming all $\mu$rings also increases with the total number of $\mu$rings. Recent studies have shown that laser and trimming power together contribute over 60% of the total on-chip network power, as shown in Table 2, which summarizes the

percentages of power spent in laser source and trimming for different crossbar designs. For example, the Corona network in 17-nm technology from HP [2, 69] is estimated to consume ∼26W in trimming $\mu$rings, out of ∼48W of the total network power. Even with optimistic $\mu$ring heating efficiency, e.g. using in-plane heaters and air-undercut [26, 25], it is estimated that $\mu$ring heating still consumes 38% of the total network power [50]. Hence, it is unwise to increase the throughput of an optical network through increasing the number of channels (and $\mu$rings) since the idle channels still consume significant static power. Instead, an *energy-efficient* optical network that achieves high throughput via improving the *channel utilization*, which does not increase static power, is preferred and should be developed in the future.

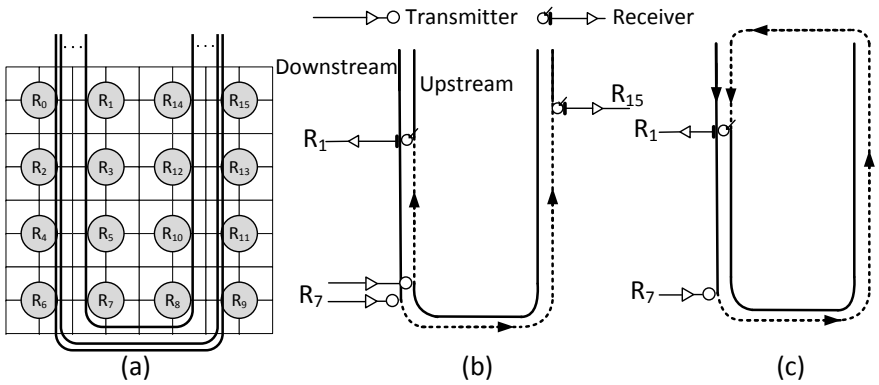



Figure 8: (a)Waveguide layout for a 16-node crossbar. (b) Single-Serpentine layout.(Data transmission: $R_7 \rightarrow R_1$ and $R_7 \rightarrow R_{15}$ via upstream and downstream channels)(c) Double-Serpentine layout.

**2.2.1.3 Waveguide Layout** I use a 16-node crossbar to show the physical layout of the optical network. In fig. 8, the ring-shaped waveguide connecting all 16 routers across the chip. There are two ways to implement a data channel, namely single-serpentine layout and double-serpentine layout [4]. They are illustrated in Fig. 8(b) and (c). The single-serpentine layout has two separate channels for upstream and downstream transmission. The direction of increasing router index is defined as downstream, otherwise the direction is defined as upstream. For example, in Fig. 8(b), $R_7$ sends message to $R_1$ through upstream channel and

to $R_{15}$ via downstream channel. The message path is illustrated by the dotted line. The alternative layout is double-serpentine that doubles the length of optical paths and lets the light traverses each node twice. During the first pass, the transmitter modulates the light to send a message. Then the receiver detects the light and converts it to digital signal in the second pass. The first option is usually adopted as it reduces the length of waveguide, the light loss and transmission latency. Hence, each channel actually is composed of two subchannels for upstream and downstream directions.

**2.2.1.4 Scalability** In typical small-scale CMPs, each tile is directly connected to a network node. However, this would be inappropriate for large-scale CMPs because the network size would be too large and for all-to-all network designs, the number of $\mu$rings increases quadratically with the network size. Also, each node's traffic injection rate is not very high because they are from a single core's private cache misses, indicating that such network is not very efficient. Therefore, one way to make crossbar design scalable is to employ the concentration or clustering technique to share the network channel among core-cache tiles [3, 24]. Downsizing the network reduces the number of $\mu$rings and static power cost. Determining an appropriate size of cluster represents the design trade-off between bandwidth and power in 1) the aggregated traffic load per cluster. If cluster size is large, then the bandwidth requirement within a cluster may be high for each optical router, resulting in contention delay and performance degradation; 2) the power consumed by $\mu$rings. Small-sized cluster leads to large network, which results in quadratic increase in number of rings; 3) the power consumed by laser source. More ring resonators on a waveguide will cause more energy loss during light propagation, which leads to higher laser power at the source.

While the optical links are utilized in global communication among clusters where long-range metal wires or multi-hop metal network are originally adopted, the intra-cluster network usually leverage metal connections as it is more power efficient for short-range traffics.

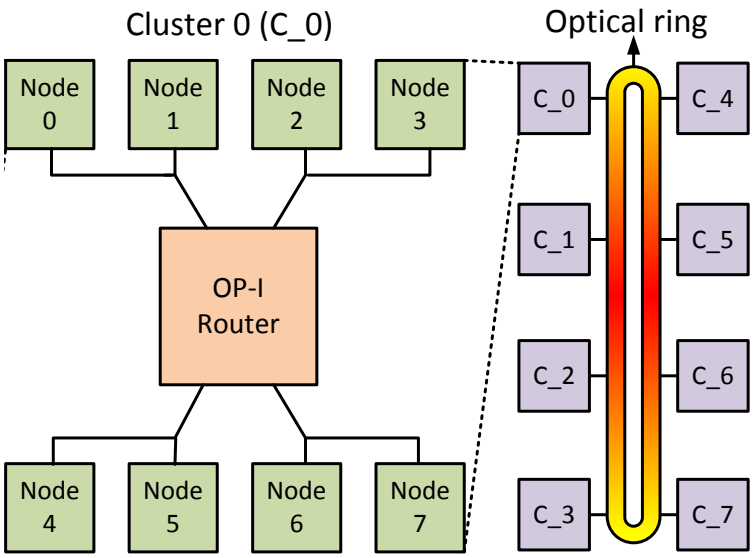



Figure 9: Hierarchical network architecture.

### 2.2.2 Optical Switch Designs

Global crossbar topology can provide contention-free communication, high performance and low design complexity [29], but large amount of $\mu$rings becomes a serious issue for high node count, requiring high laser and thermal tuning power. Many recent works build switch-based topologies such as Clos [25], mesh [16], etc. to reduce the number of optical devices. The wavelength-based routing [30] proposed a 2D torus topology with passive wavelength-routers. Retransmission is applied when multiple senders communicate with the same receiver at the same time, which brings long contention delay. Shacham et. al. [60] proposed a circuit-switch based photonic network that arranges large messages transmitted through optical network and small messages are delivered by electrical wires to improve overall energy efficiency. However, the setup time overcomes the benefit of optical transmission and makes it suitable for off-chip communications.

But there are two main challenges in designing the switch-based network. The first one is that optical crosstalk noise limits the scalability of optical network [72]. Crosstalk noise is caused by the undesirable coupling among optical signals when they pass $\mu$rings and

waveguide crossings. During crosstalk, a small portion of the power of one optical signal is directed to another optical signal and becomes noise. Since the more routers an optical signal passes, the more insertion loss it will suffer and the more crosstalk noise will be accumulated, which eventually leads to transmission error. Thus there is a limitation on the largest number of optical routers the optical signal can pass. Another challenge is high loss of waveguide crossing. It is inevitable to have waveguides crossings in the optical switch. The design constraint on the input power per waveguide limits the maximum number of wavelengths transmitted in the waveguide. The light passing more switches requires higher power, which results in less number of wavelength and network bandwidth. Koka et. al. showed that all-to-all network has better power and performance characteristics than switched network under the design constraints [33].

# 3.0 MINTRIM: TOLERATING PROCESS VARIATIONS IN NANOPHOTONIC ON-CHIP NETWORKS

I have reviewed basic knowledge of nanophotonics interconnects and recent work in previous chapters. In this chapter, I will present the work on robust and reliable on-chip optical network design. Section 3.1 introduces prior arts on reliability issues of optical network. In section 3.2, I will describe the proposed suite of solutions starting from improving the success rate of trimming while minimizing the static power, to ultimately provisioning near-full bandwidth for an optical network under PV. The PV modeling and experimental results are analyzed in Section 3.3 and Section 3.4, respectively. Section 3.5 summarizes this chapter.

## 3.1 BACKGROUND

The key elements in an optical network includes a laser source, which generates laser of different wavelengths; waveguides, which propagate laser signals across the chip; modulators, which imprint binary signals on laser of certain wavelengths, and detectors, which receive optical signals and convert them to electrical signals. The laser source is responsible for generating phase-coherence and equally spaced wavelengths. It is expected that such laser source could produce 64 or even more wavelengths per waveguide for a DWDM network [31, 74].

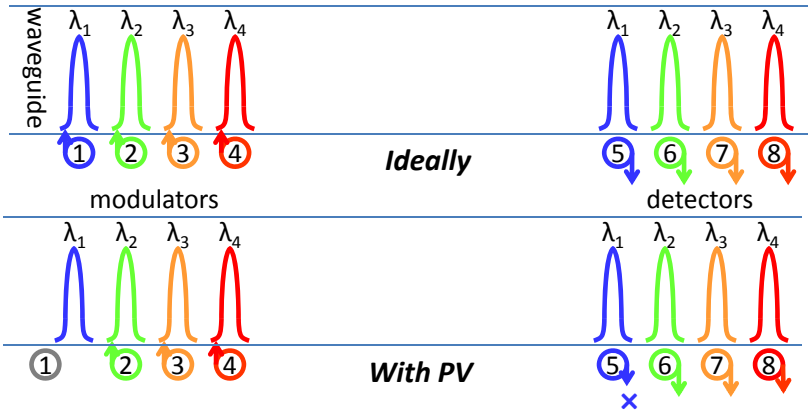



Figure 10: Bandwidth loss due to PV-drift.

### 3.1.1 A Motivating Example

If $\mu$rings are fabricated perfectly, a sender and a receiver can modulate and extract optical signals correctly without any loss. The upper part of Fig. 10 illustrates such an ideal scenario where the sender uses $\mu$rings #1 $\sim$ #4 to modulate their nominal four wavelengths $\lambda_1 \sim \lambda_4$, and the receiver uses $\mu$rings #5 $\sim$ #8 to detect and extract the same wavelengths respectively. Note that ring #5 and #1 have the same resonance, so do #6 and #2 etc. Under ideal situation, both sender and receiver can utilize 100% of their bandwidth for transmission. When PV is present, some $\mu$rings are off from their resonance due to imprecise dimension, e.g. waveguide width. Fig. 10 shows the same example with $\mu$ring #1 being off from $\lambda_1$. As a result, it cannot resonate at $\lambda_1$, downgrading the sender's bandwidth to 75%. Consequently, ring #5 at the receiver cannot receive any signal. Such a bandwidth loss is a static loss meaning that this sender loses 25% bandwidth permanently.

### 3.1.2 Current Approaches and Challenges

There are mainly two types of approaches to trimming the drifted resonant wavelength of $\mu$rings. The first one is power trimming. Heating and carrier injection can shift the resonant wavelength of a $\mu$ring up and down respectively [2]. In Figure 10, $\mu$ring #1 can be

corrected towards red using heating and shifted towards blue using current injection. This type of method can fine tune the resonance of $\mu$rings. However, there are three fundamental limitations to power trimming:

**Challenge 1: Power trimming incurs high static power consumption.** Many existing work have shown that the static power for trimming the $\mu$rings is a significant portion, or even dominant portion, of the total optical network power. For example, the Corona network in 17-nm technology from HP [69, 2] is estimated to consume $\sim$26W in power trimming, out of $\sim$48W of total network power. Even with the most optimistic $\mu$ring heating efficiency, e.g. using in-plane heaters and air-undercut [25, 26], it is estimated that $\mu$ring heating still consumes 38% of total network power [50]. For this reason, many work also focused on reducing the amount of $\mu$rings on-chip to reduce the power needed for trimming [50, 25].

**Challenge 2: Power trimming can only correct limited resonance drifts.** Even though the resonance wavelength can be corrected towards red or blue, blue shifts is still limited no matter how much power we are willing to pay. This is because blue shifts is achieved through carrier injection, which heats up the $\mu$rings and causes red shifts that need further carrier injection for correction, forming a positive feedback loop and thermal runaway [45]. In addition, more carrier injection degrades the extinction ratio and creates more power loss of the signal, e.g.$\sim$0.4nm tuning in wavelength results in 1dB signal loss [17, 39]. Hence, the achievable amount of blue shift is far less than of red shift [45]. For this reason, many work just use heating to keep all $\mu$rings at a constant temperature [25, 26, 48], which should be close to the peak temperature of the chip to avoid blue shifts.

The second class of trimming is done post-fabrication by changing its refractive index of the $\mu$ring directly, or adjusting the stress level of the cladding material. The advantage of such physical trimming is that, if successful, no additional power is required for correcting PV-drifts. However, the challenge is:

**Challenge 3: Physical trimming is immature and less commercially practical.** All physical trimmings require precise control of irradiation dose and energy, which is different from $\mu$ring to $\mu$ring. Given that there are thousands to millions of $\mu$rings on-chip, it is currently difficult to do physical trimming in mass fabrication which is critical for commercial purposes. Whereas, the power trimming saves *tuning effort* from that required for physical

trimming with the receive-data driven control circuit [19], which can tune the $\mu$rings without external intervention. Second, SOI has a key advantage over other core material: high refractive index contrast between silicon (core) and cladding, which enables small bend radii and dense integration. Hence, resonators built in non-silicon material are less attractive for future photonic networks. However, with SOI, trimming the cladding material ($SiO_2$) is unstable as a subsequent red shift of 0.15nm was observed 5 days after the irradiation. Moreover, the quality factor Q of the $\mu$ring decreased by 21$\sim$41.2% with a 1$\sim$2nm correction [58], which would increase the BER of the optical signal or require higher laser source power to overcome signal attenuation.

There are also proposals that do not rely on physical or power trimming to overcome PV. A dynamic regulation method was proposed [39] in which adjusting chip temperature is used to compensate chip-wise PV-drifts (i.e. systematic variations). For example, if the PV-drift of $\mu$rings in a chip region are toward blue, then the regulator would heat up, i.e., red shift, the region via e.g., dynamic voltage/frequency scaling (DVFS). Such coarse-grained regulation cannot overcome random PV-drifts, e.g., both red and blue drifts, among different $\mu$rings within the region. Also, DVFS comes at non-trivial performance cost, especially when cooling the chip region is required. Nitta *et al.* proposed to use error detection/correction code to tackle faulty $\mu$rings that are due to either PV-drifts, or temperature induced resonant wavelength drifts, or insufficient trimming [15]. However, such schemes can only handle small number of faulty $\mu$rings since the overhead of error correction coding, in both performance and extra optical bandwidth requirement, would be daunting otherwise. As we will show in our experiments, even conservative estimation of PV-drifts indicates that more than half the $\mu$rings could become faulty, which cannot be solved using coding mechanisms. A tuning control circuit that allows $\mu$ring to resonate at its closest wavelength instead of the original assigned one through bit re-shuffling was developed [19]. We adopt the same circuit design in the experiments and use their tuning strategy as one of the baselines to compare against ours.

Next I describe the proposed suite of solutions starting from improving the success rate of trimming while minimizing the static power, to ultimately provisioning near-full bandwidth for an optical network under PV.

## 3.2 PROCESS VARIATION TOLERANT METHOD

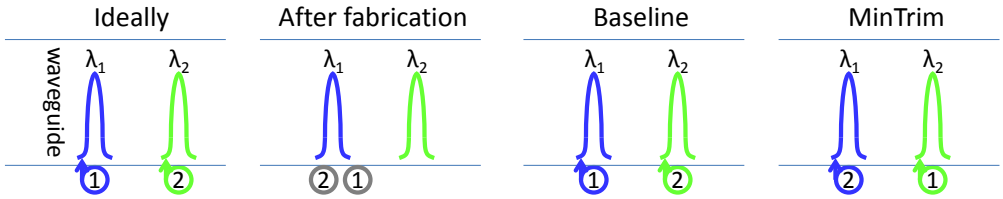
The first drawback of power trimming is high static power, since all $\mu$rings need to be kept at a constant temperature to be functional, which would require continuous heating power or current injection power (effective "cooling" through power) to cancel the effect of on-chip temperature fluctuation. With PV, a $\mu$ring may be off its nominal resonant wavelength, so *additional power trimming* is required to correct it back, on top of the power to keep it thermally stable, exacerbating the already high static power of the optical network. A $\mu$ring's resonance wavelength typically drifts by $\sim$0.1nm/°C [55, 54, 80]. Hence, an average of 1nm of PV-drift [58, 17, 75] would require equal amount of power for regulating the $\mu$ring temperature within 10°C fluctuation range. Hence, PV-drifts add significant power overhead to the network, which is what we will minimize in MinTrim.

Second, even with unlimited power supply, current injection can shift the resonant wavelength towards the blue end of the spectrum, but can also degrade trimming efficiency and even trigger thermal runaway [17, 45, 39]. Hence, it can only correct small PV-drifts, e.g. 0.4nm which also results in 1dB signal loss [39]. With PV, a $\mu$ring's resonant wavelength may be shifted towards red beyond the correctable range. This is the main reason for the network to lose bandwidth since such $\mu$rings and the corresponding nominal wavelengths cannot be used. As we will show later, our sample network architecture loses more than 40% bandwidth because 32% of the $\mu$rings are uncorrectable due to PV. MinTrim strives to turn uncorrrectable into correctable scenarios to achieve maximum bandwidth.

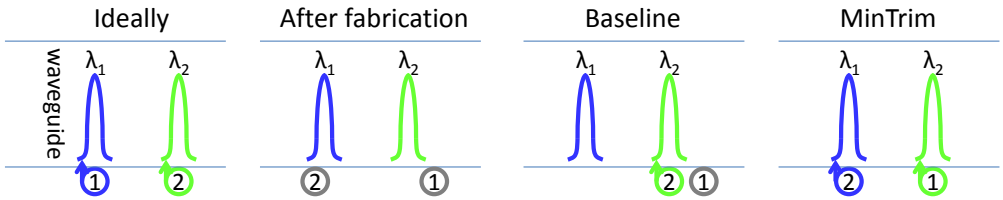
We discuss MinTrim using three types of wavelength-$\mu$ring organization of optical buses and crossbars, namely single-writer-multiple-reader (SWMR), multiple-writer-single-reader (MWSR), and multiple-writer-multiple-reader (MWMR) [29, 69, 51, 36, 76, 4]. In SWMR or MWSR, network nodes have exclusive sets of wavelengths for transmitting or receiving data. In these two architectures, modulators and detectors of each node use complementary sets of wavelengths. In MWMR, all modulators and detectors of a node use all wavelengths, increasing the network bandwidth over the other two. Both MWSR and MWMR require arbitration before sending data while SWMR does not. MinTrim is applicable to all these three architectures.

### 3.2.1 An Optimization Problem

The first step in MinTrim is developed based on the observation that a $\mu$ring does not have to be trimmed to its nominal wavelength as it may be far from the $\mu$ring's resonant wavelength. With PV, the distribution of the resonant wavelengths of $\mu$rings are somewhat random. Hence, as long as we can generate an association between $\mu$rings and wavelengths, such that the number of usable wavelengths for each node is maximized, then we can achieve the highest bandwidth. In order to keep the trimming power low, the most intuitive way is to trim a $\mu$ring to a *nearby* wavelength, rather than its nominal wavelength, to reduce the trimming distance which linearly affects the trimming power. More importantly, such nearby-mapping can reduce the number of uncorrectable $\mu$rings as their trimming distances are now smaller.




(a) Less trimming power for PV.




(b) Higher bandwidth.

Figure 11: Two advantages of trimming $\mu$rings to a nearby wavelength.

Figure 11 illustrates these two advantages with a simple example. Here the nominal wavelengths of $\mu$ring#1 and #2 are $\lambda_1$ and $\lambda_2$ respectively. In 11(a), suppose PV causes $\mu$ring#1 and #2 to be closer to $\lambda_2$ and $\lambda_1$ respectively. The baseline design trims the two $\mu$rings back to their nominal wavelengths. In MinTrim, $\mu$ring#1 will be trimmed to $\lambda_2$, and $\mu$ring#2 to $\lambda_1$, which clearly consumes less trimming power than in the baseline. In 11(b), suppose $\mu$ring#1's resonant wavelength is too far from $\lambda_1$ to be correctable using current

injection. The baseline would lose $\lambda_1$ as no $\mu$ring can resonate at it, but MinTrim would actually make $\mu$ring#1 correctable by trimming it to $\lambda_2$ since it is closer, and $\mu$ring#2 to $\lambda_1$, salvaging all available bandwidth.

However, if the resonant wavelength of a $\mu$ring is roughly in the middle of two channels, say $\lambda_i$ and $\lambda_{i+1}$, MinTrim needs to determine which wavelength should the $\mu$ring be trimmed to. The decision is based on which map would generate higher bandwidth and require lower trimming power. Since the decision for one $\mu$ring affects other $\mu$rings, MinTrim needs to generate a globally optimal solution, which can be solved by an optimization tool such as integer linear programming (ILP). ILP is a powerful method for optimizing a certain objective function through determining a set of decision variables, subject to some constraints. Note that MinTrim is a post-fabrication procedure to alleviate the PV-induced damage. No further reconfigurations are needed at runtime. Hence, running an optimization algorithm incurs only a one-time cost, and is worthwhile since it improves the yield of the chip effectively. We will now describe how to formulate MinTrim into an ILP problem by defining the decision variables, objective functions and constraints.

**3.2.1.1 Decision Variables** Since we are trying to decide which wavelength should a $\mu$ring be trimmed to, the decision variables of our problem are simply boolean variables, map($r_n$, $w_m$, *node*), representing whether $\mu$ring $r_n$ of a *node* should be trimmed to wavelength $w_m$, 1 being yes and 0 being no.

**3.2.1.2 Objective Function** MinTrim tries to achieve two objectives: maximal bandwidth and minimal trimming power. Given that ILP can only maximize (or minimize) one goal, we let maximal bandwidth take higher priority over minimal power, but the reverse can also be formulated under a different chip design goal. That is, if there are two solutions, one with higher bandwidth and the other with lower trimming power, MinTrim will return the former as the solution. To achieve this, we iteratively run ILP with a bandwidth in descending order starting from 100%. The granularity of descreasing bandwidth is losing one wavelenth for a node at a time. The algorithm terminates when a solution is found, i.e. the requested bandwidth is satisfied and the trimming power is the lowest within the

available solution pool. Trimming power is calculated as the following formula, where $\lambda_{act}[r_n]$ is a parameter of $r_n$ to represent the actual wavelength of $r_n$ post fabrication. The difference between actual and target wavelength, $w_m$, determines how much trimming power is required.

$$\sum_{\substack{\forall n,\forall m \\ \forall node}} map(.) \times \begin{cases} 0.13 \times (\lambda_{act}[r_n] - w_m) & \text{if} \lambda_{act}[r_n] \geq w_m, \\ 0.24 \times (w_m - \lambda_{act}[r_n]) & \text{if} \lambda_{act}[r_n] < w_m. \end{cases}$$

The coefficients, 0.13mW/nm and 0.24mW/nm, are unit power required for current injection and heating respectively [45]. We will use $map(.)$ rather than the full length of the $map$ function for brevity, since they are all in one form.

**3.2.1.3 Constraints** There are two constraints on trimming $\mu$rings to wavelengths. For every node using a waveguide, (1) every $\mu$ring of the node should resonate with at most one wavelength in the waveguide; and (2) every wavelength in the waveguide should be resonant with at most one $\mu$ring of the node:

$$\forall r_n,\ \forall node, \qquad \textstyle\sum_{w_m \in \{\text{all } \lambda\text{'s}\}} map(.) \leq 1 \tag{3.1}$$

$$\forall w_m,\ \forall node, \quad \textstyle\sum_{r_n \in \{\text{modulators in } node\}} map(.) \leq 1, \quad \sum_{r_n \in \{\text{detectors in } node\ \}} map(.) \leq 1 \tag{3.2}$$

To enforce that modulators and detectors of each node use complementary set of wavelengths in SWMR and MWSR, we have:

$$\forall node, \qquad \text{Let } S = \{\lambda\text{'s assigned to } node \text{ for modulation}\},$$

$$\forall w_m \notin S, \quad \textstyle\sum_{r_n \in \{\text{modulators in } node\}} map(.) = 0 \tag{3.3}$$

$$\forall w_m \in S, \quad \textstyle\sum_{r_n \in \{\text{detectors in } node\}} map(.) = 0 \tag{3.4}$$

Those are not needed for MWMR since it does not have this constraint.

Another set of important constraint is on the trimming distance. In the thesis, we assume 0.4nm as the constraint for current injection [39]. For trimming through heating, the constraint depends on the chip power budget since heating power increases linearly with trimming distance. A 2nm of wavelength shift requires the temperature of the $\mu$ring to be

20°C above the ambient temperature [45]. In addition, allowing a wide range of heating brings challenges to thermal insulation among the $\mu$rings. Therefore, in this section, we will put constraints on trimming distance through heating, termed "Rlimit", and show in the results the trend of trimming power and network bandwidth with varying allowable distance. Hence, the constraints for trimming distance are:

$$\forall n,\ \forall m,\ \forall node,$$
$$map(.) \times (\lambda_{act}[r_n] - w_m) \leq 0.4,\ \text{if } \lambda_{act}[r_n] \geq w_m,$$
$$map(.) \times (w_m - \lambda_{act}[r_n]) \leq \text{Rlimit},\ \text{otherwise.}$$

In addition, the constraint for bandwidth is:

$$\forall node,$$
$$\sum_{r_n \in \{\forall \mu\text{rings in } node\}, w_m \in \{\text{all } \lambda\text{'s}\}} map(.) \geq Bandwidth_{min}$$

where $\text{Bandwidth}_{min}$ is reduced incrementally, starting from 100%, during the interactive search procedure.

This first ILP step is able to dramatically improve the success rate of trimming $\mu$rings and the number of usable wavelengths. As will be shown later, the number of usable $\mu$rings improved from 68% in the baseline to 97%, resulting in a bandwidth increase from 59% to 81%. To salvage the remaining bandwidth loss, we now introduce the next step in MinTrim.

### 3.2.2 Supplementing $\mu$rings with Spares

The next simple method is to supplement the existing $\mu$rings with spares. Having more $\mu$rings creates more opportunities for selecting correctable $\mu$rings, as illustrated in Figure 12 where $\mu$ring#1 is supplemented with #2 which is closer to $\mu$ring#1's nominal wavelength $\lambda_1$, under PV. MinTrim will trim $\mu$ring#2 to $\lambda_1$. The rationale behind this idea is that when fabricating two $\mu$rings of the same nominal wavelength instead of one, there is always a better one for MinTrim to pick. The advantages are again two fold: (1) reduced trimming power with closer rings and (2) improved successful trimming; since the $\mu$ring with less

trimming distance will be selected. Incorporating spare $\mu$rings in ILP formulas is as simple as increasing the set of modulators and detectors in Equation (3.1)-(3.4), without any further changes.

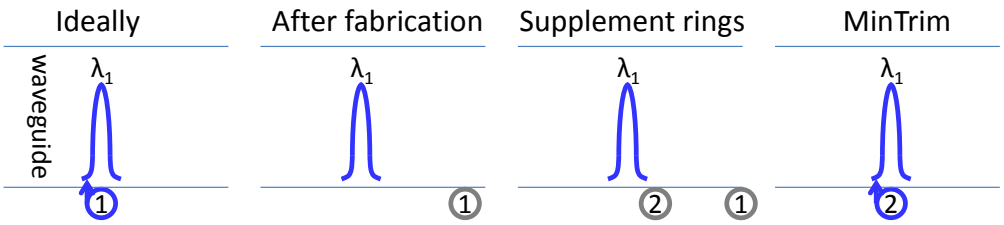



Figure 12: Supplementing $\mu$rings with spares.

The first question to address is how many spares to provide for a node with N $\mu$rings and resonant wavelength $\lambda_1, ..., \lambda_N$. We do not have to backup every $\mu$ring because many of them might already be good enough. Suppose we provide M supplemental $\mu$rings, M<N. Ideally, these M $\mu$rings should be the backups for those with large PVs. Unfortunately these are not known prior to fabrication and, hence, there are a number of alternatives for assigning nominal wavelengths to $\mu$rings. For instance, we can assign N $\mu$rings to $\lambda_1, ..., \lambda_N$ and assign the remaining M $\mu$rings to M wavelengths chosen uniformly among $\lambda_1, ..., \lambda_N$. However, this alternative is likely to benefit only two wavelengths closest to a spare's resonant wavelength. Hence, in our experiments, we will explore the following strategies:

1. The nominal wavelengths of all N+M $\mu$rings are uniformly distributed across the entire wavelength spectrum $\lambda_1 \sim \lambda_N$, to hopefully generate the best coverage. We term this strategy `Even` as shown Fig. 13(a).
2. Observing that it is more difficult for MinTrim to correct $\mu$rings on the two ends of the wavelength spectrum because they can only be trimmed in one direction while others can be trimmed towards either red or blue, it is also natural to supplement $\mu$rings on the two ends with more spares than in the middle. We term this strategy `Double_ends_even_middle`, or DEEM, meaning that we assign 2R spares with nominal wavelengths of $\lambda_1 \cdots \lambda_R$ and $\lambda_{N-R+1} \cdots \lambda_N$, and distribute the remaining M-2R $\mu$rings across the spectrum of $\lambda_{R+1} \sim \lambda_{N-R}$. Fig. 13(b) shows an example of DEEM.

3. If M = N, then two $\mu$rings can be assigned to each of $\lambda_1 \cdots \lambda_N$. We term this strategy `Double` as illustrated by Fig. 13(c).

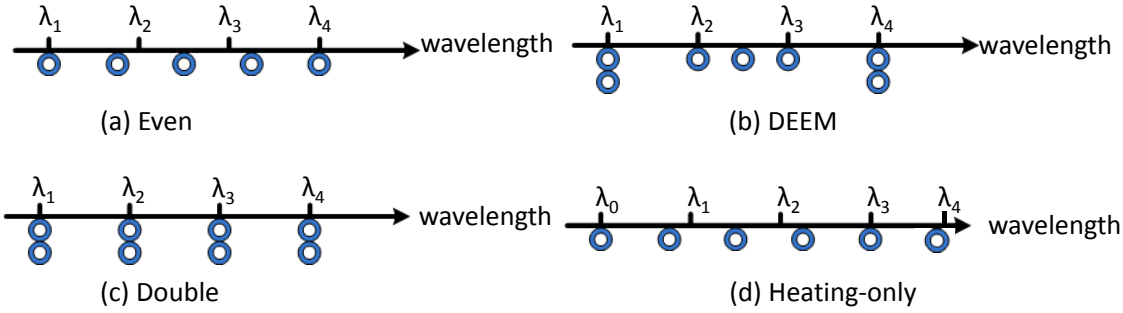



Figure 13: Different strategies for spare $\mu$rings placement.

In SWMR optical crossbar architecture, modulators have larger impact on network bandwidth than receivers. Because losing one modulator results in the bandwidth loss of all links connected to the local node. Whereas the failure of one receiver only causes the bandwidth degradation for one link between two nodes. In addition, the modulators is much less than the receivers at each node. Due to these reasons, adding redundant modulator is more cost-efficient than supplying spare receivers. Hence, when M > N, we implement a triple-sender-double-receiver strategy termed as `3S2R` that required 4 more modulators per node per waveguide over `Double`.

For the optical network adopting the heating-only trimming method [25, 26, 48], $\lambda$s can only be trimmed towards red. To improve the possibility of successful wavelength mapping with heating, we proposed to supply the extra rings at the left side of the spectrum in addition to the ones inside the spectrum. Fig. 13(d) shows that the $\mu$ring of resonant wavelength $\lambda_0$ that does not belong to the designated wavelength set is added to optical network. This strategy could handle the conditions when the red shifts caused by PV can not be corrected due to the limitation of trimming. Then the supplementary rings with smaller $\lambda$s, e.g. $\lambda_0$ can be trimmed to the ones inside the spectrum, e.g. $\lambda_1$ with heating. We assumed that K supplemental $\mu$rings' resonate wavelengths are outside the spectrum range, K<N. The experiment results in section 3.4 indicate that a small value of K can result in 20% of bandwidth improvement, but 20% more trimming power.

The second question relates to the possibility that adding more $\mu$rings may increase the power consumption of the network. If N out of N+M $\mu$rings are selected by MinTrim, the remaining M $\mu$rings may cause ∼1dB light loss each in the waveguide [26], especially when a $\mu$ring is close to a wavelength. For this reason, those M $\mu$rings should be tuned off, by bringing their resonance wavelengths to the closest mid-points between two channels. When a $\mu$ring is tuned off, it generates ∼1.5e-3dB light loss in the waveguide [45], which results in a total of 0.55% laser power loss. Since off-tuning is done through trimming, this amount of trimming power overhead is more of a concern. We measured through our experiments that the average trimming distance for this part is 0.205nm, about $\Delta\lambda/4$ since the trimming distance is within [0,$\Delta\lambda/2$]. In fact, our experimental results will show that more spare $\mu$rings lead to total trimming power (trimming N $\mu$rings + tuning off M $\mu$rings) reduction because the power required to trim one $\mu$ring by 2nm through heating is equivalent to the power for tuning off 9 unselected $\mu$rings. As a result, the `DEEM` strategy of sparing results in the best bandwidth, ∼90%, with the lowest power requirement, as will be shown in Section 3.4. Last, the spare $\mu$rings do not increase the die area since the waveguides extend across the entire die and there is plenty of space between $\mu$rings to accommodate spares.

### 3.2.3 Flexible Wavelength Assignment for Network Nodes

To recover the remaining bandwidth, we develop the third step of MinTrim. Observe that in both SWMR and MWSR, a node (either a modulator in SWMR or a detector in MWSR) does not use all wavelengths in a waveguide to transmit or receive data. Each node is assigned N/X wavelengths for transmission (SWMR) or receiving (MWSR), where N is the total number of wavelengths in a waveguide shared by X network nodes. In this paper, we set N to be 64 and X to be 16. With perfect fabrication process, i.e. no PV, it does not matter which N/X wavelengths are assigned to each node. With PV, however, determining which N/X wavelengths are assigned to a node is crucial since a wavelength may not be usable by one node but usable by another. Hence, a node should be assigned with those N/X wavelengths that are usable by this node. We term this technique flexible_wavelength_to_node_assignment, or `Flexible_assignment`.

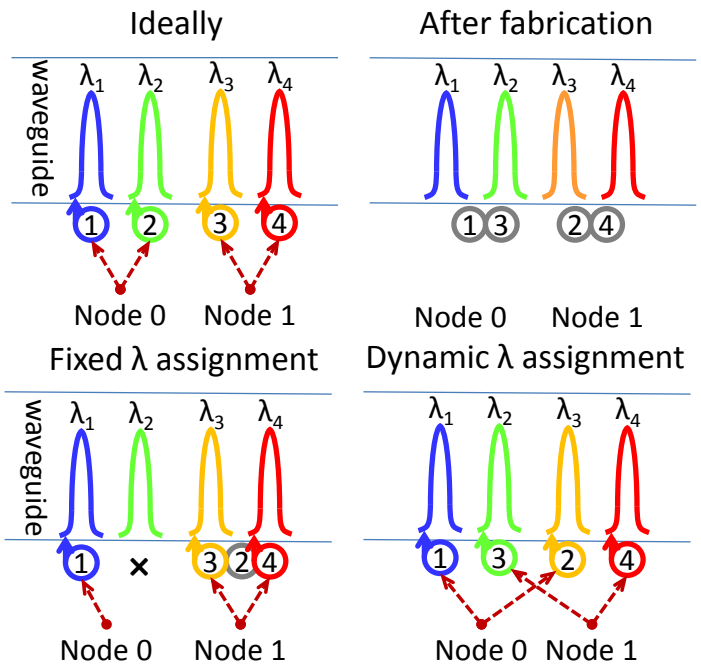



Figure 14: A case for flexible assignment between wavelengths and nodes.

Fig. 14 explains why having a flexible assignment is effective in bandwidth recovery. $\text{Node}_0$ has two $\mu$rings: #1 and #2, and $\text{Node}_1$ has #3 and #4. Due to PV, the resonance wavelengths of the four $\mu$rings are drifted as shown in "After fabrication". When original wavelength-node assignment is used ("Fixed $\lambda$ assignment"), ILP will search *within the local pool of $\mu$rings* to find ones that can resonate at $\lambda_1$ and $\lambda_2$. Since $\mu$ring #2 is drifted beyond correctable range of current injection, $\lambda_2$ becomes unusable. However, note that the optimum assignment between these wavelengths and nodes is: ($\mu$ring) #1$\rightarrow \lambda_1$, #3$\rightarrow \lambda_2$, #2$\rightarrow \lambda_3$, #4$\rightarrow \lambda_4$. With a fixed wavelength-node assignment, $\text{Node}_0$ cannot use $\lambda_2$ because $\mu$ring #3 is physically local to $\text{Node}_1$. However, $\mu$ring #2 is physically local to $\text{Node}_0$, and can resonate at $\lambda_3$, $\text{Node}_0$ can hence use $\lambda_1$ and $\lambda_3$, $\text{Node}_1$ can use $\lambda_2$ and $\lambda_4$, as shown in the figure.

To achieve flexible assignment between wavelengths and nodes, we extend the ILP formulation with new constraints. First of all, while Equation (3.1) and (3.2) still hold, Equation (3.3) and (3.4) cannot be used since the set of modulating wavelengths of each node is

no longer pre-defined. MinTrim needs to search for such set for each node. A new constraint we establish is that a wavelength can be assigned by at most one node:

$$\forall w_m, \quad \sum_{\forall node} \sum_{\substack{r_n \in \{\text{modulators (SWMR) } in \text{ } node\} \text{ or} \\ r_n \in \{\text{detectors (MWSR) } in \text{ } node\}}} map(.) \leq 1 \tag{3.5}$$

For detectors in SWMR, their resonant wavelengths should be the union of all modulating wavelengths of all other nodes. The same principle applies to modulators in MWSR.

$$\begin{gathered} \text{Let} R = \{\forall \text{dectecotrs (SWMR)}\} \text{ or } \{\forall \text{ modulators (MWSR)}\} \\ \forall w_m, \ \forall node, \\ \sum_{r_n \in R} map(.) \leq \sum_{r_n \neq node} \sum_{r_n \in \bar{R}} map(.) \end{gathered} \tag{3.6}$$

Finally, since we have spare $\mu$rings which can also be applied with flexible assignment, we define the following constraint to avoid having too many modulators or detectors per node.

$$\forall node, \ \sum_{r_n \in \bar{R}} \sum_{\forall w_m} map(.) \leq N/X \tag{3.7}$$

As I will show in the results, flexible assignment can recover almost all the remaining lost bandwidth. Lastly, MWMR does not need this step, so only the first two steps (ILP with spares) will be sufficient. This is because both modulators and detectors already have the full bandwidth spectrum to resonate. There is no need to reassign wavelengths among nodes since every node already has all available wavelengths.

### 3.2.4 Wrap Around Scheme

Given that the resonance of $\mu$ring repeats in each free spectral ranges (FSR), the separations between peaks of wavelength transmissivity, prior works exploited the ring resonance repetition by wrapping around the next resonance for rings [10, 19]. Fig. 15 shows an example that applying wavelength warp around scheme to improve the successful rate of trimming. After fabrication, the resonances of ring #1 $\sim$ #4 all shift toward red due to systematic variation. The resonance of ring #4 in next FSR is drawn with the dash circle, which is close to $\lambda_1$—the first wavelength channel inside the spectrum. Ring #4 is trimmed to $\lambda_1$ instead of $\lambda_4$ to meet trimming constraint. Then ring #1$\sim$ #3 are all shifted by one channel. Through exploring the mapping opportunities with resonance repetition, closer resonant wavelengths could be found to improve bandwidth and reduce trimming distance.

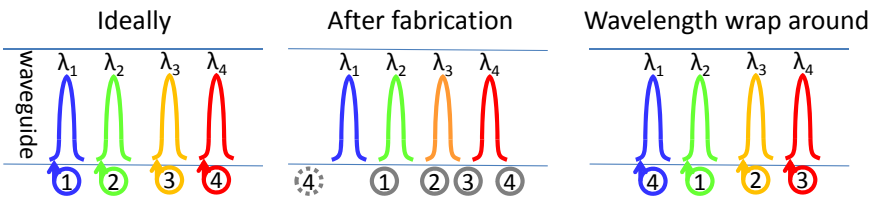



Figure 15: A case for wavelength wrap around. Extra resonance of $\mu$ring #4 is depicted as dash circle.

However, wrap around approach has some limitations on network designs. For example, it requires that the range of wavelength spectrum covered by the modulators should be close to the size of FSR. Otherwise, the resonance in the neighboring FSR would be too far to reach. While the network node usually only needs a subset of wavelengths to send one flit, then the wrap around scheme is useless for such design. In addition, FSR of each ring determined by the dimensional size is different from each other, which makes the tuning control logics even more complicated.

## 3.3 MODELING PV OF $\mu$RINGS

To evaluate the effectiveness of the proposed MinTrim, we first need to model an optical network subject to PV. The resonant wavelength of a $\mu$ring is determined by several factors including material used for waveguide and cladding, waveguide cross-section dimensions, circumference of the $\mu$ring, temperature etc. [59, 58]. Especially, for a fixed material and constant temperature, the wavelength is sensitive to the width and height of the waveguide. The variations of the wavelength is approximately linear to the width variation and height variation in waveguide. For example, 1nm of variation in width and height leads to 0.58$\sim$1nm [59, 39, 70] and $\approx$2nm [59] shift in resonance wavelength of the $\mu$ring respectively. Due to fabrication imperfection, the variations of critical physical dimensions, such as width, height or the thickness of silicon are inevitable. Hence, to characterize the PV of resonance wavelength of $\mu$rings, we will develop a variation model for the physical dimensions of the optical waveguide. Recent laboratory fabrications of optical devices show that physical dimensions variations can be classified into die-to-die, or D2D, (a.k.a. intra-die) and within die, or WID (a.k.a. inter-die) variations [59, 48]. The D2D variation refers to non-uniformity of devices between dies that are on the same or different wafers. This is generally caused by the fabrication tool and process design. The WID variation refers to such non-uniformity between identical devices within a single die. This is generally caused by die-level processes such as lithography and dry etch [27, 52]. Within a die, each step of the process may create spatial (systematic) and random variations in physical dimensions of the waveguide. Since the characteristics of the variations in optical devices are close to process variations in CMOS devices [13, 43] which also present D2D, WID including systematic and random variations among transistors, we adopt VARIUS [56], a PV modeling infrastructure for CMOS technology, based on the statistic tool `R` and its package `geoR` to model both WID and D2D variations.

VARIUS uses Normal (Gaussian) distribution to characterize on-chip process variations. The key parameters are mean ($\mu$), variance ($\sigma^2$), and density ($\phi$) of a variable that follows Normal distribution. Since wavelength variations are approximately linear to dimension variations of waveguide, we assume they follow the same distribution. The mean ($\mu$) of

wavelength variation of a $\mu$ring is its nominal wavelength. We use a spectrum of 64 wavelengths in a WDM network starting at 1550nm [48] and a channel spacing of 0.8nm. Hence, those wavelengths are the means for each $\mu$ring modeled.

Table 3: Two sets of PV parameters. WID variation=$\sqrt{\text{systematic var.}^2 + \text{random var.}^2}$ [56].

| | WID Variation (nm) | | D2D Variation (nm) | |
|---|---|---|---|---|
| | small die | large die | small die | large die |
| $PV_1$ | 0.57 [59] | 0.61 | 1.08 [59] | 1.01 |
| $PV_2$ | 0.37 [48] | 0.39 | 1.6 [48] | 1.40 |

The variance ($\sigma^2$) of wavelength variation is determined based on laboratory fabrication data [59, 48] and our target die size. Since optics are more cost-effective for many-core CMPs, we choose to model a 64-core chip with die size $400mm^2$ [29, 69]. There are no readily available variation data for such a die size, and measurements for small die sizes cannot be directly used because variations in small region is different from those in a large region. In [59], the standard deviation, $\sigma$, is 0.15nm for two $\mu$rings that are only 25$\mu$m apart, and 0.55nm if they are 1.7mm apart. The former characterizes the random variations within a die, and the latter describes systematic variations for a small die, e.g. $2\times2mm^2$. The D2D die variation in a $200mm$ wafer is also reported to be 1.08nm. To derive corresponding parameters for a $400mm^2$ die, we first generated 3K dies of $2\times2mm^2$ using the above variation parameters: $\sigma_{D2D} = 1.08nm$, $\sigma_{WID-systematic} = 0.55nm$, $\sigma_{WID-random} = 0.15nm$. Then we sort the dies according to their resulted mean values, and selected 100 (400/4) dies with close mean values to assemble a large die. This is because previous experiments demonstrated strong within-die spatial correlations of dimension variations [59, 48]. Hence, the 100 small dies that are next to each other should be strongly correlated as well. From the assembled large die, we then derive the WID and D2D variations that are used in our experiments. Finally, the density $\phi$ is a parameter that determines the range of WID spatial correlation. It is expressed as a fraction of chip's length in one dimension in VARIUS. As the spatial correlation of two devices decreases as their distance grows, $\phi$ is the distance at which the correlation drops to zero. Typical value for $\phi$ is 0.5/1.0 and for a large/small die.

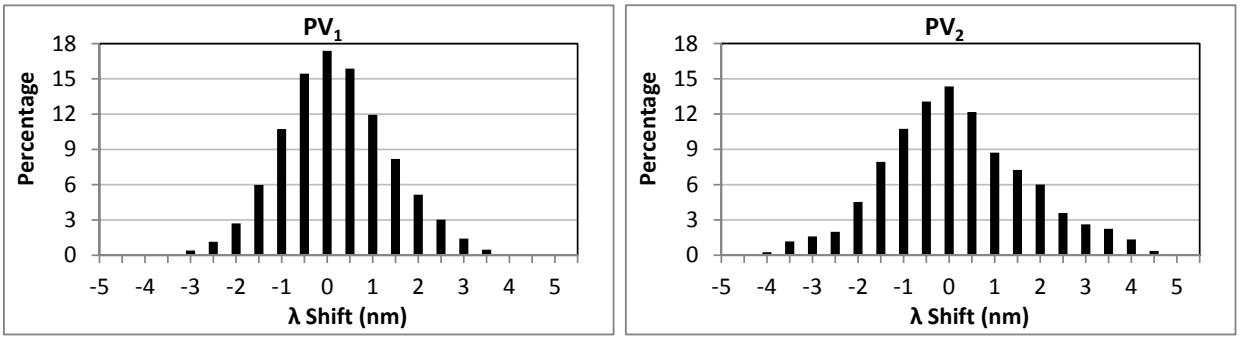



Figure 16: Distribution of wavelength shift for two sets of PV parameters in Table 3.

We generated two sets of variation parameters based on two different fabrications results [59, 48], using the same methodology since both of them use small dies ($2\times2.2mm^2$) in [48]. Table 3 compares the published results and our derivation for larger die sizes. As we can see, when dies size is larger, WID variation increases since some portion of D2D variation is now WID. Consequently, D2D variation decreases a little since it loses a portion to WID. We input these two sets of parameters into VARIUS and generated 100 sample dies of $400mm^2$ each. Each sample contains over one million points indicating the wavelengths of $\mu$rings. We then extracted those along the optical waveguide according to the physical layout of an optical crossbar [4]. The total number of points picked from the samples are equal to the number of $\mu$rings. Fig. 16 shows the distribution of wavelength shifts under $PV_1$ and $PV_2$. As we can see, the total effective variance, including both WID and D2D, of $PV_2$ is larger than of $PV_1$, so the bell shaped distribution is wider than for $PV_1$, meaning that more shift is present on-die which creates more bandwidth loss.

### 3.4 EVALUATIONS AND RESULTS

We use an SWMR crossbar, shown in Figure 17, as an example to demonstrate the effectiveness of MinTrim, although it is applicable to MWSR and MWMR as elaborated in Section 3.2. Our optical network is composed of 4 identical waveguides, each supporting 64

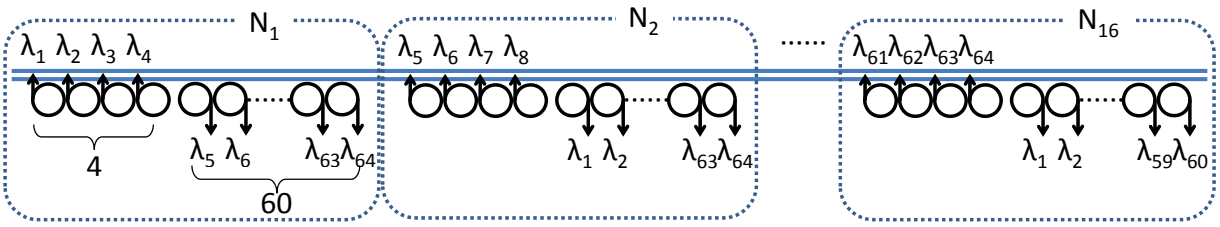



Figure 17: An SWMR network architecture used for evaluating MinTrim.

wavelengths denoted by $\lambda_1, ..., \lambda_{64}$. Each waveguide is shared by 16 network nodes. Since this is a single writer architecture, each node is exclusively assigned 4 $\lambda$'s for transmission. Hence no contention can occur during a write. Four $\mu$rings are used as modulators to resonate with these 4 $\lambda$'s. Every node can simultaneously read from all other 15 nodes, hence "multiple readers", requiring a total of 60 $\lambda$'s for reception, and 60 $\mu$rings detectors as shown in the figure.

The physical layout of the crossbar employed here is a symmetric design as each waveguide has exactly the same placement of $\mu$rings next to it. There are asymmetric designs such letting a subset of network nodes share one waveguide. For example, each node sends data via 16 wavelengths traversed in one specific waveguide instead of 4 wavelengths per waveguide and the light transmitted in each waveguide can be modulated by $\mu$rings connected to different set of nodes. However in such case, the bandwidth loss might be less than the symmetric scenario because it is less likely to have 16 failed rings than 4 ones. In addition to low design complexity, the reason that we select the symmetric layout is to show that the proposed solutions are able to recover most of network bandwidth even the design is vulnerable to PV. Furthermore, the first two steps of MinTrim: ILP and sparing can still be applied to other configurations. The third step, flexible wavelength mapping requires some minor modification, such as letting four nodes share 64 wavelengths in the waveguide instead of 16 nodes. However, if the connection of each node is separated by using different bundles of waveguides to obtain high transmission bandwidth, only ILP and spare can be applied to the optical network. Later we will show that previous two schemes dominate

the contribution of bandwidth improvement, so we expect that the final bandwidth to still approach 98% because our baseline configuration has more bandwidth loss than the original settings.

The variations of all $\mu$rings are generated as described in the previous section. Results are averaged over 100 sample dies. MinTrim computes solutions using the state-of-the-art ILP solver `lpsolve` [7]. The constraints and objective functions in the ILP problem are formulated using the front-end AMPL language [18].

We use *total network bandwidth* as a metric to evaluate MinTrim under different settings. The total network bandwidth is defined as the number of working channels (pair-wise tuned senders and receivers), summed over all possible sender-receiver pairs of the network. This is important because under PV, a sender and a receiver must have the same $\lambda$'s to communicate. Hence, only the common $\lambda$'s between the two nodes are counted towards effective bandwidth. As we can see, to have high total bandwidth, each node must be able to use as many $\lambda$'s as possible. Total network bandwidth of a perfect network without PV is 100%, and MinTrim strives to approach that.

In addition, we measure the power consumption of the network since another major advantage of MinTrim is power reduction. The power trimming techniques we employ requires 0.13mw/nm for current injection [2] and 0.24mw/nm for heating [45]. We assume current injection can correct up to $0.5\Delta\lambda$ towards blue [39] for power trimming. For the design just use heating to keep all $\mu$rings at a constant temperature [25, 26, 48], no blue shifts are allowed. For Rlimit (or Rlimit for short), we assume that the chip has certain power budget that limits this amount and we gradually relax such constraint to see if, using MinTrim, a large power budget is necessary to achieve high network bandwidth. Power measurement includes both the trimming power used to correct $\mu$ring's $\lambda$'s and the power required to tune-off unused $\mu$rings, i.e., power overhead.

### 3.4.1 Baseline Bandwidth Results

With PV, large amount of $\mu$rings are off from their nominal resonance $\lambda$, leading to a significant bandwidth loss. Assume an optimistic error tolerance of 10% $\Delta\lambda$, i.e., if the

actual $\lambda$ of a $\mu$ring is within 10% of the nominal $\lambda$, the $\mu$ring can still work. If no trimming is applied, the average total bandwidth is only 0.6% for both $PV_1$ and $PV_2$. In other words, the network does not work at all. Hence, we adopt power trimming in our baseline, and first compare two different ways of such trimming: (1) trim the $\mu$ring to the *closest* $\lambda$; and (2) trim the $\mu$ring to its nominal $\lambda$, both under trimming distance constraints. Note that (1) is different from trimming a $\mu$ring to a nearby $\lambda$, as is done in MinTrim, because a nearby $\mu$ring may not be the closest one, and searching for a good nearby $\lambda$ requires global optimization. Trimming to the closest $\lambda$ minimizes trimming power, but does not optimize bandwidth.

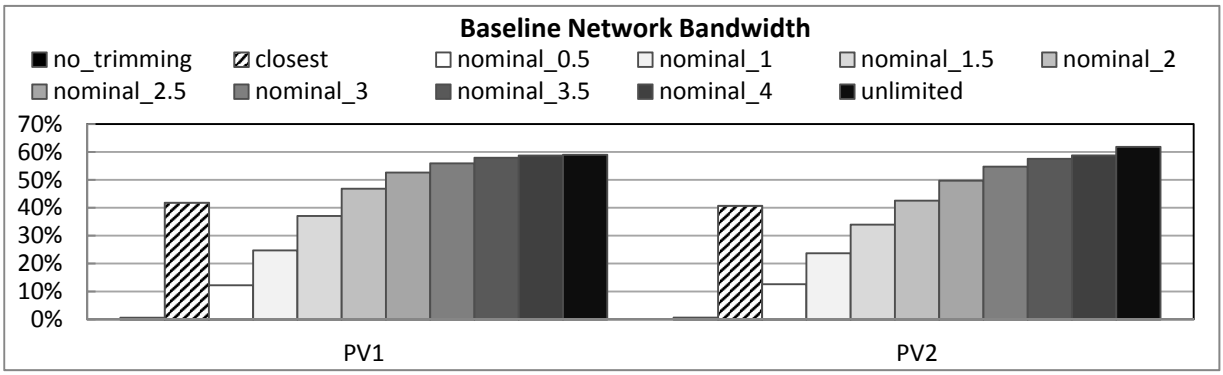



Figure 18: Average baseline network bandwidth comparison. Numbers following `nominal_` are Rlimit in unit of $\Delta\lambda$.

The bandwidths after these trimmings are shown in Fig. 18. As we can see, the "closest" bars can recover bandwidth from 0.6% in "no_trimming" to ~42%. The advantages of "closest" is that it does not require large trimming distance, and has the lowest trimming power as will be shown in Figure 37. It loses bandwidth when (1) more than one $\mu$rings are trimmed to the same $\lambda$, so one has to be removed and no spare is available for making this up; and (2) a sender and a receiver's $\mu$rings are trimmed to different $\lambda$'s, and only the common $\lambda$'s can be used for communication. Those two cases can be avoided by trimming the $\mu$rings to their nominal $\lambda$'s, labeled as "nominal_Rlimit". However, the "nominal"s also have limited capability in bandwidth recovery under tight heating power budget, e.g. below $2\Delta\lambda$. Progressively better bandwidth can be achieved when we relax Rlimit: 59%~62% for $PV_1$ and $PV_2$ respectively with unlimited Rlimit. We will use both "closest" and "nomi-

nal_"s in our later results. Also, although $PV_1$ and $PV_2$ show noticeable $\lambda$ shift distribution (Figure 16), the resulting baseline bandwidths differ only slightly. We will show results for $PV_1$ in the following discussion for clarity.

### 3.4.2 MinTrim Bandwidth Results

**3.4.2.1 First step: ILP.** When ILP is applied, great bandwidth improvement can be achieved immediately, as shown in Figure 19. The error bars show the minimum and maximum results from the 100 samples we experimented with. ILP achieves a bandwidth of 74% and 81% when Rlimit is $2\Delta\lambda$ and unlimited respectively. The reason of this improvement was illustrated in Figure 11: ILP can reduce the uncorrectable $\mu$rings by finding a good nearby $\lambda$. However, the improvement diminishes with a larger power budget. This problem can be addressed by having spare $\mu$rings as shown below.

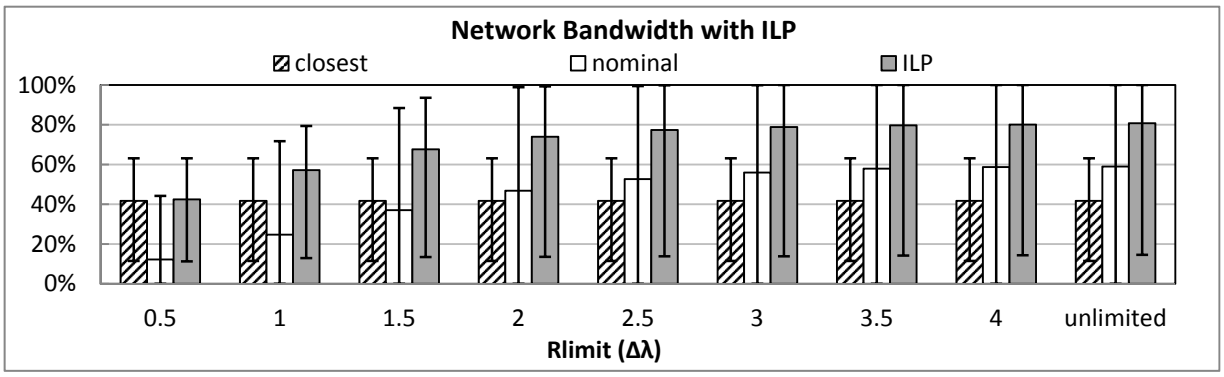
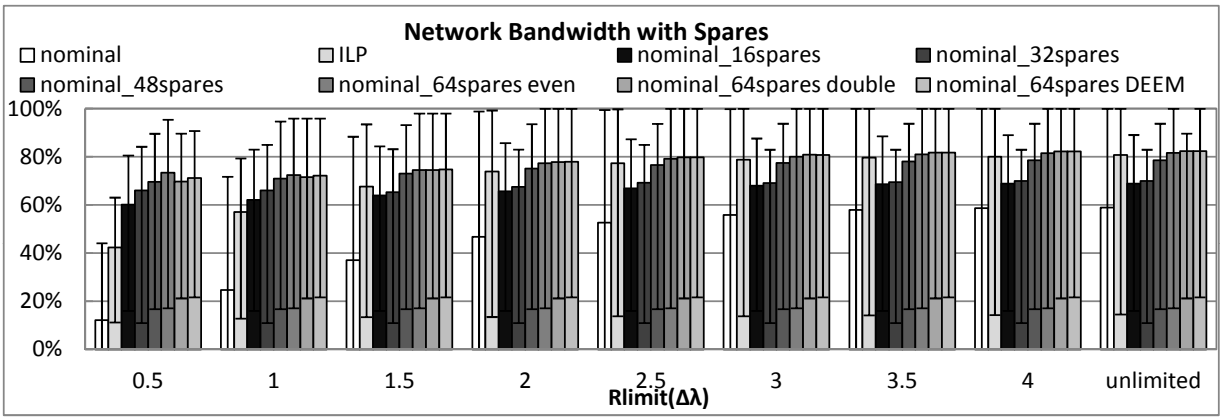



Figure 19: Bandwidth comparison among "closest", "nominal", and ILP-only.

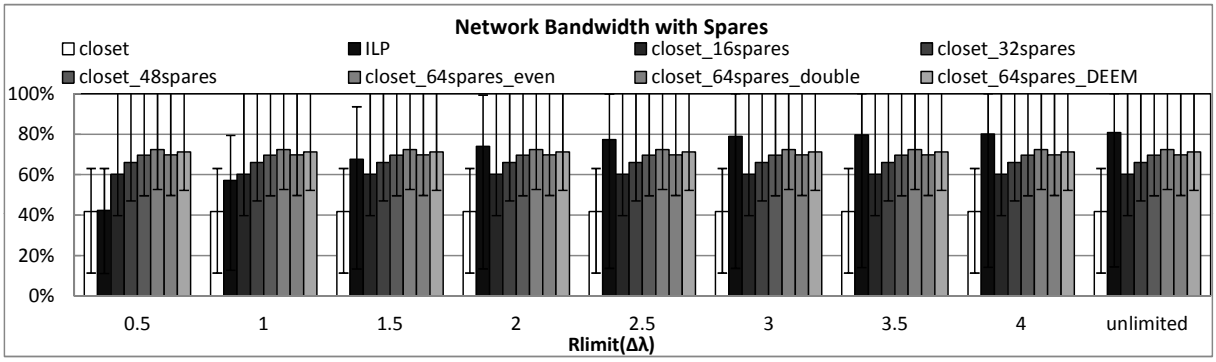
**3.4.2.2 Second step: Using spare $\mu$rings.** Each node in baseline and ILP only has 64 $\mu$rings. We now show the bandwidth results with different number of spare $\mu$rings, 16, 32, 48, and 64, on top of the original 64. When the number of spares is less than 64, we use the `Even` distribution as introduced in Section 3.2.2. When there are 64 spares, we applied all three distribution methods: `Even`, `Double` and `DEEM`. Recall that each node originally has 4 modulators and 60 detectors. With `DEEM`, we `Double` the 4 modulators, and 8 detectors (4 on each end of spectrum), and use `Even` for the remaining 104 filters. We treat modulators

and detectors separately because they are built differently. As we can see from Fig. 20, having spares effectively recovers more bandwidth than using "nominal", "closest" or ILP alone. More spares result in more improvement. From nominal to having 64 spares using `Even`, the bandwidth improvements are 500%∼38% when Rlimit increases from $0.5\Delta\lambda$ to unlimited. The `Double` method is more effective than `Even` because doubling $\mu$rings at their nominal $\lambda$'s have higher chances of getting a working $\mu$ring, as indicated by the $\lambda$ shift distribution in Fig. 16. Whereas, in `Even`, the nominal $\lambda$'s are not the 64 channels in the waveguide. Finally, the `DEEM` method stands out as the best one because the $\mu$rings on the ends of a spectrum are more difficult than in the middle. So doubling those $\mu$rings while using `Even` for middle $\mu$rings, given the same number of spares as in `Double`, achieves the best tradeoff. The bandwidth of `DEEM` reaches 73%∼82% when Rlimit increases from $0.5\Delta\lambda$ to unlimited based on "nominal" mapping. Figure 20(a) also shows that without additional $\mu$rings, ILP is able to recover similar amount of bandwidth as the approach of spare rings under loose trimming constraint(Rlimit $\geq 2.5\Delta\lambda$). ILP is even more effective than spare ring scheme when "closest" is adopted and the trimming constraint is beyond $1/5\Delta\lambda$), shown in Figure 20(b).

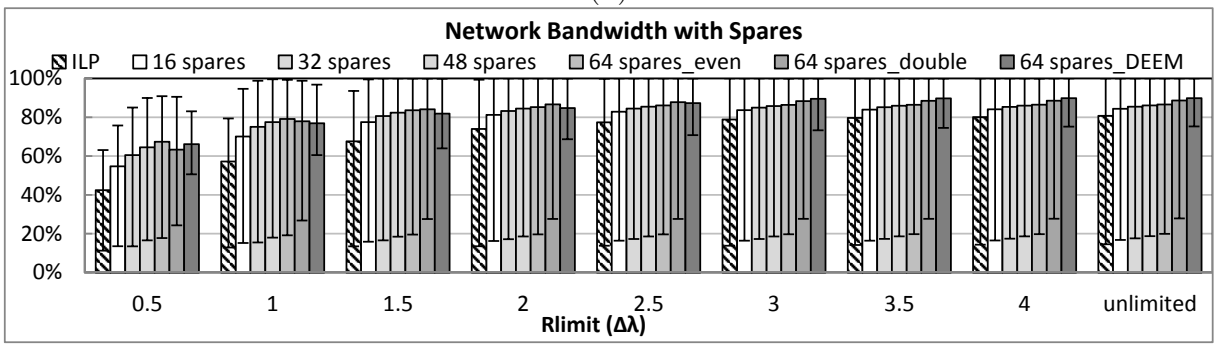
**3.4.2.3 Third step: Flexible $\lambda$ assignment to nodes.** To evaluate the effectiveness of the proposed third scheme, we applied the flexible assignment between $\lambda$'s and network nodes to "nominal" and compare it with the other two schemes: ILP and "spare"—64 $\mu$spare rings with `DEEM` in Fig. 21. Not surprisingly, providing flexibility in $\lambda$-node assignment generates more bandwidth than the baseline. As shown in Figure 21, all the three schemes could improve the network bandwidth significantly, while having spares performs slightly better than ILP and flexible mapping. In addition, applying 64 spare $\mu$rings to ILP can achieve 18%∼35% more bandwidth than using ILP alone. Flexible mapping could recover more bandwidth than ILP alone when the trimming constraint is tight, while spare ring scheme performs best with small heating range. Adding flexible mapping on top of ILP with spares increases bandwidth by 8%∼12%. When Rlimit is $2.5\Delta\lambda$, the bandwidth is 98.2%, close to 98.4% at unlimited Rlimit. Hence, with flexible assignment, having a power budget corresponding to Rlimit=$2.5\Delta\lambda$ is sufficiently good. More interestingly, the flexible




(a)




(b)




(c)

Figure 20: Bandwidth comparison among "nominal", ILP and varying amount of sparing in addition to (a) "nominal", (b) "closest" and (c) ILP.

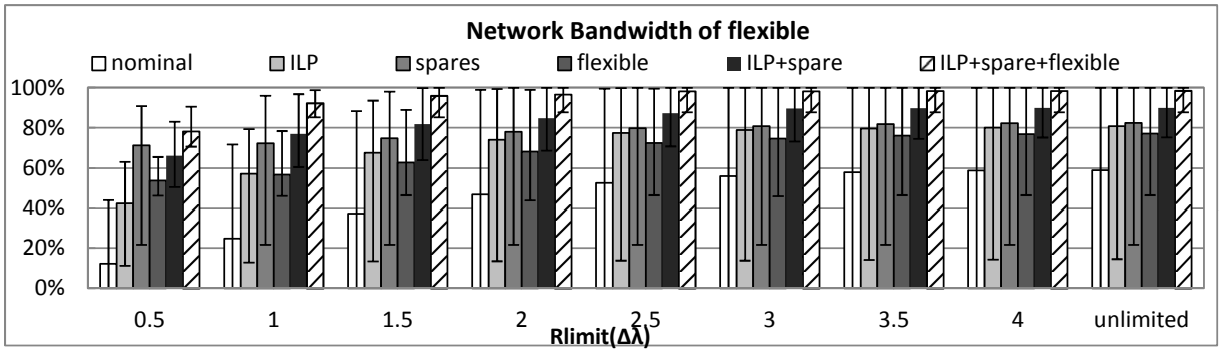



Figure 21: Bandwidth comparison between fix and flexible wavelength assignment.

assignment scheme results in much smaller error range (from 100 samples we generated), meaning that by allowing the nodes to select most suitable $\lambda$'s, the success rate of finding an assignment with high bandwidth is increased. This indicates that MinTrim provides a robust method to salvage network bandwidth under PV.

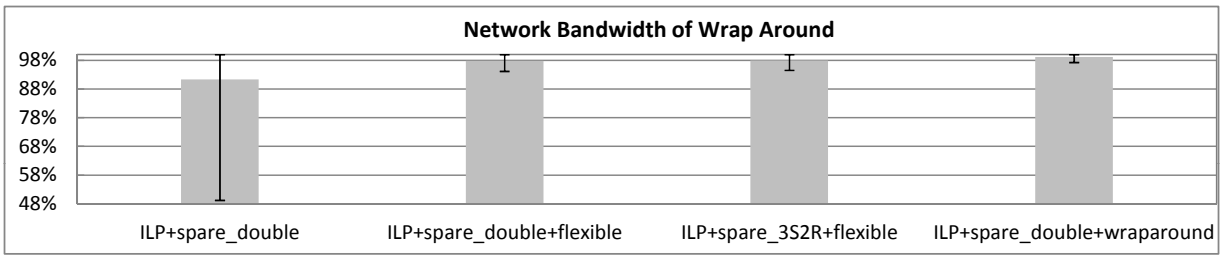
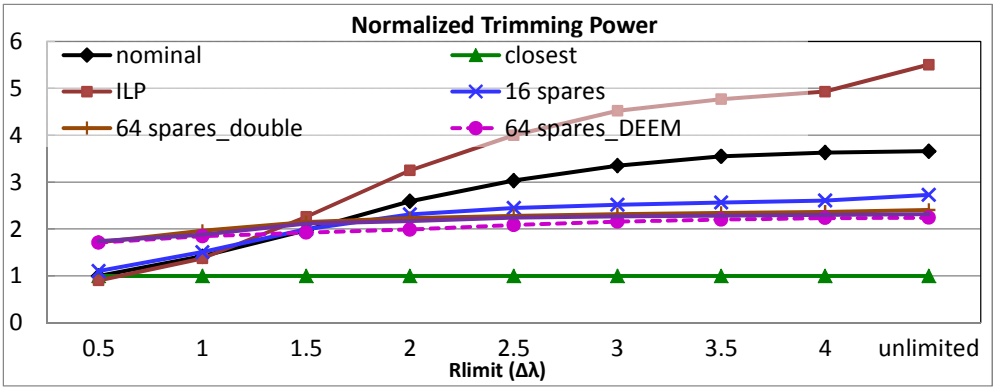



Figure 22: Bandwidth achieved with wrap around scheme, Rlimit is $2\Delta\lambda$.

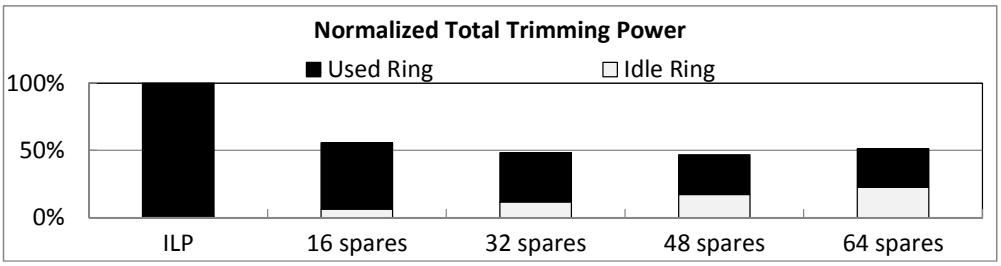
**3.4.2.4 Compared to Wrap Around Scheme** To compare MinTrim with the wrap around scheme proposed by prior arts [19], we need to implement it in addition to our flexible wavelength allocation approach. Because in baseline design, each node only uses a fixed subset of 64 wavelength channels, which does not meet the requirement of wrap around. Whereas MinTrim allows the node to use the non-nominal wavelengths and make wrap around feasible. Figure 22 compares the network bandwidth results among variety

of designs and shows that the bandwidth improvement of wrap around is only 0.2% over MinTrim on average. The reason is that wrap around mainly addresses the systematic variations, whereas ILP and flexible wavelength mapping in MinTrim can also mitigate this type of PV. The bandwidth achieved by the strategy `3S2R` does not contribute more bandwidth (only 0.1%) improvement since the flexible $\lambda$ assignment has already been able to improve the wavelength matching rate of modulators and no extra spares is necessary.

### 3.4.3 MinTrim Power Consumption Results




(a) Trimming power normalized to baseline-nominal.




(b) Power breakdown with Rlimit=3$\Delta\lambda$.

Figure 23: Power analysis of different MinTrim schemes.

The other major advantage of MinTrim is the trimming power reduction due to decreased total trimming distance. Fig. 23(a) shows the power comparison among different schemes normalized to baseline schemes "nominal" at Rlimit=0.5$\Delta\lambda$. For clarity, we do not show all sparing settings because their results overlap heavily in the figure. As we can see, baseline-closest requires lowest power among all schemes, but it can only achieve 41.8% of total

bandwidth. MinTrim-ILP at $0.5\Delta\lambda$ consumes even lower power (10% lower) than in baseline while achieving similar bandwidth (42.4%). However, ILP consumes the highest power, and baseline-nominal is the $2^{nd}$ highest among all when Rlimit increases because they both can trim more $\mu$rings at further distances but ILP has higher priority in bandwidth so it trims more $\mu$rings at further distances than in baseline. Once we add spare $\mu$rings, e.g. starting at 16 spares, the power consumption immediately drops at all Rlimits beyond $1.5\Delta\lambda$. This is because solutions can be found with closer $\mu$rings that help to decrease trimming distance. However, before $1.5\Delta\lambda$, higher power is consumed because again, higher bandwidth is more important. So MinTrim halts when there is a solution for high bandwidth, even when the power is higher. Overall, having more than 32 spares consumes about the same power, with 48-spares being the lowest. For example, A 37%/39% power reduction is observed for using 48 spares, compared with "nominal" when Rlimit is $3\Delta\lambda$/unlimited. `Double`, `DEEM` and flexible assignment do not differ significantly. The conclusion from these results is that having spares is effective in lowering power and improving bandwidth.

Fig. 23(b) shows the power breakdown for MinTrim, between trimming useful $\mu$rings (used ring) and tuning-off unused $\mu$rings (idle ring), with different number of spares from 0 to 64. The results are normalized to total trimming power of ILP, i.e. with 0 spares. The trend clearly shows that although adding spares increases the power for off-tuning unused $\mu$rings, the amount of active power for trimming useful $\mu$rings is greatly reduced, resulting in a large total reduction. Also, having 64 spares is sufficient because having more spares would slowly increase the total power because the useful power is stabilizing while the off-tuning power increases steadily.

### 3.4.4 MinTrim Quality Assessment through Network Connectivity Evaluation

As discussed earlier, MinTrim with flexible $\lambda$-node assignment is a robust method for improving network bandwidth because its worst cases (worst solutions due to severe PV in the 100 generated samples) are much better than using fixed assignment. Since the achieved bandwidth is still not 100%, another important metric is the probability of completely losing connectivity between two nodes. That is, no single $\lambda$ is common between the two nodes.

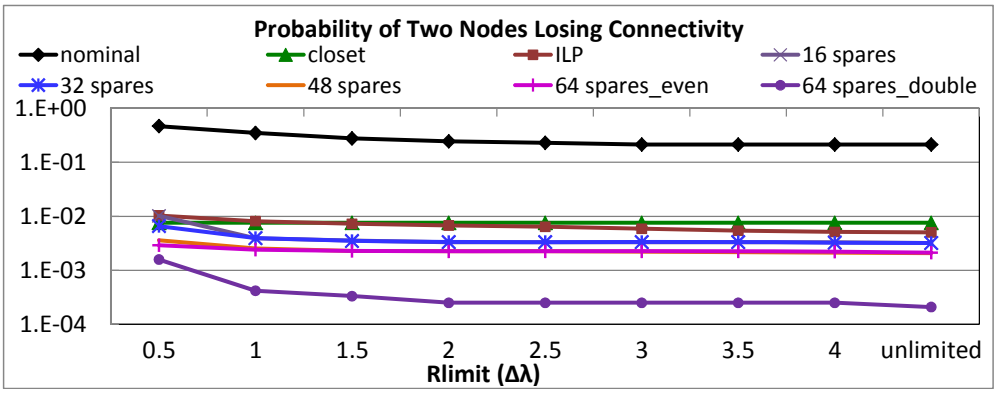



Figure 24: Probability of losing connectivity between two nodes.

Fig. 24 shows such probability on logarithmic scale. The data are collected through counting how many such pairs occur in the entire 100 samples, each having 16×15 node pairs. We did not find any disconnected node pairs in "64-spares-DEEM" and "64-spares-double-flexible". Although MinTrim does not guarantee connectivity, our experiments do show that the probability for the two schemes are very low. The next best scheme is "64-spares-double" using fixed $\lambda$-node assignment. The probability of losing one pair is $10^{-4} \sim 10^{-3}$. The next batch of schemes have similar probabilities: $10^{-3} \sim 10^{-2}$. These schemes include those with spares, ILP, and baseline-closest. Baseline-nominal has the highest disconnection rate, nearly 2 orders of magnitude worse than other schemes. This is because pair-wise disconnection often occurs in worst PV scenarios. The worst cases for "nominal" is worse than for "closest" and ILP, as shown in Fig. 19. For example, if all $\mu$rings of a node drifted too far to be correctable, "nominal" bails out but "closest" and ILP may still find a solution. In summary, MinTrim with enough spares and flexible assignment are among the best schemes in terms of network connectivity.

We illustrate the bandwidth improvement of MinTrim with an randomly selected sample in Fig. 25 with the X axle being the index of the network node. For a specific node, it may have different number of wavelengths to communicate with other nodes because the wavelengths used for transmission might not be available at each receiver node. Hence,

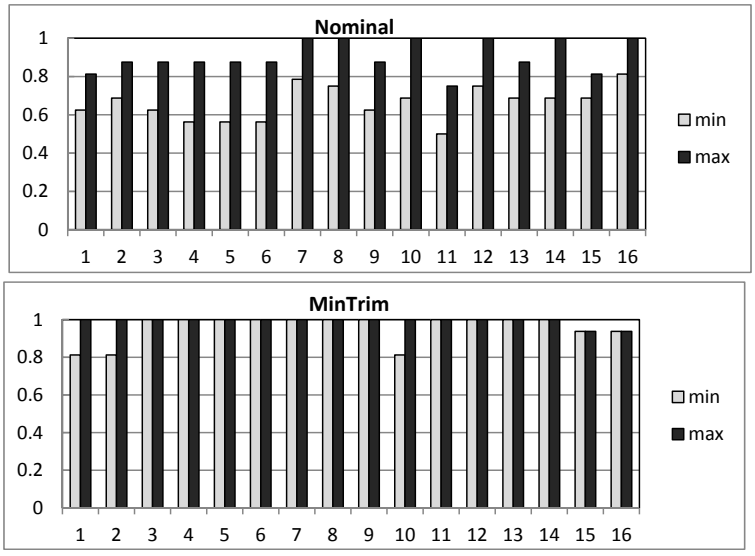



Figure 25: An example to show the maximum and minimum connection bandwidths of each node for "nominal" and MinTrim, respectively.

Fig. 25 shows the minimum and maximum connection bandwidth of each network node under baseline design and MinTrim, respectively. We can observe that after applying MinTrim, both worst and best case connection bandwidth is improved and become uniform.

### 3.4.5 Heating-only Trimming

**3.4.5.1 Normalized Bandwidth** Since correction ability on blue shift is far less than of red shift [45], many work only use heating to keep all $\mu$rings at a constant temperature [25, 26, 48], which should be close to the peak temperature of the chip to avoid blue shifts. However, it could not alleviate the red shifts introduced by PV. We measured the normalized network bandwidth achieved by baseline design, illustrated in Fig. 26 with RLimit being $2\Delta\lambda$. Compared to the one allowing $0.5\Delta\lambda$ of blue shifts shown in Fig. 18, bandwidth is degraded by 12%. After applying ILP, the normalized bandwidth reaches 70%. Adding spare rings also helps improve the successful rate of associating $\mu$rings and wavelengths, which leads to higher network bandwidth. In Fig. 26, "ILP+32spares+KL" indicates K extra rings at

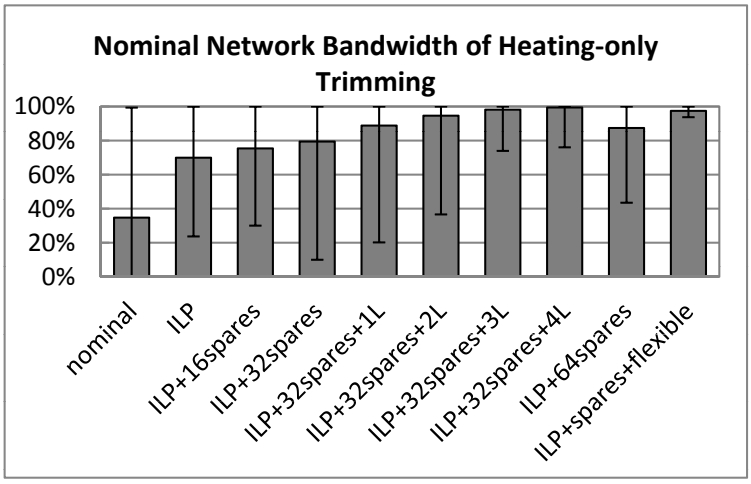



Figure 26: Normalized bandwidth achieved by heating-only trimming.

the left side of the spectrum. We can observe from the figure, the network bandwidth is close to 100% with only 4 $\mu$rings, which means that the asymmetric placement of $\mu$rings corresponding to imbalanced trimming ability can effectively mitigate the bandwidth loss. Whereas supplying more rings inside the spectrum indicated by "ILP+64spares" can only produce 87% of network bandwidth on average and flexible wavelength mapping leads 10% more of bandwidth improvement, which is still less effective than asymmetrical spare ring strategy.

**3.4.5.2 Trimming Power** Fig. 27 shows the comparisons on the trimming power generated by baseline and MinTrim under heating-only trimming method at Rlimit=$2\Delta\lambda$. MinTrim-ILP consumes higher power than "nominal" since it is able to correct much more $\mu$rings, which results in larger cumulative trimming distance. The power consumption immediately drops when spare $\mu$rings approach is applied, same as the normal power trimming method. However, power cost increases quickly with the number of the $\mu$rings placed at the left side of the spectrum because even adding one $\mu$ring might have significant impact on the $\mu$ring mapping solution generated by ILP. While supplying $\mu$rings inside the wavelength spectrum could help reduce both bandwidth loss and power consumption by 45%

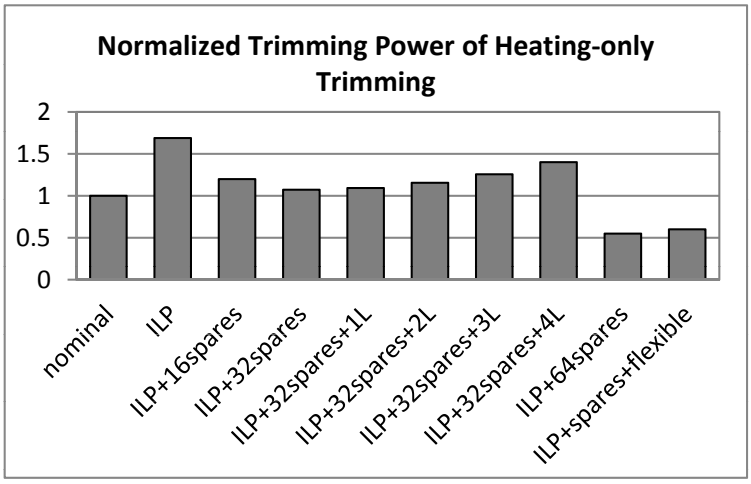



Figure 27: Normalized trimming power required by heating-only trimming.

and 63% compared to baseline design, respectively. Overall, all the spare ring strategies can effectively mitigate bandwidth loss. The tradeoff is that adding rings inside the spectrum can reduce trimming power significantly but requires doubling the $\mu$rings while supplying a couple of rings with resonate wavelengths outside the spectrum is enough to provide nearly full bandwidth but needs higher trimming power.

## 3.5 SUMMARY

PV in optical networks is a serious problem. A network can be paralyzed by PV due to variations of device dimensions and changes in resonance wavelength of $\mu$rings. Current power trimming techniques cannot solve this problem, as shown by our experiments. Our proposed technique, MinTrim, is shown to be effective in tolerating PV. The key ideas of MinTrim include using redundancy and allowing flexibility, which are natural approaches to handling variations. MinTrim improves bandwidth from 59% to 98.4% in the best cases. We also found that using redundancy is not only effective in improving bandwidth, but also in reducing power consumption which is a critical factor in optical network. A 39% trimming

power reduction is observed through MinTrim. For network architectures that do not belong to SWMR, MWSR, or MWMR, we emphasize that the first two steps of MinTrim, ILP and sparing, can always be applied. Hence, MinTrim is a general method that can be tailored to a network architecture.

MinTrim was proposed to target only the PV problem in this chapter. The ILP solvers are computationally extensive and thus are only suitable for determining the trimming configuration off-line. If trimming is to be decided on-line to address temperature (dynamic) variations, then faster algorithms have to be applied. I will thus explore simple heuristics to obtain feasible, but probably suboptimal, solutions to the above stated problems and compare these solutions with the ones obtained from the corresponding ILP in the following chapter.

# 4.0 BANDARB: MITIGATING THE EFFECTS OF THERMAL AND PROCESS VARIATIONS IN NANOPHOTONIC ON-CHIP NETWORKS

MinTrim provides an efficient PV-tolerant solution to improve the reliability of on-chip photonic networks. However, in addition to the dimensional variations of $\mu$rings caused by fabrication imperfection, temperature fluctuations (TF) across the whole chip at runtime causes dynamic variations of resonant wavelengths. Due to PV and TF, future optical networks should be designed to adapt to resonant wavelength shifts dynamically. MinTrim cannot deal with thermal variations since it uses an ILP approach which is prohibitively complex for on-line computation. Hence, this chapter introduces a bandwidth arbitration scheme, BandArb, which dynamically reassigns wavelengths to $\mu$rings and communicating nodes taking into consideration both thermal and process variations as well as network traffic. The low computational complexity of the scheme makes its suitable for run-time invocation.

The remainder of this chapter is organized as follows. Section 4.1 introduces a background about PV and TF effects as well as previous work related to addressing wavelength deviations. Section 4.2 discusses the details of the proposed BandArb methodology including static assignment and dynamic wavelength allocation. Section 4.3 shows the comparisons on bandwidth, throughput improvement and computation overhead with baseline trimming methods . Finally, Section 4.4 summaries this chapter.

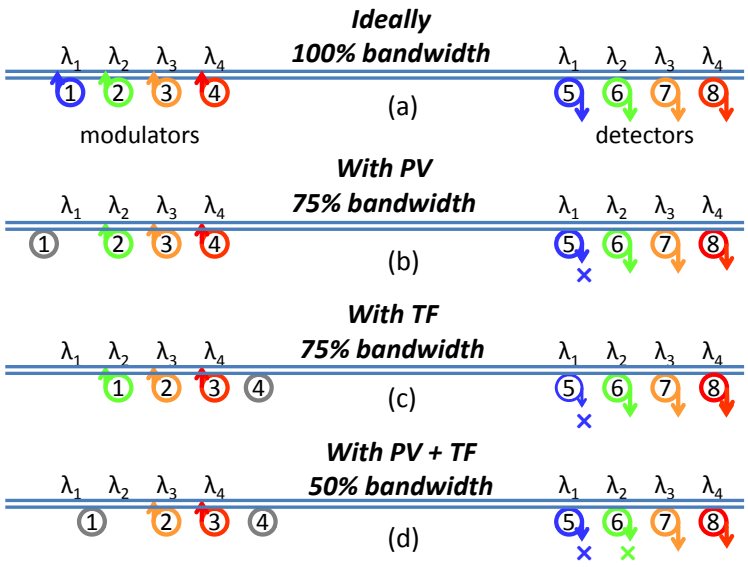



Figure 28: An example showing bandwidth loss due to process variation (PV) and temperature fluctuation (TF).

## 4.1 BACKGROUND AND RELEVANT STATE-OF-THE-ART

### 4.1.1 Severity of PV- and Thermal-shifts

The $\mu$ring based modulator at the sender side and the corresponding detector at the receiver side should resonate at the same $\lambda$. In the ideal case, a $\mu$ring fabricated perfectly can modulate and extract optical signals of its designated $\lambda$ (called nominal $\lambda$) correctly without any loss. We use a simple example to illustrate this in Figure 28(a). The sender uses $\mu$rings #1 $\sim$ #4 to modulate four wavelengths $\lambda_1 \sim \lambda_4$ indicated by different colors, and the receiver uses $\mu$rings #5 $\sim$ #8 to detect and extract the same wavelengths respectively. Under an ideal situation, both the sender and the receiver can utilize 100% of their bandwidth for transmission. Figure 28(b) shows an example when PV is present: ring #1 shifts from $\lambda_1$ and, as a result, the sender cannot resonate at $\lambda_1$. Consequently, ring #5 at the receiver does not receive any signal, which downgrades the communication bandwidth to 75%. Such a loss is static meaning that the 25% bandwidth loss is permanent.

At runtime, processor temperature tends to fluctuate (denoted as TF) and the resonance of the $\mu$rings will change with temperature. Figure 28(c) illustrates the same example when the temperature of the sender node increases (assuming no PV is present). All rings drift towards the red end of the spectrum. For ease of illustration, we assume that they all drift by $\Delta\lambda$ which is the spacing between neighboring wavelengths. Hence, ring #1 $\sim$ #3 now resonate at $\lambda_2 \sim \lambda_4$ respectively. As a result, the sender loses $\lambda_1$, 25% of bandwidth, because it has no rings to resonate at $\lambda_1$. Such loss is dynamic since the resonance drift is linear with temperature and temperature could either increase or decrease. For example, the sender could continue to lose $\lambda_2 ... \lambda_4$ if temperature continues to increase, and a similar effect takes place if temperature decreases below the nominal value (causing what is called blue shift). Note that if the $\mu$rings do not drift by multiples of $\Delta\lambda$, then the sender would lose its entire bandwidth. Finally, when we consider both PV and TF, bandwidth loss is likely to be compounded. Figure 28(d) illustrates an example where the sender loses both $\lambda_1$ and $\lambda_2$ which account for 50% of the bandwidth. If we consider PV and TF for receiver's $\mu$rings, then the bandwidth loss is even higher since only common wavelengths between a sender and a receiver constitute usable bandwidth.

For a thermal shift rate of 0.1nm/°C [17, 39, 45, 54, 80], a temperature fluctuation of 20$\sim$40°C would result in a 2$\sim$4nm of wavelength shift. The minimum spacing between adjacent wavelengths in a WDM spectrum, $\Delta\lambda$, is 0.16nm [45] $\sim$ 0.8nm [61] for silicon $\mu$ring resonators. This is determined by the bandwidth available for WDM (depending on the minimum radius of $\mu$rings) and the loss and crosstalk between two adjacent wavelength channels. Hence, the thermal shift of 2$\sim$4nm may span between 3 and 28 channels. The example in Figure 28(c) has a shift of 1 channel, as a result of 1.4$\sim$8°C temperature increase. As for PV-shift, it is reported to exceed 1nm [17, 58, 75]. In a recent demonstration of a photonic platform leveraging the existing state-of-the-art CMOS foundry infrastructure [48], the measured PV is as much as 600GHz across a wafer. Using less aggressive channel spacing of 0.8nm (100GHz) [61], 600GHz of variation corresponds to a shift of 1.25$\sim$6 channels.

### 4.1.2 Current Approaches and limitations

Without any correction, PV and TF could result in a high probability of optical channel failure. Thus, power trimming is necessary to adjust the drifting resonance of $\mu$rings. However, instead of shifting a $\mu$ring to its designated (nominal) wavelength, clever proposals were made to reduce the trimming power by tuning a $\mu$ring to the wavelength closest to its current resonant wavelength or to the nearest wavelength in a predetermined grid [19, 48]. Both methods use bit re-shuffling to reduce power by reducing the trimming distance, but do not address the limitations of trimming range, which may result in significant bandwidth losses.

MinTrim [79] is a post-fabrication scheme to realign the resonant wavelengths of $\mu$rings to achieve maximum available network bandwidth under the effects of PV. Nevertheless, one-time calculation cannot address the dynamic drift of $\lambda$ introduced by TF. So MinTrim needs to be applied to re-organize the wavelengths after thermal drifting to handle PV and TF together. However, the ILP optimization algorithm used in MinTrim introduces a latency that is unaffordable at run time.

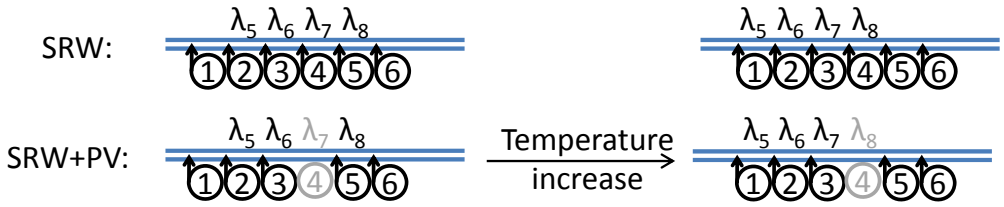



Figure 29: The limitations of SRW [45] in the presence of PV. Grey $\lambda$s are not used.

Sliding Ring Window(SRW) [10, 45] is a technique that adds thermal $\mu$rings at both ends of the spectrum, as shown in the upper part of Figure 29, to improve the power efficiency and address non-correctable $\lambda$ shifts caused by TF. In this example, the network node uses $\lambda_5 \sim \lambda_8$ for transmission and two extra $\mu$rings, ring #1 and #6 are added to the left and right of the original group of four $\mu$rings. When temperature increases, all $\mu$rings shift towards red, and if the shift is over $0.5\Delta\lambda$, power trimming would further red-shift them to the next

channel instead of blue-shift them back to nominal positions. As a result, ring #1∼#4 will resonate at $\lambda_5 \sim \lambda_8$ respectively, preserving 100% of the bandwidth. Symmetrically, ring #6 is used when temperature decreases. However with PV, SRW would still not be able to fully preserve the bandwidth. This is illustrated in the lower part of Figure 29, where it is assumed that ring #4 has drifted away from $\lambda_7$. Neither thermal $\mu$rings #1,#6 nor normal $\mu$ring #4 is able to resonate at $\lambda_7$. Hence, when all $\mu$rings' resonance shift to the next channel due to TF, there is no $\mu$ring correctable to $\lambda_8$. Although the indices of resonance wavelengths change with the temperature to accommodate TF, the aggregated bandwidth is still reduced by 25%.

Since SRW and MinTrim provide an efficient solution to deal with TF and PV, respectively, one intuitive way for addressing both variations is to adopt the static mapping solution generated by MinTrim to realign the $\mu$rings (including the thermal $\mu$rings) to overcome PV, and then applying SRW as a thermal adjustment to compensate for the changes in temperature. This method is termed as "$MinTrim + SRW$" in this chapter and used as one of the baseline designs. Specifically, when temperature increases by $\Delta T$, the resonant wavelength for each $\mu$ring shifts from the wavelength assigned to it by MinTrim, say $\lambda_r$, to a new wavelength $\lambda_r + \delta$. Then, SRW trims the $\mu$ring to the next wavelength higher than $\lambda_r + \delta$ if $\delta > 0.5\ \Delta\lambda$ or blue-shift the $\mu$ring back to $\lambda_r$, otherwise. Assuming that $\Delta\lambda = 0.8nm$ and $\delta\lambda/\delta T = 0.1$nm/°C), then $\Delta\lambda$ corresponds to 8 degrees and SRW trims the $\mu$ring to $\lambda_k$ where

$$k = r + (int)((\Delta T + 4)/8). \tag{4.1}$$

However, the trimming distance between the nominal $\lambda$ of the $\mu$ring and $\lambda_k$ may be larger than the correction range allowed by the power trimming technology, even though the distance between that nominal $\lambda$ and $\lambda_r$ is within that range. In such a case, the SRW correction fails. For example, if we assume a temperature variation of 3 degrees in Figure 30, it may not be possible to trim $\mu$ring #4 back to $\lambda_8$ if PV had originally caused it to drift from $\lambda_8$ by the allowable correction range and it was trimmed back to $\lambda_8$ by MinTrim. In this example, TF causes an increase of lost bandwidth from 25% to 50%. In section 4.3, we will show that even small temperature variations may result in noticeable bandwidth degradation.

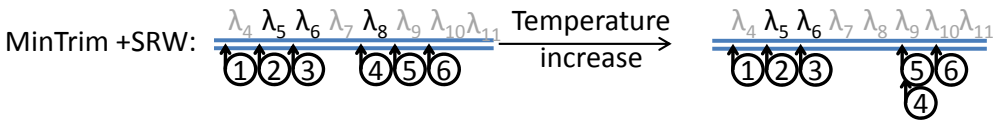



Figure 30: The bandwidth loss under "$MinTrim + SRW$".

Instead of combining individual schemes designed for PV and TF, we propose, in the next section, a scheme for simultaneously mitigating the effects of PV and TF at any given temperature with reasonable computation time.

## 4.2 BANDARB:DEALING WITH BOTH PV AND TF

Before proposing solutions for providing near-full bandwidth under TF and PV, we first introduce the optical network architecture that we will use to illustrate and evaluate our designs. We also introduce some notation that will be used to precisely describe the proposed solutions.

### 4.2.1 Network Architecture

The technique proposed in this chapter for mitigating the effects of PV and TF, BandArb, is illustrated using the single-write-multiple-read (SWMR) crossbar architecture proposed in [29, 50]. In this architecture, there is a total of $n$ nodes, $N_1$,...,$N_n$, connected by WDM optical waveguides, each supporting $w$ wavelengths, $\lambda_1, ..., \lambda_w$. Every node uses an exclusive set of $m = w/n$ contiguous wavelengths to send data and can received data via the remaining $w-m$ wavelengths. Hence, all nodes can transmit simultaneously without the need for arbitration and each node can simultaneously receive data from the other $n-1$ nodes. We define the sets $\Lambda_{i,t}$ and $\Lambda_{i,r}$ as the set of wavelengths that are assigned to node $N_i$ (nominal wavelengths) for transmission and reception, respectively. In our evaluation, we will use an example network

with $n = 16$ nodes and $w = 64$ wavelengths. Hence, $\Lambda_{i,t} = \{\lambda_{m(i-1)+1}, ..., \lambda_{mi}\}$, and $\Lambda_{i,r} = \{\lambda_1, ..., \lambda_w\} - \Lambda_{i,t}$. That is, $N_1$ uses $\lambda_1 ... \lambda_4$ for transmission, $N_2$ uses $\lambda_5 ... \lambda_8$ , and so on. Figure 31(a) shows, as an example, the wavelengths used by $N_2$ for transmission and reception.

We assume that the SWMR crossbar uses the low power mechanism proposed for Firefly [51], where all the detector $\mu$rings stay turned off (tuned off) by default. A sender will notify the receiver to turn on its detector $\mu$rings prior to a transmission through a reservation broadcast bus.

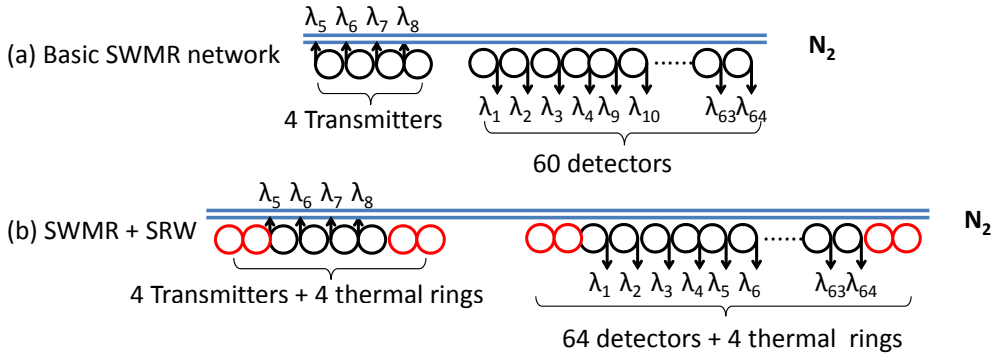



Figure 31: SWMR design of $N_2$ with and without SRW [45].

As in SRW, thermal $\mu$rings [45] are added to mitigate TF. We use 4 extra thermal $\mu$rings (2 per end of spectrum) to cover a wide range of temperature variation. Note that in the basic SWMR design [29, 50], $\mu$rings at the receiver side in each node are tuned to the wavelengths that are not used for transmission by that node. Thus, the detectors might not resonate a set of contiguous wavelengths. For example, the detectors of $N_2$ receive data via $\lambda_1 \sim \lambda_4$ and $\lambda_9 \sim \lambda_{64}$. Hence when temperature increases, $\lambda_9$ become unavailable due to the "hole" among the wavelengths. To avoid bandwidth loss, 4 more detectors are added to construct a consecutive wavelength group for detectors. Thus, for every node, $\Lambda_{i,r} = \{\lambda_1, ..., \lambda_{64}\}$. Figure 31(b) shows the SRW design for $N_2$. There are total of 8 transmitters and 68 detectors for each node to maintain bandwidth within a wide temperature window. Each node, $N_i$, has transmitting $\mu$rings (used for signal transmission) that resonate at the wavelengths in $\Lambda_{i,t}$ and receiving $\mu$rings (used for signal reception) that resonate at the wavelengths in $\Lambda_{i,r}$.

Ideally, in the absence of PV and TF, each node resonates at its nominal wavelength and hence, all the wavelengths in $\Lambda_{i,t}$ can be used for transmission and all the wavelengths in $\Lambda_{i,r}$ can be used for reception. However, due to PV and TF, the resonance wavelength of each $\mu$ring (including thermal $\mu$rings) may deviate from its nominal wavelength, and power trimming can be used to shift its resonance either back to its nominal wavelength or to some other wavelength in $\lambda_1, ..., \lambda_w$. However, there is a severe limit on the correction ability of power trimming which translates into a constraint on the wavelength distance that can be shifted for correction purpose. Hence, with PV and TF, the trimming distance constraint determines the set of wavelengths that can be used for transmitting signals. We call this set the set of "potential resonant wavelengths", $\Pi_{i,t}$. A wavelength, $\lambda_k$ is in $\Pi_{i,t}$ if any of the transmitting $\mu$rings at node $N_i$, including the thermal $\mu$rings, can be power trimmed to $\lambda_k$. The set $\Pi_{i,r}$ is similarly defined for the $\mu$rings used for detecting signals.

Clearly, not all the wavelengths in $\Pi_{i,t}$ can be used since a $\mu$ring can only be tuned to one wavelength in $\Pi_{i,t}$. However, after power trimming is applied to the $\mu$rings (according to some $\mu$ring-to-wavelength re-alignment algorithm), we define the set $\Sigma_{i,t}$ as the subset of wavelengths in $\Pi_{i,t}$ that, can actually be used for transmission at $N_i$. That is, a wavelength is in $\Sigma_{i,t}$ if some transmitting $\mu$ring is actually power trimmed to that wavelength as a result of the re-alignment algorithm. The set $\Sigma_{i,r}$ is similarly defined for the wavelengths used for detection.

Note that SWMR avoids arbitration by ensuring that two nodes do not use the same wavelength for transmission. Noting that for any two nodes, $N_i$ and $N_j$, the sets $\Sigma_{i,t}$ and $\Sigma_{j,t}$ can include a common wavelength, we specify $O_{i,t}$ as the subset of wavelengths in $\Sigma_{i,t}$ owned by $N_i$ such that $O_{i,t}$ and $O_{j,t}$ for any $i$ and $j$ do not intersect. Hence, no transmission interference will result if each $N_i$ transmits using the $\mu$rings tuned to wavelengths in $O_{i,t}$. Obviously, $O_{i,t}$ can be specified as the intersection of $\Sigma_{i,t}$ and $\Lambda_{i,t}$, which is what is used in SRW. In Table 4, we summarized the notations for the 4 different wavelength sets defined in this section.

Table 4: Summary of the wavelength sets notation

| Set | Definition |
|---|---|
| $\Lambda_{i,t/r}$ | The set of nominal wavelengths that are initially assigned to node $N_i$ for transmission and reception, respectively |
| $\Pi_{i,t/r}$ | The set of "potential resonant wavelengths" within the correction range of $\mu$rings |
| $\Sigma_{i,t/r}$ | The set of wavelengths corresponding to the tuned $\mu$rings at $N_i$ after power trimming |
| $O_{i,t}$ | The set of wavelengths owned by $N_i$ and used for transmission |

### 4.2.2 Coarse-Grained BandArb (CG-BandArb)

In this section, we introduce coarse-grained BandArb, a two-step methods to mitigate the bandwidth loss caused by PV and TF. The first step is to maximize the sets $\Sigma_{i,t}$ and $\Sigma_{i,r}$ by re-aligning $\mu$rings and wavelengths locally within each node $N_i$ (section 4.2.2.1). As discussed in the previous section, the set $O_{i,t}$ of wavelengths used for transmission by $N_i$ can be computed as the intersection of $\Sigma_{i,t}$ and $\Lambda_{i,t}$. However, this local scheme restricts the set $O_{i,t}$ of wavelengths owned by $N_i$ (used for transmission) to be a subset of $\Lambda_{i,t}$, which may result in bandwidth degradation. The available transmission bandwidth at each node may be increased by removing this restriction and computing $O_{i,t}$ at each node using global knowledge of $\Sigma_{i,t}$ for $i = 1, ..., n$. In section 4.2.2.2, we introduce the second step of CG-BandArb, which is a global algorithm for enlarging the sets $O_{i,t}$, $i = 1, ..., n$.

**4.2.2.1 Local Wavelength Re-alignment** When a node looses bandwidth due to TF, it can locally re-align its $\mu$rings to wavelengths to reclaim some of the lost bandwidth. Figure 32 shows the same example used in Figure 3 where $\mu$ring #4 could not be tuned to

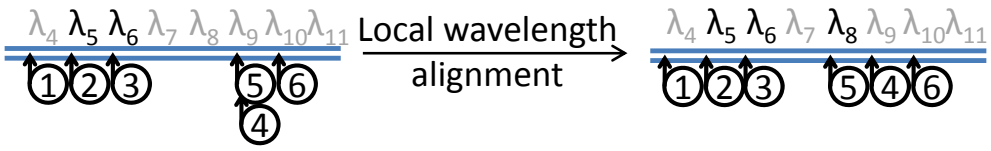



Figure 32: Increasing the bandwidth using local re-alignment.

$\lambda_8$. With local re-alignment, it may be possible to tune $\mu$ring # 5 to $\lambda_8$ if this is within the correction range, thus recovering 25% more bandwidth compared to "$MinTrim + SRW$". In this case, $\Sigma_{2,t} = \{\lambda_4, \lambda_5, \lambda_6, \lambda_8, \lambda_9, \lambda_{10}\}$ and $O_{2,t} = \{\lambda_5, \lambda_6, \lambda_8\}$.

In what follows, we present three local realignment algorithms. The first two have been proposed in the literature in the context of optical links without thermal $\mu$rings, while the third one is a new algorithm that achieves a tradeoff between realignment efficiency and algorithm complexity. All three algorithms can be used for aligning the transmitting as well as the receiving $\mu$rings, but will be described only for the transmitting $\mu$rings.

**Nominal**

This algorithm realigns each $\mu$ring to its nominal wavelength. If the realignment exceeds the allowable correction range, the $\mu$ring is tuned off. The same thermal compensation mechanism shown by Equation 4.1 is applied to nominal.

**Closest**

This algorithm trims each $\mu$ring to resonate at its closest wavelength instead of its originally assigned nominal wavelength [19]. This algorithm results in a low trimming power due to the short trimming distances. However, its realignment capability is also limited, which results in over 50% of bandwidth loss [79].

Simple mapping schemes such as Closest and Nominal introduce a small circuit overhead. The temperature of neighboring $\mu$rings changes together as a group [45], so only few hardware sensors are necessary to detect drifting $\mu$rings. However, the solutions generated by these simple algorithms may fail to find $\mu$rings for the wavelengths generated by the laser source which wastes precious optical network resources.

**Wavelength Matching (WM)**

The goal of this local realignment heuristics is to maximize the size of $\Sigma_{i,t}$ and $\Sigma_{i,r}$ at $N_i$, which maximizes the number of $\mu$rings that can be used for transmission and detection. This has the potential of maximizing the size of $O_{i,t}$, the set of wavelengths owned by $N_i$ for transmission, especially when the global wavelength re-allocation algorithm described in the next subsection is applied.

**Algorithm 1** WM: Compute $\Sigma_{i,t}$ at node $N_i$.

```
for (∀λ_k ∈ Π_{i,t}) do
  if R[k]= {mr} /* only ring mr can be tuned to λ_k */ then
    match(mr, λ_k)
  else
    if (Nominal(mr) = λ_k for some mr ∈ R[k]) then
      match(mr,λ_k)
    else
      match (mr,λ_k) for any mr ∈ R[k]
    end if
  end if
end for
```

The Wavelength Matching heuristics, WM, is illustrated in Algorithm 1. When the algorithm is applied at node $N_i$, the set $R[k]$ is defined as the set of $\mu$rings at node $N_i$ that can be tuned to $\lambda_k$. Also, for any $\mu$ring, $mr$, $Nominal(mr)$ is defined as the nominal resonant wavelength of $mr$. As defined earlier, $\Pi_{i,t}$ stands for the potential resonant wavelengths in $N_i$ which includes all the reachable wavelengths. Finally, the procedure Match($mr$,$\lambda$) tunes $mr$ to $\lambda$, adds $\lambda$ to $\Sigma_{i,t}$ and removes $mr$ from any set $R[k]$ that includes $mr$ (because one $\mu$ring cannot resonate at multiple wavelengths simultaneously). In WM, if only one $\mu$ring can resonate at $\lambda_k$, then naturally this $\mu$ring is assigned to $\lambda_k$. Otherwise $\lambda_k$ is mapped to the nominal ring if it is in $R[k]$. If not, one of the rings that can be trimmed to $\lambda_k$ is selected. We evaluated the effectiveness of different options for this last step: selecting the $\mu$ring with lowest or highest index or the one belonging to the fewest number of $R[k]$ sets. We observed that different selection options have little impact on the resulting bandwidth.

**4.2.2.2 Global Wavelength Re-allocation** The second step in BandArb is to reclaim more of the bandwidth lost due to TF by flexibly specifying the owner sets, $O_{i,t}$, $i = 1, ..., n$, and removing the restriction that $O_{i,t}$ should be a subset of $\Lambda_{i,t}$. The condition that the sets $O_{i,t}$ are not intersecting should still be enforced to ensure that transmission wavelengths of different nodes are disjoint. However, this global re-allocation requires knowledge of the wavelengths that each node can use for transmission after the $\mu$rings are locally tuned (the sets $\Sigma_{i,t}$, for $i = 1, ..., n$). Figure 33 illustrates the additional benefit of global re-allocation after local wavelength alignment demonstrated in Figure 5. Specifically, global re-allocation may add $\lambda_9$ to $O_{2,t}$ if that wavelength is not used for transmission by any other node.

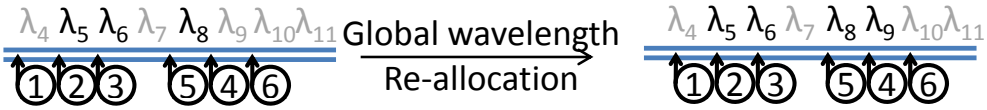



Figure 33: The effect of global wavelength re-allocation.

Since major TF spans milliseconds or even seconds, CG-BandArb is needed at a coarse granularity and can be triggered only when bandwidth loss reaches a certain threshold value (for example after the loss of one channel). The proposed heuristic to re-allocate wavelengths among nodes is shown in Algorithm 2. In that algorithm, CanUse(k) is defined as $\{N_i; \lambda_k \in \Sigma_{i,t}\}$, the set of nodes that can use $\lambda_k$ for transmission. If two nodes can use $\lambda_k$ for transmission, the algorithm prefers the original owner $N_i$ of $\lambda_k$ ($\lambda_k \in \Lambda_{i,t}$). If the original owner node of $\lambda_k$ does not have a $\mu$ring tuned to $\lambda_k$, then any node that can be tuned to $\lambda_k$ is chosen as its owner. To improve the success rate of allocating all wavelengths, the node which owns the least number of wavelengths can be selected as the owner of $\lambda_k$.

We propose Algorithm 2 to globally re-allocate wavelengths after any local wavelength alignment algorithm is applied. In Section 4.3, we evaluate the effectiveness of global re-allocation after each of the three local alignment algorithms described in the previous section, and we call the three resulting schemes BandArb_nominal, BandArb_closest and BandArb_WM. It will be shown that applying global re-allocation always improves the bandwidth, irrespective of the local alignment algorithm used. It will also be shown that the

**Algorithm 2** Compute $O_{i,t}$, $i = 1, ..., n$ from $\Sigma_{j,t}$, $j = 1, ..., n$.

```
for k = 1 → w (w is total number of wavelengths) do
  Construct CanUse(k) from Σ_{i,t}, i = 1, ..., n
  if CanUse(k) = {N_i} /* only N_i can use λ_k */ then
    add λ_k to O_{i,t}
  else
    if λ_k ∈ Λ_{i,t} and N_i ∈CanUse(k) then
      add λ_k to O_{i,t}
    else
      add λ_k to O_{i,t} for some N_i ∈ CanUse(k)
    end if
  end if
end for
```

performance of the two-steps BandArb heuristics is able to achieve a bandwidth that is comparable to the optimum global ILP approach which is too complex to execute at run time after each TF.

**4.2.2.3 Implementation of CG-BandArb** Local wavelength realignment and global wavelength reallocation are triggered either when the drift in a $\mu$rings reaches to a certain threshold value that makes it uncorrectable, or equivalently, if a large temperature variation is sensed. As indicated in [10, 19], it is reasonable to assume the existence of a circuit that detects the status of the $\mu$rings' resonance to trigger CG-BandArb. The network can still be utilized with degraded bandwidth during the execution of CG-BandArb. Finally, a control module should control the $\mu$ring tuning process after receiving the updated wavelength alignment results and notify the detection unit once the resonant wavelengths are stabilized.

When CG-BandArb is invoked at a node, $N_i$, due to an excessive drift in a receiving $\mu$ring, the updated $\Sigma_{i,r}$ can be locally computed and broadcast to all other nodes. Every node, $N_j$ needs to know $\Sigma_{q,r}$ for every other node, $N_q$, in order to determine the wavelengths that can be used for sending messages to that node. Specifically, even if a wavelength, $\lambda_k$ is

in $O_{j,t}$, $N_j$ cannot use $\lambda_k$ to send messages to $N_i$ if $\lambda_q$ is not in $\Sigma_{q,r}$. The information about $\Sigma_{q,r}$ for $q = 1, ..., n$ can be kept in a Receiver Wavelength Availability Table (RWAT) of $n$ rows and $w$ columns using $nw$ bits, where $w$ is the number of wavelengths. Each node can keep a copy of this table and update it after receiving a broadcast message with an updated $\Sigma_{i,r}$ from some node, $N_i$.

When CG-BandArb is invoked at $N_i$ due to an excessive drift in a transmitting $\mu$ring, the updated $\Sigma_{i,t}$ can be locally computed at $N_i$ and broadcast to other nodes, but the updated owner set, $O_{i,t}$, cannot be computed without the knowledge of $\Sigma_{j,t}$ for all nodes, $N_j$, $j = 1, ..., n$. For this reason, each node keeps track of these sets in a Sender Wavelength Resonance Table (SWRT) similar the the RWAT and updates this table after receiving a broadcast message with an updated $\Sigma_{i,r}$ from $N_i$. Given that a change in $\Sigma_{i,r}$ can change the set $O_{j,t}$ for multiple nodes, every node should execute Algorithm 2 after receiving the updated $\Sigma_{i,r}$. Alternatively, only $N_i$ can compute all the sets $O_{j,t}$, $j = 1, ..., n$ and broadcast them to the other nodes.

Since SWRT or RWAT table in each node is updated by the same broadcast message, there is no consistency problem when messages reach every node within one cycle. If the network frequency is higher, e.g. 5GHz, it is possible that some nodes receive the message earlier than other remote nodes do. So we enforce that the table is updated in the same cycle at which all nodes receive the message.

### 4.2.3 Fine Grained BandArb (FG-BandArb)

In addition to allowing effective re-allocation of wavelengths to nodes at a coarse granularity, the thermal $\mu$rings provides the opportunity for each node to transmit using a $\mu$ring that is tuned to a wavelength that is owned by a different node, as long as this wavelength is idle. This is the goal of FG-BandArb. For example, in Figure 34 it is assumed that CG-BandArb resulted in $O_{2,t} = \{\lambda_5, \lambda_6, \lambda_8, \lambda_9\}$ (see Figure 6). With FG-BandArb, however, $N_2$ can also use $\lambda_4$ and $\lambda_{10}$ for transmission, as long as these wavelengths are not used by other transmitters, thus increasing the transmission bandwidth to 150%. Because of the thermal $\mu$rings, the set $\Sigma_{i,t}$ at any node $N_i$ may be a super set of $O_{i,t}$. Hence, $N_i$ can dynamically

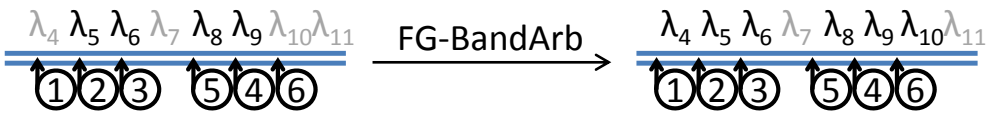



Figure 34: wavelength borrowing in FG-BandArb.

borrow a wavelength, $\lambda_k$ that is in $\Sigma_{i,t}$ - $O_{i,t}$ for transmission since, as a result of CG-BandArb, it already have a $\mu$ring tuned to $\lambda_k$. However, because $\lambda_k$ may also be in $\Sigma_{j,t}$ of some other node, $N_j$, an arbitration algorithm should be used to guarantee that $\lambda_k$ is used for one transmission at a time. Moreover, the availability of wavelengths to borrow depends on the network traffic conditions that may vary every cycle. Hence, arbitration should be applied every cycle.

With wavelength borrowing allowed, multiple nodes might want to access the same wavelength in any given cycle. For example, if $\lambda_{10}$ is owned by $N_3$ and $N_3$ is not be able to release it directly to $N_2$ because $N_4$ might have a temperature decrease and may also be able to access $\lambda_{10}$. Therefore, the major question is how to borrow wavelengths from their owners to regain lost bandwidth or even obtain extra wavelength channels?

FG-BandArb is a distributed arbitration mechanism in which each node relies on the global wavelength information stored in the binary tables, RWAT and SWRT, generated by CG-BandArb to reach the same consensus about the availability of a wavelength for borrowing. Moreover, in order to give priority in arbitration to the owner node of a wavelength, a third table, the Wavelength Ownership Table (WOT) is used to store the sets $O_{i,t}$ for $i = 1, ..., n$. As was the case with the other two tables, the size of WOT is $nw$ bits.

**4.2.3.1 The Wavelength Arbitration algorithm** To arbitrate borrowing of wavelengths, a node, $N_i$, broadcasts a 2-bit request before it transmits during a given cycle. One bit to indicate that it will use the wavelengths that it owns (the ones in $O_{i,t}$) and the other to specify whether it wants to borrow wavelengths that it does not own. Because we assume a Reservation-assisted SWMR crossbar [51], this broadcast is embedded in the wake-up mes-

sage that is sent to allow the receiver to turn on its detectors. Since this request is carried on a broadcast bus, all nodes receive it and, given RWAT, SWRT and WOT use the same algorithm to reach the same decision.

The arbitration algorithm is a priority-based scheme that assigns priorities to the nodes that can use the same wavelength. Specifically, the owner node for a $\lambda$ has the highest priority to use it. Hence, if $N_i$ requests to send a message in a cycle, all wavelengths in $O_{i,t}$ are not available to other nodes during that cycle. However, if the owner of $\lambda$ does not request transmission, then the nodes that have $\lambda$ is their $\Sigma_{j,t}$ sets are next with descending priorities given to the node with the smaller owner set (ties are broken in favor of smaller id). For example, suppose that WOT and SWRT show that $\lambda_5$ is owned by $N_2$ but can be requested by $N_1$ and $N_3$. Then $N_2$ has the highest priority, and $N_1$ has the next priority if $O_{1,t}$ is smaller than $O_{3,t}$.

Because all priority information is implicitly given in the bit tables, every node can locally calculate if it can obtain a wavelength as long as it knows its competitors. The arbitration algorithm can be made more effective if the information about the destination of the requester (transmitted as part of the wake up message in [51]) is taken into consideration during the arbitration. Specifically, $N_i$ looses its priority for a $\lambda$ if it is sending a message to a destination $N_j$ that cannot receive on $\lambda$ ($\lambda$ is not in $\Sigma_{j,t}$). For example, if $N_1$ can use $\lambda_1 \sim \lambda_4$ to send data but the receiver of its message cannot receive on $\lambda_1$. Then the arbiter at each node marks that only $\lambda_2 \sim \lambda_4$ are not available for borrowing.

The "notification" step for broadcasting requests can be parallelized with other routing stages. Hence, the optical router can still have 4 pipeline stages : (1) buffering/routing/notification, (2) reservation/arbitration, (3) crossbar traversal and (4) transfer to the remote router or a local node if the destination node shares the same optical router. To complete the request transmission in a single cycle, the cycle time should accommodate the transmission delay in the worst case plus the small control delay. Considering a 400 $mm^2$ die size, the speed of light is 10.45ps/mm and the latency of optical/electric/optical conversion is 75ps [29]. Therefore, transmission can complete at a 1GHz network frequency.

#### 4.2.3.2 Adaptive Transmission Based on Availability of Wavelengths

Because resources are allocated at the wavelength-level in BandArb, it is reasonable to reconsider the packaging of the network messages. In the baseline without resonance deviations, one node always uses the full bandwidth ($m = n/w$ wavelengths) to send a flit at a time. With PV, TF and run-time wavelength allocation, each source-destination pair (or each transmission in the case of FG-BandArb) may have available for transmission a different number of wavelength varying between 1 and $m + 4$ (assuming 4 thermal ring). Hence, for a better utilization of available bandwidth, we allow the transmission unit size to vary with the available number of channels. Suppose that each node needs one clock cycles using 4 $\lambda$s to transmit one packet. If $N_1$ needs to send four packets and could claim six wavelengths by borrowing two wavelengths from $N_2$ successfully, while $N_3$ are contending for the wavelengths owned by $N_2$ but fails due to lower priority. Then $N_1$ will use all six wavelengths for the first two cycles and will not request the two additional wavelength for the third cycle, thus allowing $N_3$ to borrow them.

Finally, we note that the BandArb design could be leveraged by other optical crossbar architectures such as Multiple-Write-Multiple-Read (MWMR) and Multiple-Write-Single-Read (MWSR). In fact, MWMR and MWSR need already an arbitration mechanism since multiple senders contend for the same set of wavelengths. Moreover, MWMR and MWSR can also benefit from adaptive transmission size to fully utilize the varying communication bandwidth among different node pairs.

## 4.3 EVALUATIONS

In this section, we evaluate the effectiveness of BandArb using both synthetic traffic and real traffic traces from PARSEC [8] and SPEC CPU 2006 [1] benchmarks.

### 4.3.1 PV and TF Modeling

The characteristics of the variations in optical devices are close to PV in CMOS devices [13, 43]. Specifically, PV can be classified into die-to-die (D2D), and within die (WID) variations including systematic and random variations among transistors. Hence, we use VARIUS [56], a PV modeling infrastructure, to model PV in $\mu$rings. We adopted the parameters of variations and modeling methodology used in MinTrim [79]: 0.61nm of WID variations and 1.01nm of D2D variations. The mean value of wavelength is the nominal wavelength of the $\mu$ring. The spectrum of the 64 wavelengths starts at 1550nm with a channel spacing of 0.8nm. Another parameter is the density $\phi$ that determines the range of WID spatial correlation and is set as 0.5. With these parameters we used VARIUS to generate 100 sample dies. We assumed that current injection can correct up to $0.5\Delta\lambda$ towards blue [39] and up to $2\Delta\lambda$ towards red (the trimming constraint on heating).

The thermal traces were generated with GEM5 [9], McPAT [38] and HotSpot [62]. The power trace of each system component was produced by the first two simulation tools with the time step being 1 ms. For the CMOS-compatible nanophotonic interconnects, we assume the same chip specification as CMOS technology for the thermal modeling in HotSpot [62], e.g. thermal conductivity. The chip layout is drawn according to the area data of different on-die components such as processor core, cache, memory controller, etc, obtained by scaling the results of McPAT [38] to 22 nm technology.

### 4.3.2 Simulation Methodology

We performed the simulation with a cycle-accurate network simulator extended from Noxim [49]. A 16-tile network with radix-16 optical crossbar similar to the one shown in Figure 31(b) is modeled. Packet sizes of 1 and 4 flits are used to send control and data message, respectively. The simulation configuration is listed in Table 5.

For CG-BandArb, the network bandwidth stays the same as long as the temperature does not change. Thus, we use the average pair-wise bandwidth as the indicator of effectiveness for wavelength alignment and reallocation schemes. Since both sender and receiver may have different sets of $\lambda$s that can not be accessed, only the common $\lambda$s between the two nodes

are counted towards the effective bandwidth. In addition to experimenting with synthetic temperature settings, we also use real thermal traces collected for PARSEC and SPEC CPU 2006 benchmarks.

To evaluate the performance aspects of FG-BandArb that depends on run-time traffic conditions for wavelength allocation, we experiment with multiple synthetic benchmarks including uniform random traffic ("UR"), where each node uniformly injects packets into the network with random destinations, as well as permutation traffic, where each node has specific destination nodes. We evaluate bit-complement("BC"), transpose("TP"), bit-reversal("BR"), shuffle("SF") and two types of hotspot traffic("HT1" and "HT2"). The first type of hotspot traffic lets all nodes send requests only to the four nodes at the left corner in the network while in the other type each tile sends messages to one node with higher probability than others. We also use GEM5 [9] to collect real network traffic traces with PARSEC and SPEC CPU 2006 benchmarks. The evaluation metrics, network throughput and latency, are measured as the average receiving rate at each core and round trip delay between sending a request and receiving its reply, respectively.

We compared CG-BandArb with several wavelength mapping designs with both simple and complicated algorithms to show the tradeoff between design complexity and effectiveness. Since ILP is too complex for run-time, the solutions produced by ILP under different temperatures are only used as an ideal baseline (denoted as "$ILP$") that cannot be realistically reached. For simplicity, we use some abbreviations of the compared designs. CG-BandArb using the $\Sigma$ sets generated by closest, nominal or WM, is denoted as "BA_closest", "BA_nominal" or "BA_WM", respectively. "MS" and "FG-BA" are short for (MinTrim+SRW) and FG-BandArb, respectively.

### 4.3.3 Evaluation of Network Bandwidth

**4.3.3.1 Comparisons of Local Wavelength Re-alignment** First, we evaluate the effectiveness of local wavelength re-alignment algorithms without the global re-allocation of wavelengths. Figure 35 compares the network bandwidth for the different local schemes (Closest, Nominal, WM, MinTrim+SWR and ILP) normalized to the bandwidth in a network

Table 5: System configuration

| | |
|---|---|
| Network organization | 16 tiles, 1 cores per tile |
| Cores | 4-thread, 2 GHz |
| Crossbar Radix | 16 |
| Private L1I/D | 32/64KB per core,<br>2-cycle hit time, write-through |
| Shared L2 | 2MB per tile, 8-way, 64B line,<br>10-cycle hit time, write-back |
| Memory Controller | Four, located in the edges of the chip |
| Router | 1GHz frequency, 4-stage pipeline. |
| Packet size | 1/4 flits |
| Buffer space | 4 flits/VC, 2VCs/input |
| Network traces | Synthetic and PARSEC,<br>SPEC CPU 2006 benchmarks |

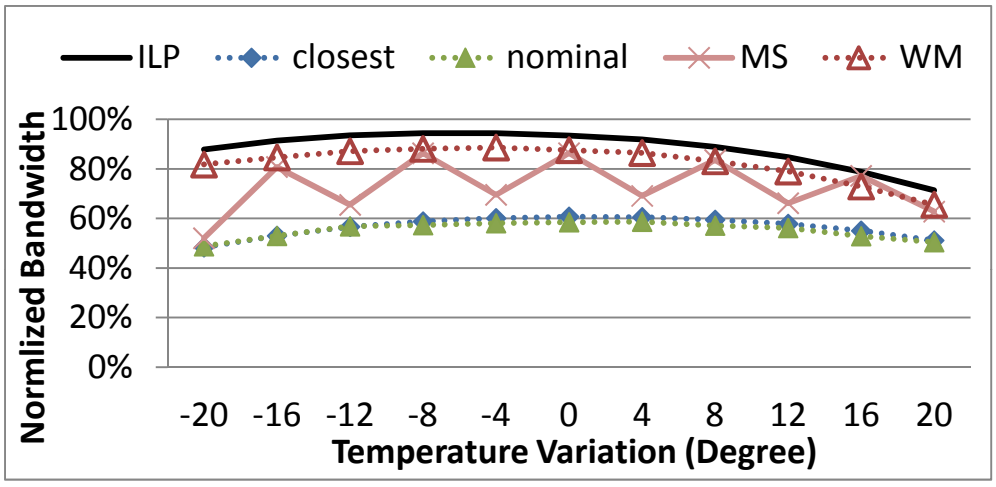



Figure 35: Network bandwidth Vs. temperature variations for local wavelength re-alignment normalized to the bandwidth in the absence of PV and TF.

without any PV and TF. The bandwidth is plotted for different temperature variations of the network nodes. From the figure, we observe that (1) the bandwidth drops when temperature variation increases because the enlarged thermal shift paralyzes more $\mu$rings; and (2) bandwidth distribution with temperature is asymmetric due to the different trimming ranges on the red and the blue directions. Overall, WM can utilize 88% of the bandwidth with no TF and 65% under large TF; while the ideal approach, ILP, utilizes 94% and 71% bandwidth in the best and worst cases, respectively. Therefore, WM, which outperforms all the other algorithms, shows a performance comparable to that of ILP.

The bandwidth curve of "MS"(MinTrim + SRW) under different temperature shows periodical zigzag shape and indicates that a change of as little as 1 degree can lead to a noticeable bandwidth loss. It is mainly because of the step function implied in Equation 4.1. To explain the trend, we use the part of the curve in the 0-8 degree range of temperature variation as an example. When the resonant wavelength of a $\mu$ring is between two channels $\lambda_i$ and $\lambda_{i+1}$, it is shifted to whichever channel is closer. Since the allowable blue shift is only $0.5\Delta\lambda$, it is less likely that the $\mu$ring is corrected back to the lower wavelength when the correction distance increases due to TF. Hence, the bandwidth keeps dropping in the range of 0 to 4 degrees. But when the temperature shift is beyond 4 degrees, the $\mu$ring is mapped to the next channel $\lambda_{i+1}$ instead of $\lambda_i$. Then the resonant wavelengths become closer to the target when temperature increases, which reduces the trimming distance and contributes to the increase of bandwidth. "MS" has low tolerance to small temperature variations because of its naive thermal-compensation mechanism.

**4.3.3.2 Effectiveness of Global Wavelength Re-allocation** Figure 36 demonstrate the effect of global wavelength re-allocation on the available bandwidth. Since Nominal and Closest provide similar performance in Figure 35, we do not show the results of nominal and BandArb_nominal in the rest of this section.

By comparing the curves for WM vs BA_WM and closest vs BA_closest, the bandwidth enhancement provided by the global re-allocation of wavelengths becomes obvious no matter which local wavelength realignment algorithm is applied. Specifically, global re-allocation improves the bandwidth significantly and results in small performance sensitivity to temper-

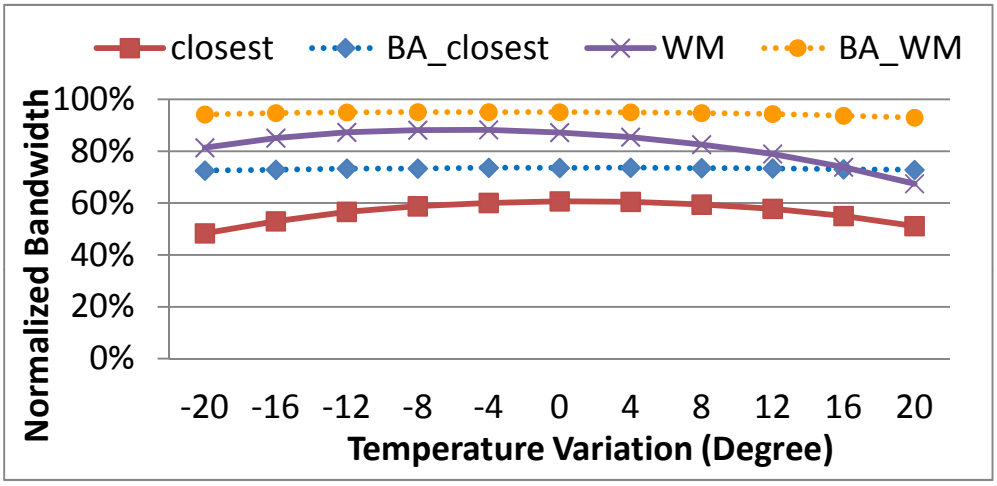



Figure 36: Network bandwidth Vs. temperature variations for CG-BandArb normalized to the bandwidth in the absence of PV and TF.

ature variations. CG-BandArb with WM as a local alignment scheme recovers 95% and 93% of the bandwidth when the range of temperature variation is 0 and 20 degrees, respectively.

### 4.3.4 Evaluation of Tuning Power and Computation Latency of Re-alignment

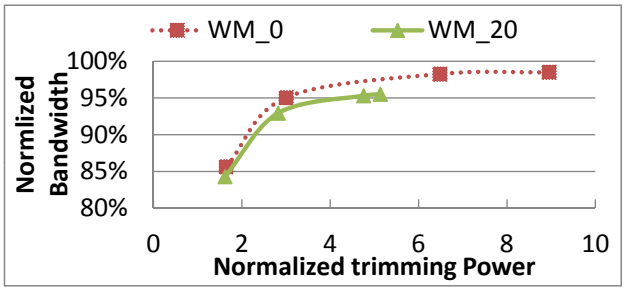



Figure 37: Trimming Power VS Normalized Bandwidth.

To tune a $\mu$ring to the target resonance, extra static power is consumed with 0.13mw/nm for current injection [2] and 0.24mw/nm for heating [45]. Since only local wavelength alignment modifies the trimming distance of each $\mu$ring, the tuning power does not change by

applying global re-allocation of wavelengths. Tuning power also is related to the trimming constraint. A relaxed limit on trimming leads to strong correction ability but leads to larger trimming distance, and thus, larger trimming power. Hence, we studied the design tradeoff between the trimming power and achievable bandwidth of the proposed WM by varying the trimming constraint on heating form $1\Delta\lambda$ to $4\Delta\lambda$. In In Figure 37 results are given when no TF is modeled during the simulation (WM_0) and when the temperature of each node is randomly generated within a 20 degree range (WM_20). As shown in the figure, relaxing the trimming distance incurs additional power, but the improvement of the available bandwidth when the trimming distance is larger than $2\Delta\lambda$ is greatly reduced. This is why we use a limit of $2\Delta\lambda$ on the trimming range in our simulations.

Table 6: Computation Time of Different algorithms

| | |
|---|---|
| ILP | 13.2 s |
| WM | 24.03$\mu$s |
| BandArb_closest | 32.84 $\mu$s |
| BandArb_WM | 56.88 $\mu$s |

Table 6 lists the computation time of the proposed wavelength mapping algorithms running at Intel© Xeon© server with 2.50GHz CPU and 16GB memory. Note that closest, nominal and (MinTrim+SRW) can be implemented with electric circuits that incurs a small area overhead, thus, producing the mapping results in several ns. ILP can deliver the optimum bandwidth results among all the evaluated mapping schemes, but it also requires a large execution time which cannot keep up with the rate of possible TF. The computation time is reduced significantly with WM and BandArb_WM which makes them suitable for run-time calculation, especially that the span of TF is in the order of millisecond at worse.

### 4.3.5 Evaluation of Network Connectivity

Another important metric to evaluate optical networks is the yield indicated by the number of node pairs losing connectivity due to PV or TF, which is possible because of the limit

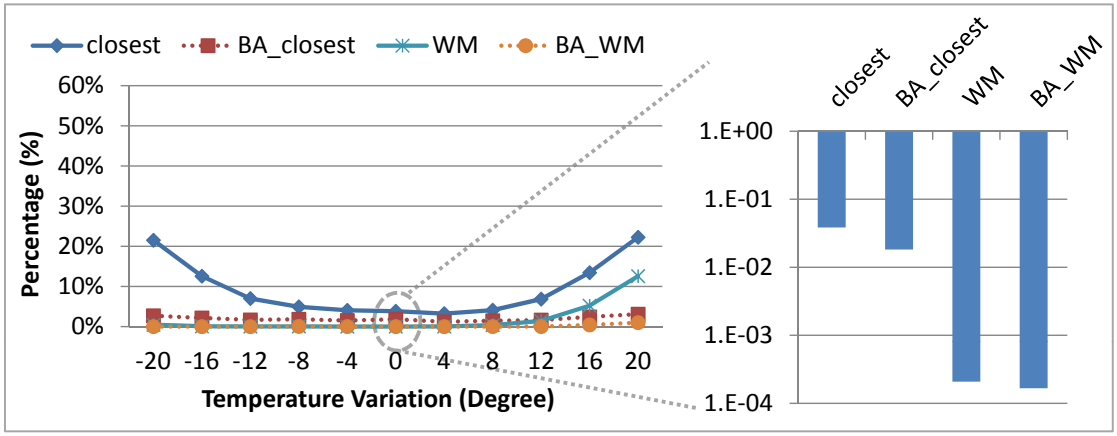



Figure 38: Probability of losing connectivity.

imposed on the wavelength correction range of $\mu$rings. Specifically, connectivity is lost when a sender and a receiver do not have any common wavelengths to communicate with each other. Figure 38 shows such probability for different alignment heuristics. The left figure shows the probability distribution as a function of the range of temperature variation, while the right figure amplifies the probabilities when no TF is present. The data are measured with the normalized number of such failure pairs for all of the 100 samples, each having 16×15 node pairs.

We can observe from Figure 38 that the naive local wavelength alignment, nominal, may result in high failure rate of the optical network, especially with large TF. WM is able to improve the robustness of the network by reducing the probability from $10^{-1} \sim 10^{-2}$ to $10^{-3} \sim 10^{-4}$. Local wavelength alignments are illustrated with solid lines and global CG-BandArb approaches are denoted with dotted lines. The figure shows that CG-BandArb reduces significantly the rate of disconnection. In addition, CG-BandArb-based schemes are demonstrated to be effective in dealing with TF since their probability of losing connectivity is stable with increasing temperature variations. FG-BandArb is not evaluated here because the connectivity is affected by the network traffics and varies significantly during the entire execution.

If at run-time, TF causes any pair of nodes to be disconnected, then a thermal management unit on the chip can halt execution until the chip cools down. Alternatively, a chip may be considered defective after fabrication if any pair of nodes loses connectivity due to PV alone and any receiver or sender is estimated to lose 4 or more wavelengths with a 40 degree temperature variation. In the experiments presented next, we only use non-defective chips (according to the above definition) from the 100 generated die samples.

### 4.3.6 Evaluations Using Traffic Traces

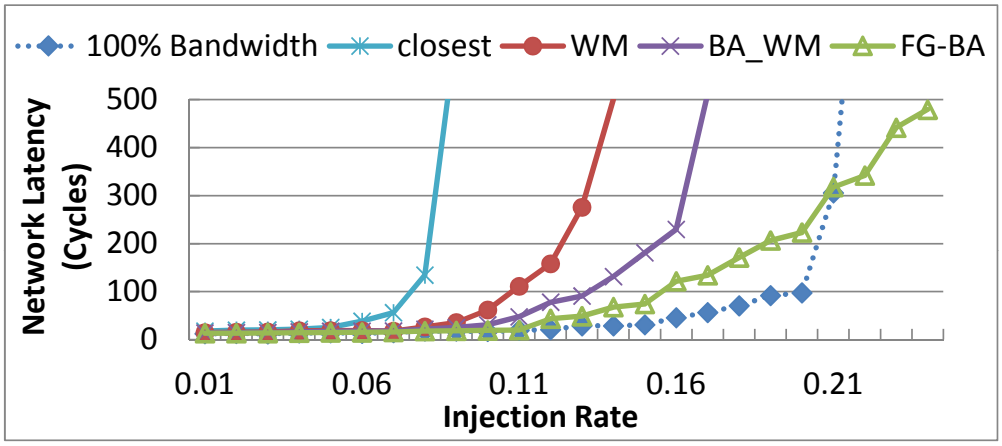



Figure 39: Network Latency under Uniform Random traffics.

**4.3.6.1 Synthetic Traffic Traces** FG-BandArb (FG-BA) allocates the wavelengths to the nodes dynamically. Figure 39 shows the network latency curve for a network that always has the full bandwidth available for communication (100% Bandwidth) and compares this latency with the latencies resulting from applying the bandwidth re-mapping algorithms. In this experiment, the temperature of each node is randomly varied within the range of 40 degree every 1ms. Although CG-BandArb(BA_WM) is able to recover most of the bandwidth by re-aligning wavelengths to the $\mu$rings and nodes, it still results in a much earlier saturation point compared to the ideal scenario. The reason is that even losing a small number of wavelengths can cause significant increase in network latency. For example, in the ideal case, transiting one packet takes 1 cycle. However, if only one $\lambda$ is lost, it takes 2 cycles.

Clearly, FG-BandArb is able to postpone the saturation point. In addition, with the ability to leverage the wavelengths made available by the thermal $\mu$rings, the network latency could be even lower than the "100% Bandwidth" case.

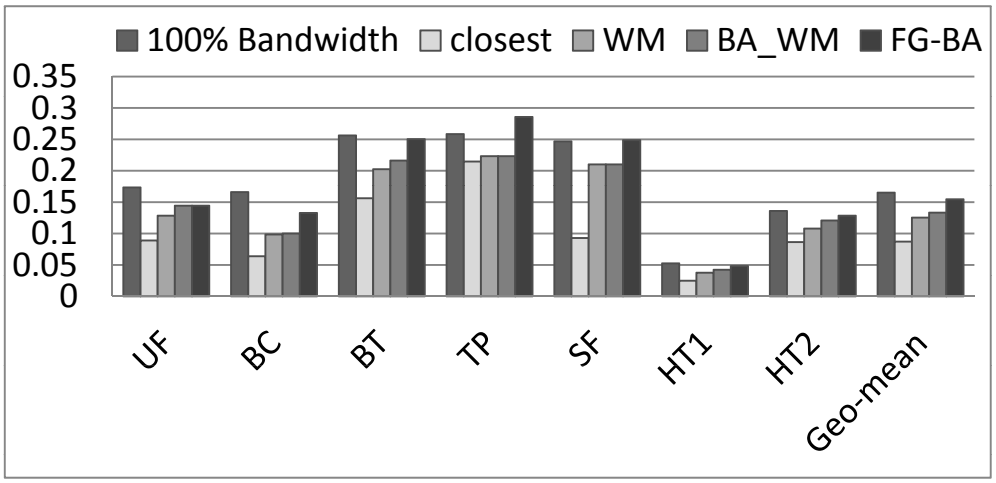



Figure 40: Network Throughput with Synthetic Traffic Trace.

Figure 40 compares the network throughput of the ideal case, WM and FG-BandArb for different temperature variations, measured via the receiving rate of each node. FG-BandArb improves the throughput by utilizing the network channels more effectively than the coarse-grained approach across all the synthetic traces used in the section. In summary, the proposed wavelength re-allocation at three different levels provides an effective method to salvage network bandwidth under PV and TF.

**4.3.6.2 PARSEC and SPEC CPU 2006 Benchmarks** We also evaluated the efficiency of BandArb using thermal traces generated by running a mixed multiprogrammed workload applications from the PARSEC/SPEC CPU benchmark suites (see table 7). In these experiments, the time step between two consecutive temperature measurements is set at 1ms. As an example, Figure 41 shows the thermal traces of 6 of the 16 network nodes generated for the workload Mix-1. As we can observe, the temperature is different at the different nodes and fluctuates in time, as well. Figure 42 shows the average bandwidth available to each communicating node pair when the effect of the thermal traces are considered.

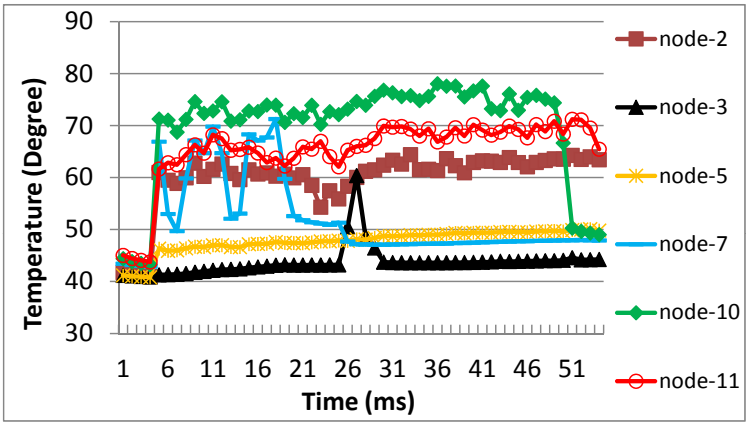
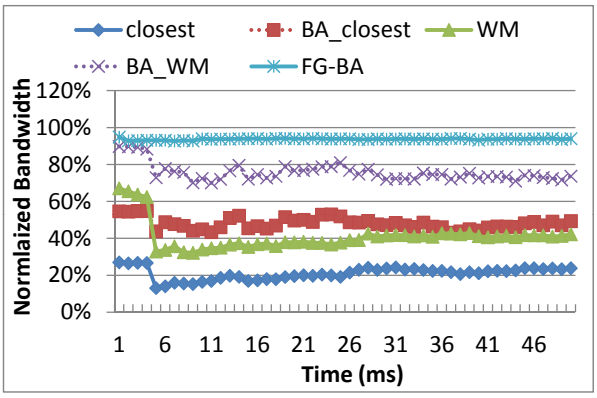



Figure 41: An example thermal trace of multi-programming benchmarks.

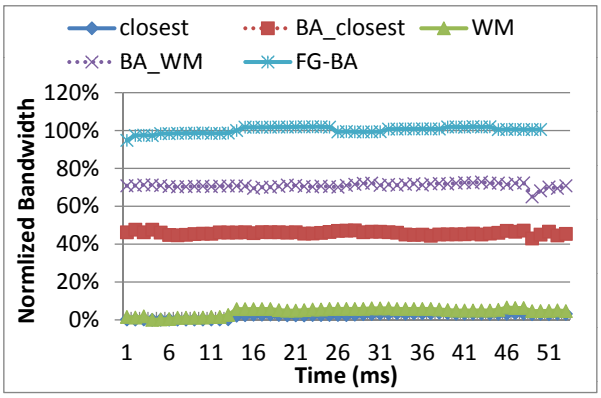
We observe that FG-BandArb outperforms all the evaluated mapping algorithms when real thermal traces are used. It is able to provide near-full or even above 100% bandwidth. From Figure 42, we also observe that global wavelength re-allocation allows the network to handle TF better than the local algorithms. Moreover, FG-BandArb leverages the improvements of global and local wavelength alignment algorithms and improves the bandwidth by and additional 27%, on average.

Table 7: Multiprogrammed workloads

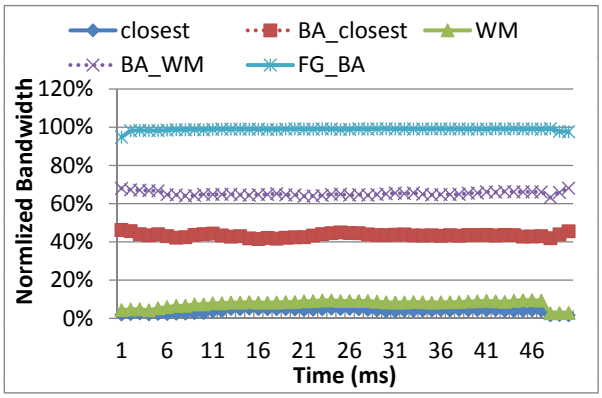
| | |
|---|---|
| Mix-1 | Blacksholes, bodytrack, canneal, dedup |
| Mix-2 | Ferret, fluidanimate, streamcluster, rtview |
| Mix-3 | Leslie3d, libquantum, namd, sjeng |
| Mix-4 | Leslie3d, cactusADM, libquantum, bzip |



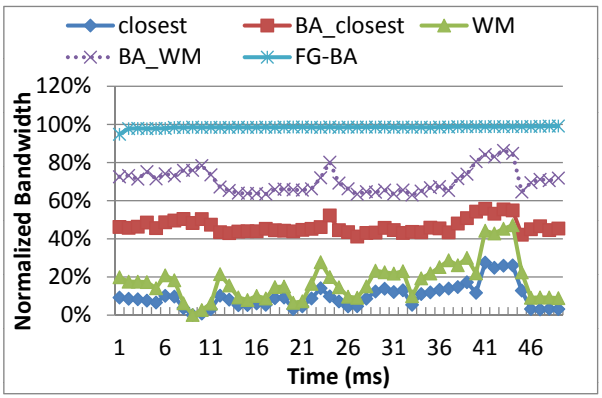

(a) Mix-1




(b) Mix-2




(c) Mix-3




(d) Mix-4

Figure 42: Normalized available bandwidth for communication with multi-programmed workloads.

## 4.4 SUMMARY

With process variation and thermal shifts, it is not realistic to assume that optical networks are perfect. Hence, future optical networks should be designed to adapt to bandwidth loss or changes. This chapter advocates that re-allocating the wavelengths to $\mu$rings and nodes according to temperature deviations, PV and traffic conditions is effective in maintaining the network bandwidth and performance. Two proposed heuristic mechanisms, WM and BandArb that re-align the wavelength, respectively, within local nodes and across the entire network upon temperature variation recover 95% of the optical bandwidth. A fine grained arbitration scheme that allocate the wavelengths to the active nodes every cycle based on the traffic status can further improve the network utilization and increase the network throughput by up to 18% and reduce the network loss by 27% on average over CG-BandArb. Therefore, the proposed BandArb technique is effective in mitigating the impacts of both process and thermal variations from an architecture perspective.

In the future, our goal will be to reduce the overhead of wavelength re-allocation schemes through simplifying CG-BandArb so that it can be realized using hardware-based solutions. Also, given that most of the wavelength sharing in FG-BandArb happens among neighboring nodes, it should be possible to design simpler arbitration schemes that remove the need for broadcasting requests. We will also investigate fine-grained wavelength allocation schemes for optical networks that already require flit-level arbitration, such as MWSR, MWMR.

# 5.0 CONCLUSIONS AND FUTURE WORK

This chapter is dedicated to distilling previous chapters and discussing future work.

## 5.1 TOWARDS RELIABLE NANOPHOTONIC INTEROCNNECTION NETWORK DESIGNS

Chip multiprocessors (CMP) have emerged as a promising microarchitecture for keeping up performance with integration density [21, 47]. Today, the number of on-chip cores has reached low hundreds, e.g. Intel's 80-core Terascale chip [66] and nVidia's 128-core Quadro GPU [46] and will be likely to reach upper hundreds or even a thousand [12] in the near future. The importance of on-chip network grow together with the size of those CMPs, in order to meet the performance requirements.

However, ITRS [23] identified limitations in using metal wires for global links: (i) the wire performance does not scale well; (ii) long RC wires require large number of repeaters that consume significant portion of total power; and (iii) the slow increase of pin count restricts the bandwidth between core and memory. In contrast, nanophotonic links can provide high bandwidth density, low propagation delay, communication-target-independent power consumption, and natural support for multicast/broadcast. Recent advances in photonic devices and integration technology have made it a promising candidate for future global interconnects.

Unfortunately, while optical interconnect provides many attractive and promising features, there are also fundamental challenges in integration and fabrication of those devices to providing robust and reliable on-chip communication.Among many challenges, the ther-

mal sensitivity and process variation of silicon photonic devices are the two most important and difficult hurdles. Studies have shown that naïve solutions could overturn the benefits of putting optics on-chip [17, 39, 45].

Due to temperature fluctuations, optical components fail to resonate designated wavelengths in the waveguide since their resonances all drift by several channels with temperature variations. Consequently, either transmitter or receiver or both cannot utilize all available wavelengths/bandwidth to send/receive data from the optical interconnect. Another major reason for refractive index change is PV. Variations of critical physical dimensions, e.g. thickness of wafer, width of waveguide, caused by lithography imperfection and etch non-uniformity of devices are inevitable. As a result, not all fabricated $\mu$rings can be used due to process variations, leading to wavelength/bandwidth loss in communication.

Current solutions to the resonance frequency shifting problem are either impractical or too preliminary and severely limited. For shifts caused by PV, post fabrication techniques have been experimented with to trim the resonance frequency of a $\mu$ring using high energy particles such as UV light or electron beam. However, given that the number of $\mu$rings on-chip is on the order of thousands to millions [2, 26, 30, 51, 69], tuning $\mu$ring one by one is impractical. Another well-known solution to the resonance frequency shifting problem is heating and current injection [2, 17, 39, 45]. But it can result in significant power consumption and has limited correction ability. Hence, there is currently no practical and economical solution to this problem. It is thus unrealistic to assume that the optical network is always perfect. I believe it is time for computer architects to start thinking about improving the reliability of optical interconnects.

In this thesis, I present our contributions on this field to make one step further towards adopting optics on-chip. The goal is to tackle the bandwidth loss problem that arise from fabrication error and runtime on-chip temperature fluctuation in an optical interconnection as high bandwidth density has always been a major advantage of optical networks over electric networks.

In Chapter 3, I propose a serial of approaches, named "MinTrim" which uses ILP to reorganize the arrangement among $\mu$rings and wavelengths, adds supplementary $\mu$rings and allows flexible assignment of wavelengths to network nodes as long as the resulting network

presents maximal bandwidth. Each step is shown to improve bandwidth provisioning with lower power requirement. Evaluations on a sample network show that a baseline network could lose more than 40% bandwidth due to PV. Such loss can be recovered by MinTrim to produce a network with 98.4% working bandwidth. In addition, the power required in arranging $\mu$rings is 39% lower than the baseline. Therefore, MinTrim provides an efficient PV-tolerant solution to improving the reliability of on-chip photonics.

In Chapter 4, new techniques "BandArb" is presented with a focus on reducing the computation cost to be applicable at runtime so that dynamic variations are addressed. Since temperature changes slowly, BandArb first re-assigns the $\mu$rings to resonant wavelengths in each network node to tackle both static (PV) and dynamic (TF) variations at a coarse granularity. Then, based on the observation that the load on the network is often not balanced, fine-grained wavelength arbitration is used to allow a transmitting node to borrow idle wavelengths from other nodes to maximize bandwidth utilization. The evaluations with both synthetic traces and SPEC2006/PARSEC benchmarks shows that the proposed two-level scheme can effectively mitigates the impacts of both PV and thermal variations.

## 5.2 FUTURE WORK

Besides the solutions presented in this thesis, there are still plenty of interesting problems that need to be solved in this exciting area. I listed a couple of topics in the following sections.

### 5.2.1 Improving Connectivity of Photonic Network

As discussed in section 3.4 and section 4.3, even with MinTrim and BandArb, there is still a small chance for two nodes in the network lose connection due to the effects of PV and TF. Once disconnection happens, the chip is no longer functional or has to stop computation until the temperature variation become smaller. Furthermore, the $\lambda$ drifts dynamically with the chip temperature, which makes it difficult to determine whether the network is failed

or not during the post-fabrication test. To improve the yield and reliability of the optical network, we could use a fault-tolerant routing algorithm to handle static and dynamic link failures.

In an all-to-all optical network, each node has a direct link to access any other node. If the link bandwidth is degraded to zero, the packet must re-route to other intermediate nodes before it reaches the destination. However, the new routing path may create deadlocks in the network, which can be avoided by using different sets of virtual channels (VCs). Hence the fault-tolerance design introduces many open questions remaining to be solved, such as how could we determine the routing path, how many VCs are necessary to setup a path if there exists a connection between two nodes, etc.

### 5.2.2 Extending BandArb to Other Crossbar Designs

MWMR and MWSR share the same behavior in the transmission side: multiple senders contend for the same set of wavelengths, defined as network link. In the baseline design, the arbitration which is already in place allows only one node to send data via the link. However, due to PV or TF drifts, it is possible that the link bandwidth can not be fully used by the arbitration winner, leading to the degradation of network throughput. While other competitors that are not allowed to use the link simultaneously can actually leverage the remaining wavelengths. Hence, we should revisit the flit-level arbitration and transmission protocol implemented in MWMR/MWSR and develop a more efficient method to fully utilize the link bandwidth and improve the throughput.

### 5.2.3 Reliable Off-Chip Optical Network Designs

For future large-scale CMP, memory latency, energy, capacity and bandwidth will be performance bottlenecks. [64] proposed a memory architecture, as shown in Figure 43 that leverages the emerging 3D stacking technology [11, 28, 37, 42, 71, 77] to improve the memory capacity by stacking multiple memory dies on top of each other and enable fast intra-chip data transmission with short and high dense Through-Silicon-Vias (TSVs). To break the pin barrier and reduce the power consumption of I/O pins, they also adopted the silicon-photonic

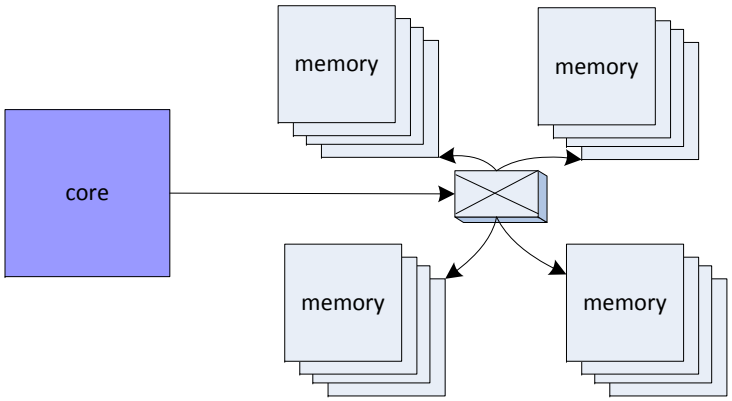



Figure 43: An example of future memory system architecture [64].

technology to connect the processor cores and multiple memory stacks. It is interesting to explore the optical network topology, network interface and memory access protocol designs for the off-chip memory traffics. Last but not least, the infrastructure should include an architecture-level approach to maintain the reliability of the proposed optical network.

## BIBLIOGRAPHY


[1] Spec cpu2006. http://www.spec.org/cpu2006.

[2] J. Ahn, M. Fiorentino, R. G. Beausoleil, N. Binkert, A. Davis, D. Fattal, N. P. Jouppi, M. McLaren, C. M. Santori, R. S. Schreiber, S. M. Spillane, D. Vantrease, and Q. Xu. Devices and architectures for photonic chip-scale integration. *Appl. Phy. A*, 2009.

[3] J. Balfour and W. Dally. Design tradeoffs for tiled cmp on-chip networks. In *the 20th International Conference on Supercomputing*, pages 187–198, 2006.

[4] C. Batten. Designing nanophotonic interconnection networks. In *Workshop on the Interaction between Nanophotonic Devices and Systems*, 2010.

[5] M. Beals, J. Michel, J. F. Liu, D. H. Ahn, D. Sparacin, R. Sun, C. Y. Hong, L. C. Kimerling, A. Pomerene, D. Carothers, J. Beattie, A. Kopa, A. Apsel, M. S. Rasras, D. M. Gill, S. S. Patel, K. Y. Tu, Y. K. Chen, and A. E. White. Process flow innovations for photonic device integration in cmos. In *In Proc. of the International Society for Optical Engineering (SPIE)*, pages 689804–689804–14, 2008.

[6] S. Beamer, C. Sun, Y.-J. Kwon, A. Joshi, C. Batten, V. Stojanović, and K. Asanović. Re-architecting dram memory systems with monolithically integrated silicon photonics. In *ISCA '10*, pages 129–140, 2010.

[7] M. Berkelaar et al. Lp solve: Opern source (mixed-integer) linear programming system (2007).

[8] C. Bienia et al. The parsec benchmark suite: Characterization and architectural implications. In *PACT*, pages 72–81, 2008.

[9] 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, M. D., and D. A. Wood. The gem5 simulator. 39:1–7, May 2011.

[10] N. L. Binkert, A. Davis, N. P. Jouppi, M. McLaren, N. Muralimanohar, R. Schreiber, and J. H. Ahn. The role of optics in future high radix switch design. In *ISCA*, pages 437–448, 2011.

[11] B. Black, M. Annavaram, N. Brekelbaum, J. DeVale, L. Jiang, G. H. Loh, D. McCaule, P. Morrow, D. W. Nelson, D. Pantuso, P. Reed, J. Rupley, S. Shankar, J. Shen, and C. Webb. Die stacking (3d) microarchitecture. In *Proceedings of the 39th International Symposium on Microarchitecture*, pages 469–479, 2006.

[12] S. Borkar. Thousand core chips - a technology perspective. In *DAC*, pages 746–749, 2007.

[13] Y. Cao and L. T. Clark. Mapping statistical process variations toward circuit performance variability: An analytical modeling approach. In *Design Automation Conference*, pages 658–663, 2005.

[14] G. Chen, H. Chen, M. Haurylau, N. Nelson, P. M. Fauchet, E. Friedman, and D. Albonesi. Predictions of cmos compatible on-chip optical interconnect. In *In International Workshop on System-Level Interconnect Prediction*, page 13C20, Apr. 2005.

[15] V. A. Christopher Nitta, Matthew K. Farrens. Reilient microring resonator based photonic networks. 2011.

[16] M. J. Cianchetti, J. C. Kerekes, and D. H. Albonesi. Phastlane: a rapid transit optical routing network. In *Proceedings of the 36th annual international symposium on Computer architecture*, ISCA '09, pages 441–450, 2009.

[17] R. K. Dokania and A. B. Apsel. Analysis of challenges for on-chip optical interconnects. In *GLSVLSI '09*, 2009.

[18] R. Fourer, D. M. Gay, and B. W. Kernighan. *AMPL: A Modeling Language for Mathematical Programming, 2nd ed.* Duxbury Press Publishing Company, 2002.

[19] M. Georgas, J. Leu, B. Moss, C. Sun, and V. Stojanovic. Addressing link-level design tradeoffs for integrated photonic interconnects. In *Custom Integrated Circuits Conference (CICC), 2011 IEEE*, pages 1 –8, sept. 2011.

[20] H. Haeiwa, T. Naganawa, and Y. Kokubun. Wide range center wavelength trimming of vertically coupled microring resonator filter by direct uv irradiation to sin ring core. *IEEE Photonics Technology Letters*, 16:135–137, 2004.

[21] L. Hammond, B. A. Nayfeh, and K. Olukotun. A single-chip multiprocessor. *Computer*, 30:79–85, September 1997.

[22] R. Ho, W. Mai, and M. A. Horowitz. The future of wires. *Proceedings of the IEEE*, 89:490C504, April 2001.

[23] ITRS. International technology roadmap for semiconductors. Technical report, 2010.

[24] W. D. J. Kim, J. Balfour. Flattened butterfly topology for on-chip networks. In *Proceedings of the 40th Annual IEEE/ACM International Symposium on Microarchitecture, MICRO 40*, pages 172–182, 2007.

[25] A. Joshi, C. Batten, Y.-J. Kwon, S. Beamer, I. Shamim, K. Asanovic, and V. Stojanovic. Silicon-photonic clos networks for global on-chip communication. In *3rd ACM/IEEE International Symposium on Networks-on-Chip*, pages 124 –133, may 2009.

[26] A. Joshi, J. Orcutt, A. Khilo, B. Moss, C. Holzwarth, M. Popovic, H. Li, H. Smith, J. Hoyt, F. Kartner, R. Ram, V. Stojanovic, and K. Asanovic. Building manycore processor-to-dram networks with monolithic silicon photonics. In *Hot Interconnects*, pages 21–30, 2008.

[27] J. Karttunen, J. Kiihamäki, and S. Franssila. Loading effects in deep silicon etching. In *International Society of Optical Engineering*, volume 4174, pages 90–97, 2000.

[28] T. Kgil, S. D'Souza, A. Saidi, N. Binkert, R. Dreslinski, S. Reinhardt, K. Flautner, and T. Mudge. Picoserver: Using 3d stacking technology to enable a compact energy efficient chip multiprocessor. In *Proceedings of the 12th International Conference on Architecture Support for Programming Languages & Operating Systems*, pages 117–128, 2006.

[29] N. Kirman, M. Kirman, R. K. Dokania, J. F. Martinez, A. B. Apsel, M. A. Watkins, and D. H. Albonesi. Leveraging optical technology in future bus-based chip multiprocessors. In *Proceedings of the 39th Annual IEEE/ACM International Symposium on Microarchitecture*, pages 492–503, 2006.

[30] N. Kirman and J. F. Martínez. A power-efficient all-optical on-chip interconnect using wavelength-based oblivious routing. In *ASPLOS '10*, pages 15–28, 2010.

[31] B. R. Koch, A. W. Fang, O. Cohen, and J. E. Bowers. Mode-locked silicon evanescent lasers. *Optics Express*, 15(18), 2007.

[32] P. Koka, M. O. McCracken, H. Schwetman, X. Zheng, R. Ho, and A. V. Krishnamoorthy. Silicon-photonic network architectures for scalable, power-efficient multi-chip systems. *SIGARCH Comput. Archit. News*, 38, 2010.

[33] P. Koka, M. O. Michael, S. Herb, C.-H. O. Chen, X. Zheng, H. Ron, R. Kannan, and V. K. Ashok. A micro-architectural analysis of switched photonic multi-chip interconnects. In *proceedings of 39th Annual International Symposium on Computer Architecture, ISCA'12*, pages 153–164, 2012.

[34] Y. Kokubun, N. Kobayashi, and T. Sato. Uv trimming of polarization-independent microring resonator by internal stress and temperature control. *Optics Express*, 18:906–916, 2010.

[35] R. Kumar, V. Zyuban, and D. M. Tullsen. Interconnections in multi-core architectures: Understanding mechanisms, overheads and scaling. In *In Proc. of the Int. Symp. on Computer Architecture (ISCA)*, pages 408–419, 2005.

[36] G. Kurian, J. E. Miller, J. Psota, J. Eastep, J. Liu, J. Michel, L. C. Kimerling, and A. Agarwal. Atac: a 1000-core cache-coherent processor with on-chip optical network. In *Proceedings of the 19th international conference on Parallel architectures and compilation techniques*, pages 477–488, 2010.

[37] F. Li, C. Nicopoulos, T. Richardson, Y. Xie, V. Narayanan, and M. Kandemir. Design and management of 3d chip multiprocessors using network-in-memory. In *Proceedings of the 33rd Annual International Symposium on Computer Architecture*, pages 130–141, 2006.

[38] S. Li, J. H. Ahn, R. D. Strong, J. B. Brockman, D. M. Tullsen, and N. P. Jouppi. Mcpat: an integrated power, area, and timing modeling framework for multicore and manycore architectures. In *Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture*, MICRO 42, pages 469–480, 2009.

[39] Z. Li, M. Mohamed, X. Chen, E. Dudley, K. Meng, L. Shang, A. R. Mickelson, R. Joseph, M. Vachharajani, B. Schwartz, and Y. Sun. Reliability modeling and management of nanophotonic on-chip networks. *IEEE Transactions on Very Large Scale Integration Systems*, pages 98–111, 2010.

[40] A. Liu, R. Jones, L. Liao, D. Samara-Rubio, D. Rubin, O. Cohen, R. Nicolaescu, and M. Paniccia. A high-speed silicon optical modulator based on a metal-oxdie-semiconductor capacitor. *Nature*, 427:615–618, 2004.

[41] J. F. Liu and J. Michel. High performance ge devices for electronic-photonic integrated circuits. 16:575–582, 2008.

[42] G. H. Loh. 3d-stacked memory architecture for multi-core processors. In *the 35th Annual International Symposium on Computer Architecture*, pages 453–464, 2008.

[43] S. R. Nassif. Modeling and forecasting of manufacturing variations. In *Asia and South Pacific Design Automation Conference*, pages 145–149, 2001.

[44] Y. Nasu, M. Kohtoku, M. Abe, and Y. Hibino. Birefringence suppression of uv-induced refractive index with grooves in silica-based planar lightwave circuits. *Electronics Letters*, 41:1118–1119, 2005.

[45] C. Nitta, M. Farrens, and V. Akella. Addressing system-level trimming issues in on-chip nanophotonic networks. In *IEEE 17th International Symposium on High Performance Computer Architecture (HPCA)*, pages 122–131, 2011.

[46] nVidia. Quadro fx 3700m. http://www.nvidia.com/object/product_quadro_fx_3700_m_us.html.

[47] K. Olukotun, B. Nayfeh, L. Hammond, K. Wilson, and K.-Y. Chang. The case for a single-chip multiprocessor. In *the $7^{th}$ International Symposium on Architectural Support for Programming Lanuages and Operating Systems*, Oct. 1996.

[48] J. S. Orcutt, A. Khilo, C. W. Holzwarth, M. A. Popovic, H. Li, J. Sun, T. Bonifield, R. Hollingsworth, F. X. Kartner, H. I. Smith, V. Stojanovic, and R. J. Ram. Nanophotonic integration in state-of-the-art cmos foundries. *Optics Express*, 19, 2011.

[49] M. Palesi et al. Noxim, an open network-on-chip simulator. http://noxim.sourceforge.net.

[50] Y. Pan, J. Kim, and G. Memik. Flexishare: Channel sharing for an energy-efficient nanophotonic crossbar. In *High-Performance Computer Architecture*, pages 1–12, 2010.

[51] Y. Pan, P. Kumar, J. Kim, G. Memik, Y. Zhang, and A. Choudhary. Firefly: Illuminating future network-on-chip with nanophotonics. In *Proceedings of the 36th annual international symposium on Computer architecture*, 2009.

[52] S. Postnikov, S. Hector, C. Garza, R. Peters, and V. Ivin. Critical dimension control in optical lithography. *Microelectronic Engineering*, 69(2-4):452–458, 2003.

[53] S. P. Qianfan Xu, Bradley Schmidt and M. Lipson. Micrometre-scale silicon electro-optic modulator. *Nature*, pages 325–327, 2005.

[54] C. Qiu and Q. Xu. Wavelength tracking with thermally controlled silicon resonators. *Optics Express*, 19, 2011.

[55] G. T. Reed, G. Mashanovich, F. Y. Gardes, and D. J. Thomson. Silicon optical modulators. *Nature Photonics*, 4:518 – 526, 2010.

[56] S. R. Sarangi, B. Greskamp, R. Teodorescu, J. Nakano, A. Tiwari, and J. Torrellas. Varius: A model of process variation and resulting timing errors for microarchitects. *Semiconductor Manufacturing, IEEE Transactions on*, 21(1):3 –13, feb. 2008.

[57] C. Schow. Optical interconnects in next-generation high-performance computers. In *OIDA 2008 Integration Forum*, 2008.

[58] J. Schrauwen, D. V. Thourhout, and R. Baets. Trimming of silicon ring resonator by electron beam induced compaction and strain. *Optics Express*, 16:3738–3743, 2008.

[59] S. K. Selvaraja. *Wafer-Scale Fabrication Technology for Silicon Photonic Integrated Circuits*. PhD thesis, Ghent University, Feb 2011.

[60] A. Shacham, K. Bergman, and L. P. Carloni. Photonic networks-on-chip for future generations of chip multiprocessors. 57:1246–1260, Sep. 2008.

[61] N. Sherwood-Droz, K. Preston, J. Levy, and M. Lipson. Device guidelines for wdm interconnects using silicon microring resonators. In *Workshop on the Interaction between Nanophotonic Devices and Systems*, 2010.

[62] K. Skadron, M. Stan, W. Huang, S. Velusamy, K. Sankaranarayanan, and D. Tarjan. Temperature-aware microarchitecture. In *ACM SIGARCH Computer Architecture News*, volume 31, pages 2–13. ACM, 2003.

[63] J. Tatum. Vcsels for 10 gb/s optical interconnects. In *In IEEE Emerging Technologies Symposium on BroadBand Communications for the Internet Era*, page 58C61, Sep. 2001.

[64] A. N. Udipi, N. Muralimanohar, R. Balasubramonian, A. Davis, and N. P. Jouppi. Combining memory and a controller with photonics through 3D-stacking to enable scalable and energy-efficient systems. In *Proceedings of the 38th annual international symposium on Computer architecture*, pages 425–436, jun 2011.

[65] S. Ueno, T. Naganawa, and Y. Kokubun. High uv sensitivity of sion film and its application to center wavelength trimming of microring resonator filter. *IEICE Transactions on Electron*, E88-C(5):998–1004, 2005.

[66] S. Vangal, J. Howard, G. Ruhl, S. Dighe, H. Wilson, J. Tschanz, D. Finan, P. Iyer, A. Singh, T. Jacob, S. Jain, S. Venkataraman, Y. Hoskote, and N. Borkar. An 80-tile 1.28tflops network-on-chip in 65nm cmos. In *In IEEE Int. Solid-State Circuits Conf.*, pages 98–590, 2007.

[67] D. Vantrease, N. Binkert, R. Schreiber, and M. Lipasti. Light speed arbitration and flow control for nanophotonic interconnects. In *Proceedings of the 42 Annual IEEE/ACM International Symposium on Microarchitecture*, pages 304–315, dec. 2009.

[68] D. Vantrease, M. H. Lipasti, and N. Binkert. Atomic coherence: Leveraging nanophotonics to build race-free cache coherence protocols. In *Proceedings of the 2011 IEEE 17th International Symposium on High Performance Computer Architecture*, pages 132 –143, 2011.

[69] D. Vantrease, R. Schreiber, M. Monchiero, M. McLaren, N. Jouppi, M. Fiorentino, A. Davis, N. Binkert, R. Beausoleil, and J. Ahn. Corona: System implications of emerging nanophotonic technology. In *35th International Symposium on Computer Architecture*, pages 153–164, 2008.

[70] K. Williams and P. Watts. Optical interconnects for nocs and off-chip communications. Tutorial, Intl. Symp. on Networks-on-Chip, 2011.

[71] D. H. Woo, N. H. Seong, D. L. Lewis, and H. S. Lee. An optimized 3d-stacked memory architecture by exploiting excessive, high-density tsv bandwidth. In *HPCA*, 2010.

[72] Y. Xie, M. Nikdast, J. Xu, W. Zhang, Q. Li, X. Wu, Y. Ye, X. Wang, and W. Liu. Crosstalk noise and bit error rate analysis for optical network-on-chip. In *in Design Automation Conference*, 2010.

[73] D. Xu. Polarization control in silicon photonic. *Topics in Applied Physics*, pages 31–70, 2011.

[74] Q. Xu, D. Fattal, and R. G. Beausoleil. Silicon microring resonators with 1.5-$\mu$m radius. *Optics Express*, pages 4309–4315, 2008.

[75] Q. Xu, S. Manipatruni, B. Schmidt, J. Shaky, and M. Lipson. 12.5 gbit/s carrier-injection-based silicon micro-ring silicon modulators. *Optics Express*, pages 430–436, 2007.

[76] Y. Xu, Y. Du, Y. Zhang, and J. Yang. A composite and scalable cache coherence protocol for large scale cmps. In *International Conference on Supercomputing*, pages 285–294, 2011.

[77] Y. Xu, Y. Du, B. Zhao, X. Zhou, Y. Zhang, and J. Yang. A low-radix and low-diameter 3d interconnection network design. In *HPCA*, pages 30–41, 2009.

[78] Y. Xu, J. Yang, and R. Melhem. Channel borrowing: an energy-efficient nanophotonic crossbar architecture with light-weight arbitration. In *Proceedings of the 26th ACM international conference on Supercomputing*, ICS '12, pages 133–142, 2012.

[79] Y. Xu, J. Yang, and R. Melhem. Tolerating process variations in nanophotonic on-chip networks. In *proceedings of 39th Annual International Symposium on Computer Architecture*, ISCA'12, pages 142 –152, jun 2012.

[80] S. Yoo. Cmos-compatible silicon photonic integrated systems in future computing and communication systems. In *OptoeElectronics and Communications Conference (OECC), 2010 15th*, 2010.