Efficient Architectures for Multichip Communication

By

Mohammad Arslan Zulfiqar

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Electrical and Computer Engineering)

at the

UNIVERSITY OF WISCONSIN–MADISON

2014

Date of final oral examination: 08/05/14

The dissertation is approved by the following members of the Final Oral Committee:

Mikko H. Lipasti, Professor, Electrical and Computer Engineering
Parameswaran Ramanathan, Professor, Electrical and Computer Engineering
Mark D. Hill, Professor, Computer Sciences
Gurindar S. Sohi, Professor, Computer Sciences
Herbert D. Schwetman, Member of Technical Staff, Oracle Labs

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

This thesis is dedicated to my parents, Mohammad Zulfiqar and Rehana Zulfiqar, for their love, support and encouragement.

ACKNOWLEDGMENTS

I owe thanks and appreciation to many people who have supported me throughout my doctoral studies. I would like to take this opportunity to thank them.

I would first like to express my deepest gratitude to my advisor Prof. Mikko Lipasti for his mentor-ship, guidance and thoughtful advice throughout my graduate studies. This thesis would not have been accomplished without his technical support and guidance. His optimism and encouragement have been instrumental in keeping me positive throughout my research. I am thankful to him for letting me conduct research in his lab.

I am also thankful to my dissertation committee members, Prof. Mark Hill, Prof. Guri Sohi, Prof. Parmesh Ramanathan, and Dr. Herb Schwetman, for their valuable feedback and guidance on my work. Interacting with each of them on various occasions has always been a learning experience for me. Their knowledge and insights are truly invaluable.

I would also like to extend my gratitude to my research collaborators at Oracle labs: Pranay Koka, Herb Schwetman, Xuezhe Zheng and Ashok Krishnamoorthy. Their guidance and support has tremendously shaped the work presented in this thesis. Both Herb and Pranay were excellent mentors to me during my internships at Oracle. I am truly fortunate to have worked with and learned so much from them. I am also thankful to my manager Alan Wood for his continuous support and encouragement during my internships at Oracle labs. I would also like to acknowledge the generous donations from Oracle in support of this work.

I would also like to give a shout-out to my graduate school friends and colleagues: Atif Hashmi, Syed Gilani, Mitch Hayenga, Dibakar Gope, Vignyan Reddy, David Palframan, Sean Franey, Andrew Nere, Zhong Zheng, Shoaib Altaf and Rehan Ahmed. Thank you for all the lively discussions and for making graduate school interesting.

Most of all, I would like to thank my parents, Mohammad Zulfiqar and Rehana Zulfiqar, and my brothers, Imran and Farhan, for their unwavering love and support.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The trend towards many-core systems continues to grow. Scaling single chip systems with higher core counts however leads to increasing fabrication costs and low process yields. Multichip systems can alleviate these concerns but require substantial chip-to-chip bandwidth to provide sustained performance. Due to the limited density of chip I/O pins and excessive power consumption of high-speed serial links, *silicon photonic technology* has been proposed as an alternative for networking multichip systems. This dissertation explores the design space of multichip photonic networks and makes several contributions.

Optical crossbars (channel sharing) designs are popular in nanophotonic literature. These architectures improve performance by allowing nodes to share the network channels. However, this sharing comes at a cost: increased optical (laser) power consumption. To explore this performance-power trade-off, an analytical model is developed in this thesis that quantifies the limits and performance gains of channel sharing techniques. Furthermore, an opportunistic channel sharing architecture called 'wavelength stealing' is proposed. The wavelength stealing architecture does not incur any arbitration overheads in accessing the shared channels and guarantees fairness.

Switched networks are ubiquitous in computer systems. In photonic networks, the switching elements (routers) can be optical or electrical. A recent paper has shown that breakthroughs are required in device development to make optical switching viable. This leaves electrical switching as an alternative design option to explore for switched photonic networks. In this context, this dissertation is the first work to provide an in-depth evaluation of electrical switching within the constraints of silicon photonic technology. Advocating a 'topology-aware' design approach, this thesis also proposes novel router designs that avoid expensive logic structures such as allocators and crossbars. Furthermore, this dissertation provides novel quality-of-service mechanisms for providing performance isolation and service differentiation between virtual machines (VMs) running on a nanophotonic system.

1 INTRODUCTION

Computer architects are in a never-ending quest to increase processor performance. The performance of a processor is typically measured in terms of how fast it can execute programs. Using Iron law, the execution time of a program can be broken down into several terms:

$$\text{Execution time of a program} = \frac{\text{\# of instructions in a program}}{\text{IPC} \times \text{frequency}} \tag{1.1}$$

For a long time[1], the primary focus of computer architecture research was to increase the product terms in the denominator thereby reducing the running time of programs. Numerous techniques were proposed to increase instructions-per-cycle (IPC) such as superscalar designs [85], out-of-order (OoO) execution [85], value prediction [63, 62] and memory disambiguation [70, 20]. At the same time, researchers were able to achieve orders of magnitude increase in frequency using techniques such as pipelining [18, 92, 85] and fabricating transistors that could switch faster with each technology generation [19, 55]. Through exceptional pace in innovation, the performance of processors increased exponentially [55]. However, at the turn of this century, researchers realized that simply relying on IPC and frequency to stay on an exponential performance curve would be challenging going forward [1, 74]. This caused a fundamental shift in the way computer systems are designed and ushered in the 'multicore era' where performance scaling is envisioned by increasing core counts [36, 94, 53].

Today, state-of-the-art processors feature multiple cores. Depending on the market segment, quad-core and octal-core chips are commonplace. Road maps and projections predict that multicore scaling will lead to hundreds if not thousands of cores on a chip [33, 39, 7]. Although on-chip wires are capable of providing high bandwidths [34, 35, 41, 68], scaling single chip systems with higher processor counts leads to two main challenges:

[1] fondly referred to as the 'single-thread performance era'

increasing fabrication costs and *low process yields* [88, 72, 50]. Coupled with the end of Dennard scaling, there is growing evidence that increasing core counts on a single chip will not be a viable option in the near future [28]. One strategy to overcome these scalability challenges is to aggregate several chips in a package [50]. This multichip approach however requires enormous chip-to-chip communication bandwidth to provide sustained performance. *The goal of this thesis is to provide novel architectures and solutions to address the communication challenges in multichip systems.*

1.1 Overcoming Die Size Limits

Placing a very large number of cores on a single piece of silicon may lead to unacceptable design complexity and yield risks while at the same time increasing fabrication costs to prohibitive levels. Instead, what a designer really wants is to 'stitch' together multiple smaller dies with a communication fabric that can achieve the same performance as a single large monolithic piece of silicon. This leads us to the central question that is the focus of this dissertation: *what is an efficient communication fabric that can close the gap between intra-die and inter-die bandwidth performance?* There a many technologies that can be deployed to interconnect multiple dies together, each with different trade-offs. They are discussed below.

Solution #1: High-speed I/O Pins

An array of chips can be interconnected using off-chip I/O pins. Unfortunately, this solution suffers from the following **limitation(s)**:

- **Limited pin density:** The density of off-chip I/O and package routes dramatically lags that of on-chip wires [9]. This gap in intra-die and inter-die bandwidths makes chip-to-chip communication the system bottleneck.

- **High power consumption:** Due to limited pin counts, off-chip I/O is typically implemented using overclocked serial links [12]. The excessive power consumption of these high-speed serial links restricts system scalability to higher chip counts.

Solution #2: 3D Integration (Die Stacking)

Another popular approach for aggregating multiple chips is to use vertical 3D stacking of dies interconnected using through-silicon-vias (TSVs) [81, 15, 93]. This technique suffers from the following **drawback(s)**:

- **Heat removal challenges:** 3D die stacking can dramatically increase thermal hotspots if two highly active dies such as processors are stacked on top of each other. Furthermore, dies that are stacked farther away from the heat sink suffer from thermal isolation leading to self-heating [17]. These thermal considerations restrict both the type and number of dies that can be stacked vertically.

- **Power delivery limitations:** Placing chips squarely atop one another restricts the amount of power that can be delivered to the dies. This means that vertical die stacking is best employed for low-power applications such as DRAM integration [6, 65].

Solution #3: Capacitive Proximity Communication (PxC)

Capacitive proximity communication (PxC) technique incorporates special silicon 'bridge' chips that employ dense on-chip (electrical) wires to carry data. This is illustrated in figure 1.1. The bridge chip is placed faced down on two logic (e.g. processor) dies as shown in the figure. The chip-to-chip connections are formed using tiny capacitive pads. Using PxC, highly dense on-chip wires on one chip can be extended across a chip-to-chip gap leading

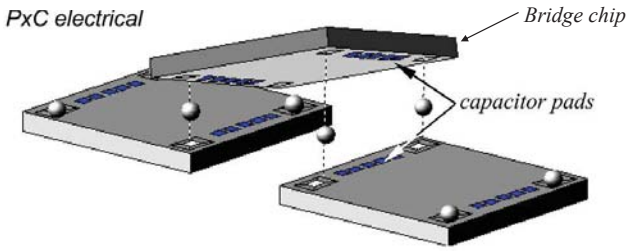


Figure 1.1: Capacitive proximity communication (PxC) (photo taken from [51]).

to a system that behaves like a logically continuous piece of silicon. The PxC approach however, suffers from several **disadvantage(s)**:

- **Low speed transmission:** Signals propagate at only $5 - 10\%$ the speed of light in standard (electrical) on-chip wires [34]. This leads to high message latencies in large-scale systems.
- **High power consumption over long distances:** PxC provides high bandwidth communication at low power consumption over *short* distances [51]. However, for distances beyond a few chips (i.e. few centimeters), the energy-per-bit consumption of electrical wires becomes expensive.

Solution #4: Optical Proximity Communication (OPxC)

In optical proximity communication (OPxC), data transmission occurs using optical signals i.e. light. Once the electrical data is converted into optical form, it can be carried in optical waveguides that are fabricated on silicon-on-insulator (SOI) chips. Light travelling in one chip can be coupled across a chip-to-chip gap using mirrored surfaces. This is illustrated in figure 1.2.

Compared to the solutions discussed earlier, optical proximity communication (OPxC) holds several **advantage(s)**:

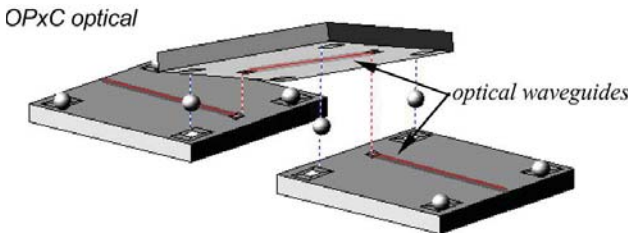


Figure 1.2: Optical proximity communication (OPxC) (photo taken from [51]).

- **High speed transmission:** Optical signals propagate at approximately 30% the speed of light in silicon [50]. Thus, compared to electrical signals, transmitting messages using optics incurs much lower latencies.
- **Low power consumption over long distances:** The energy cost of optical communication is projected to be much less than electrical communication, especially over distances larger than a few centimeters [51, 8].
- **High bandwidth densities:** Optical communication enables unprecedented bandwidth densities as each waveguide in the system can carry multiple parallel streams of information [50]. Such plentiful bandwidth availability is important to scale the system to higher chip counts.

1.2 Macrochip - A Multichip Communication Substrate

This section gives a brief description of the *macrochip* system, a wafer-scale technology platform for aggregating multiple dies [51, 50]. This system employs optical communication and was developed at Oracle (formerly Sun) labs. The macrochip system is depicted in figure 1.3.

The macrochip architecture consists of an array of sites (also called nodes). Each site can house a conventional die such as a processor and/or memory chip. The die sits face up in a site opening and a special optical bridge chip is mounted on top of it facing down. The

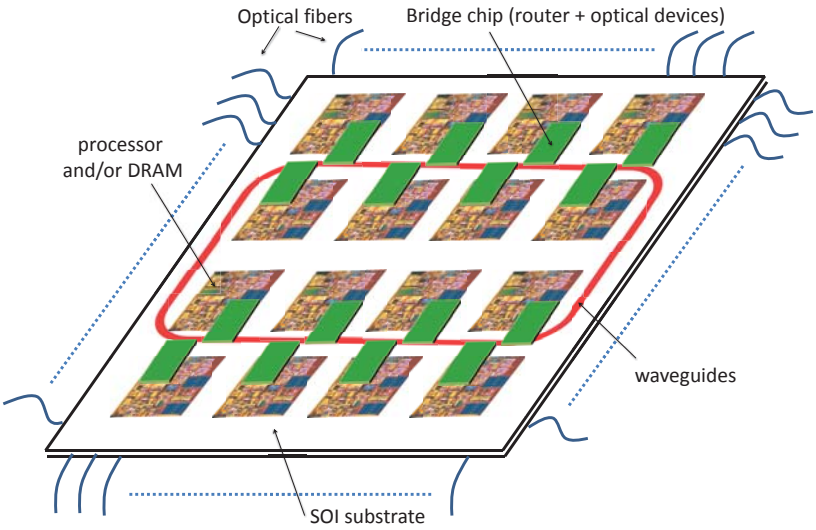


Figure 1.3: A 16-site macrochip system.

bridge chips house optical devices to perform electrical to optical conversions and vice versa along with the associated router logic. Site-to-site communication occurs via waveguides that are fabricated on an silicon-on-insulator (SOI) substrate. Power is delivered to each site from a top plate and heat is removed from the package from the opposite side (bottom).

The macrochip system uses a combination of capacitive proximity communication (PxC) and optical proximity communication (OPxC) to communicate data between the sites. The raw electrical data originating at a site is first communicated across the site-to-bridge gap using PxC. This data is then converted into optical form and transferred to a waveguide in the substrate layer using OPxC. The waveguides then carry the optical signals to the intended destination where the data is converted back into electrical form and delivered to the receiving site. By converting the data into optical form, the macrochip system is able to exploit the latency, bandwidth and energy advantage offered by optical communication when transferring data over long distances (tens of centimeters).

The macrochip system is used as the baseline architecture in this thesis. However, the contributions of this dissertation go beyond the macrochip and are broadly applicable to any silicon photonic interconnect.

1.3 Thesis Contributions

The research presented in this dissertation addresses the following problem: *For a macrochip like system, what is an efficient network design?* The goal in conducting this research was to devise novel architectures that provide robust performance (latency/ throughput) given cost metrics such as area and power consumption. What makes this research challenging is that traditional electrical solutions are not directly applicable in the optical domain as the opportunities afforded by this technology as well as the constraints imposed by it are quite different. In this regard, this dissertation makes the following contributions.

- **Evaluation of 1-hop optical network topologies:** The topology of a photonic interconnect impacts both its performance and laser power requirements. The 1-hop fully-connected point-to-point topology offers arbitration-free connectivity with low energy-per-bit consumption, but suffers from low node-to-node bandwidth. Alternatively, another class of 1-hop topologies called channel sharing or optical crossbar designs improve inter-node bandwidth but incur higher laser power consumption in addition to the performance costs associated with arbitration and contention. As part of this dissertation, an analytical model is developed that demonstrates the limits and gains of channel sharing techniques over a fully-connected topology under realistic device loss characteristics.

- **Providing arbitration-free accesses on shared channels:** This dissertation proposes a novel photonic interconnect architecture that uses 'opportunistic' channel sharing. Specifically, this network design does not incur any arbitration overheads and guarantees fairness. Evaluation of this interconnect architecture using detailed simulation in the context of a 64-chip macrochip system reveals that this new approach achieves up to 28% better energy-delay-product (EDP) compared to a fully-connected topology for some high-performance-computing (HPC) applications.

- **Investigation of electrical switching ($\geqslant 1$ hops) techniques in photonic networks:** Although 1-hop designs such as channel sharing networks incur low dynamic energy costs by reducing the number of electrical-to-optical (E/O) and optical-to-electrical (O/E) conversions along a communication path, the high laser power consumption of these networks make them less attractive for adoption in the near-term. Relaxing the 1-hop constraint, this dissertation presents an in-depth evaluation of electrical switching ($\geqslant 1$ hops) designs within the constraints of silicon photonic technology. Both low-radix and high-radix topologies are investigated and it is demonstrated that an adaptively routed fully-connected network capable of non-minimal (2-hop) routing provides significantly higher performance compared to popular topologies such as flattened butterfly or fat-tree.

- **Novel low-cost router architectures:** The extreme radix of a fully-connected topology presents a significant challenge in terms of router design. It is demonstrated that the traditional input-queued crossbar router becomes prohibitively expensive in terms of area and power consumption when scaled up for a fully-connected topology. Instead, this dissertation advocates a 'topology-aware' router design approach and proposes novel routers that avoid expensive logic structures such as allocators and crossbars. It is shown that compared to a naive traditional router, the proposed designs provide 95% and 83% savings in power and area respectively without compromising performance or throughput.

- **Incorporating quality-of-service (QoS) guarantees in optical networks:** This dissertation proposes several novel mechanisms to provide quality-of-service (QoS) guarantees in a photonic network. This enables a hypervisor to map virtual machines (VMs) with different bandwidth demands to appropriate regions in the network.

1.3.1 Relation to Previously Published Work

This thesis encompasses work, both published and unpublished, at the time of this writing. The published work is described below:

- **Wavelength Stealing: An Opportunistic Approach to Channel Sharing in Multi-chip Photonic Interconnects (MICRO - 2013).** This paper [109] explores channel sharing or optical crossbar designs and proposes a novel arbitration-free interconnect architecture called 'wavelength stealing'. This work is discussed in detail in chapter 4. This paper was coauthored by Pranay Koka, Herb Schwetman, Mikko Lipasti, Xuezhe Zheng and Ashok Krishnamoorthy.

1.4 Thesis Organization

The rest of the dissertation is organized as follows. Chapter 2 provides background information on silicon photonic technology. Chapter 3 presents prior work related to this dissertation. A detailed investigation of 1-hop silicon photonic network designs is presented in chapter 4. Electrical switching in optical networks is explored in chapter 5. Finally, chapter 6 concludes this thesis and discusses possible directions for future work.

2 PHOTONIC TECHNOLOGY

To make optical communication a reality in multichip computing systems, two types of challenges need to be addressed: device-level and architectural. Device-level challenges involve design and fabrication of optical devices that are low-loss and high speed. Such optical devices include components such as modulators, drop-filters, couplers, waveguides, etc. and constitute the building blocks of a silicon photonic network. Fabrication of these devices is under extensive on-going development and many components have been demonstrated in the literature [107, 102]. From an architectural standpoint, the main challenge is to design an interconnect that is energy efficient and yields the best performance on the target applications. Although this dissertation is primarily focused on the architectural aspects of photonic network design, a basic understanding of the various optical components involved in building these networks is important. The goal of this chapter is to provide such a foundation to the reader.

The rest of the chapter is organized as follows. Section 2.1 describes major optical components that are employed in silicon photonic networks. The basic operation of a photonic link is discussed in section 2.2. Sections 2.3 & 2.4 briefly highlight the considerations and opportunities afforded by silicon photonic technology. Finally, section 2.5 concludes this chapter.

2.1 Optical Devices and Components

In photonic networks, a channel (logical connection) between a sender and a destination is formed using one or more waveguides. Each waveguide can support multiple wavelengths (links) using a technique called wavelength-division-multiplexing (WDM). These wavelengths carry bit information in the form of modulated light. Along a communication path however, there may be many optical components that interact with the light signal. The

goal of this section is to describe these optical components and discuss their performance characteristics. The main components employed in photonic networks include: lasers, waveguides, modulators, multiplexers, drop-filters (de-multiplexers), receivers, interlayer couplers, splitters and optical switches.

Lasers

Laser sources generate the unmodulated light that is imprinted with information in optical channels. Light from external laser sources is brought to the edge of the system boundary via optical fibers and is coupled into the waveguides that are fabricated on a SOI routing fabric using either edge coupling [13, 4] or grating couplers [66, 103]. A WDM-capable laser source can inject light at many wavelengths ($\lambda_1, \cdots, \lambda_i, \cdots, \lambda_n$) thereby providing many parallel independent streams for modulating data. Furthermore, each wavelength of light (say λ_i) typically carries enough power such that it can be split up further once it is brought into the system fabric and used to power the same wavelength (λ_i) link in multiple waveguides [50]. By using this optical power sharing technique, the total number of laser sources required in the system can be reduced.

One of the biggest challenges facing silicon photonic technology is that generating laser light is currently expensive. This is because the efficiencies of commercially available WDM lasers is low: $1 - 5\%$ [108, 14, 54]. This has important implications in terms of the total bandwidth that is available in a photonic network. These considerations are discussed in section 2.3 as well as in chapters 4 and 5.

Waveguides

Waveguides are the 'wires' in a photonic network. They serve to route optical signals from a source site to the respective destination site. As discussed earlier, waveguides are capable of supporting many wavelengths using WDM. Hence, many parallel independent streams

of information can be supported in a single waveguide leading to high bandwidth densities in the system. The propagation speed of light in waveguides fabricated in an SOI substrate is approximately 30% the speed of light in vacuum. Furthermore, the propagation losses in waveguides is low: $0.05 dB/cm$ [60]. This largely obviates channel length considerations which is in stark contrast to traditional electrical networks where an important design goal is to avoid long global wires [42].

An important consideration in the design of silicon photonic networks is to avoid waveguide crossings in the SOI routing substrate as they introduce significant crosstalk and power loss [49]. This constraint has important implications both in terms of waveguide layout as well as the number of routing layers employed in the routing substrate. Furthermore, the amount of optical power that a silicon waveguide can carry as well as the number of wavelength channels (WDM factor) it can support is limited due to certain device-level considerations [80, 22, 27, 37, 26, 24]. These factors affect the number of waveguides needed to build a network topology and have important implications on the available bandwidth. These considerations are discussed in section 2.3.

Modulators

Modulators convert an electrical bit stream to an optical data stream, i.e. electrical-to-optical (E/O) conversion. One of the most promising candidates for implementing a modulator device is the reverse-biased, carrier-depletion ring resonator [107]. This ring modulator is fully compatible with standard complementary-metal-oxide-semiconductor (CMOS) fabrication processes and can operate at data rates between $10 - 20 Gbps$.

A significant issue for ring resonators is that they are highly sensitive to fabrication inaccuracies and ambient temperature variations [73, 101]. To correct the drifts in resonance frequency arising from process variations and temperature fluctuations, a tuning mechanism is incorporated with the ring resonator devices. Following the methodology used in

prior papers [50, 109], this dissertation models the tuning cost as a static component in a photonic network's total power consumption. Specifically, the tuning power is modelled as $0.3mW/ring$.

Multiplexer

A multiplexer device combines wavelengths of different channels into a single waveguide. Mutiplexer devices are typically employed when there are two types of waveguides in the system. For example, the macrochip system discussed in chapter 1 employs short local waveguides of smaller pitch in the bridge chips and low-loss global waveguides of larger pitch in the SOI routing substrate. In this case, the wavelengths of multiple local waveguides can be multiplexed into a single global waveguide using a multiplexer. One way to implement a multiplexer is to use cascaded ring resonators [105]. Using ring resonators to build multiplexers however incurs tuning costs similar to ring modulator devices.

Drop-filter (De-multiplexer)

A drop-filter or de-multiplexer is used to demultiplex a single wavelength from a shared multi-wavelength waveguide. That is, this device has two outputs: one output extracts the selected wavelength while the other output contains the remaining wavelengths. Drop-filters can be fabricated using ring resonators [79]. Consequently, they suffer from the same tuning issues as modulator ring resonators.

Receiver

A receiver is made up of a photo-detector, amplifier and thresholding circuit [64]. The photo-detector converts the received optical signal into electrical form, i.e. optical-to-electrical

(O/E) conversion. This electrical signal is then passed through an amplifier circuit. Finally, a thresholding circuit is used to decide whether a bit 0 or 1 was sent. The receiver sensitivity in this dissertation is modelled as $-21\mathrm{dBm}$ similar to what has been assumed in prior work [50].

Interlayer Coupler

An interlayer coupler or optical-proximity-communication (OPxC) couples the light traveling in a waveguide on one chip to a waveguide on another chip provided the chips are placed face-to-face (see section 1.1). There are two popular technologies to accomplish interlayer coupling: mutually aligned waveguide gratings [67, 97] or mutually aligned reflecting mirrors [52, 106]. This dissertation assumes the latter approach which incurs a device loss of about $2-3\mathrm{dB}$ per coupling.

Splitter

An optical splitter is a single input, dual output broadband device that splits the intensity of light travelling on all wavelengths at an input waveguide into two portions, one for each output waveguide [30]. Specifically, the number of wavelengths at the input and output waveguides remain the same. It is just that the optical power across all incoming wavelengths is split in half and transferred to each of the two outputs. By halving the intensity of incoming light, an optical splitter imparts a device loss of $3\mathrm{dB}$ per output. Network designs that rely on broadcast capability require optical splitters to provide this functionality.

Optical switches

Optical switches are broadband devices that route light from different input to output channels depending on how they are configured. Optical switches can be built using multiple optical device technologies. A Mach-Zehnder interferometer (MZI) [29, 90] can be used to implement an optical switch with suitable switching speeds albeit at the cost of high area and power consumption. Alternatively, periodic resonances of ring resonators can be used to create a smaller optical switch compared to the MZI approach [56]. However, the tuning power requirements of ring resonators along with the high optical losses incurred by these switches in the current technology generation make optical switching difficult to implement in the near-term [49]. Furthermore, since optical buffering is not yet feasible, optical switches along a path have to be setup in advance before sending data. This limits their applicability to circuit-switched-style networks which incur high message latencies.

2.2 Basic Operation of a Photonic Link

Bits in optical channels are represented by the presence or absence of light. A bit 1 is represented by light; whereas, a bit 0 is indicated by the absence of light.

Sending information - Modulation: The direct modulation of laser light for encoding information is performed by ring modulators (denoted by 'M' in figure 2.1). The modulator is placed next to a waveguide and is tuned to a particular wavelength. The ring modulators have two modes of operation, on-resonance and off-resonance, used to write a bit value of 0 and 1 respectively. In on-resonance, any light passing by on the tuned wavelength is coupled (drops) into the cavity of the ring modulator and is attenuated. Hence, during the on-resonance mode, a bit 0 is written on the wavelength as light is absorbed by the ring as shown in figure 2.1a. During off-resonance, the ring modulator simply allows the light to

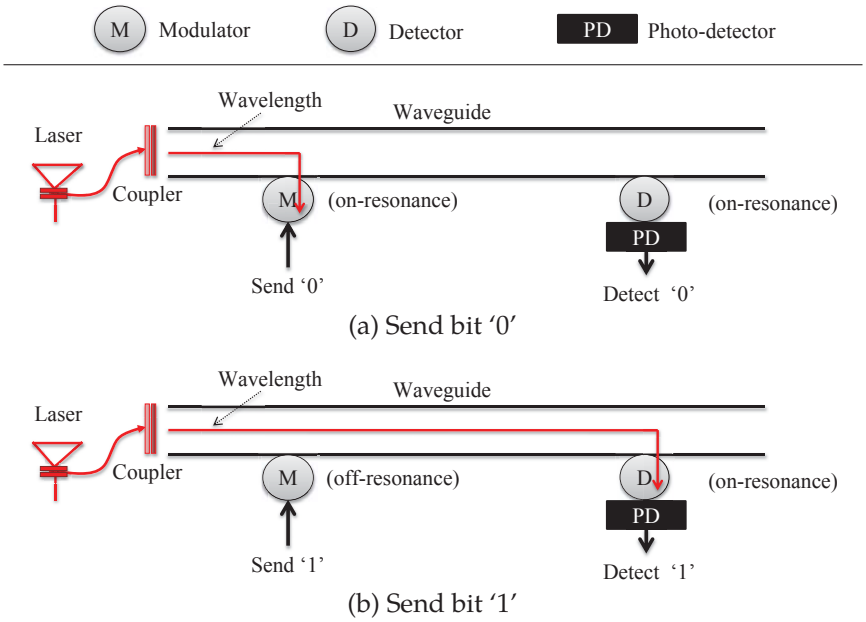


Figure 2.1: Example depicting operation of a photonic link when (a) sending a bit '0' or (b) sending a bit '1'.

pass by thereby writing a 1 on the link as shown in figure 2.1b. The ring modulator can be brought in and out of resonance by controlling the electrical charge or temperature applied to it.

Receiving information - De-modulation: On the receiver side, a drop-filter and photo-detector device are used (denoted by 'D' and 'PD' in figure 2.1 respectively). The drop-filter ring resonator is tuned to the same wavelength as the modulator ring and always kept on-resonance as shown in the figure 2.1. However, instead of dissipating the absorbed light, it is ejected out to a photo-detector (O/E conversion) which informs the receiver whether a bit 0 or 1 was sent.

2.3 Technology Considerations

This section surveys some important design considerations in nanophotonic systems. Understanding these constraints is necessary to explore the design trade-offs associated with this technology. These technology constraints can be broken down into three categories: link-level, network-level and system-level.

2.3.1 Link-level Considerations

The pertinent link-level considerations in photonic networks are described below.

- **Source enough laser light to overcome link losses:** Typically, each optical component is characterized by optical losses which represent the degradation in intensity of light as it passes by the device. Optical power that is injected by the laser source experiences degradation due to these losses. Eventually, enough laser power has to reach the receiver such that it can reliably differentiate between a bit 1 and a 0. Consequently, photonic links have to be provisioned with enough laser light so as to overcome the worst-case losses along a light-path.

- **Incorporating more optical devices along a link leads to higher optical loss:** The higher the number of optical components incorporated along a light-path, the higher the optical loss becomes for that link. To overcome this loss, this link needs to be sourced with more laser light. This increased laser power consumption has important implications on network designs that employ a large number of optical components along a link to get higher performance. These trade-offs are explored in detail in chapter 4.

- **Waveguides have limits in terms of optical power and WDM:** Waveguides are responsible for carrying information in photonic links. Due to device physics, waveg-

uides are limited in terms of the amount of laser light they can carry as well the number of wavelengths they can support [49]. These limitations together with the bandwidth requirements of the desired topology ultimately determine the total number of waveguides required to build a network. Now, depending on area limitations, it may or may not be possible to fabricate these many waveguides on the routing substrate. These scalability considerations are discussed at length in chapter 5.

2.3.2 Network-level Considerations

The following network-level constraints must be carefully considered when designing photonic networks.

- **Generating laser light is expensive:** Light generated by off-chip laser sources is used to convey information in optical networks. Generating this laser light is expensive however. This is because efficiencies of commercial WDM lasers is only $1-5\%$ [108, 14, 54]. This means that up to 95% of the wall-socket power flowing into the laser source is completely wasted. Due to this expensive power overhead, all photonic network designs in this dissertation are evaluated under a fixed laser power budget.

- **Input laser power is limited:** In addition to the constraints involved in generating laser light, delivering this optical power to the routing substrate is challenging as well. For example, the amount of input laser power that can be delivered to the macrochip system is limited by the number of optical fibers that can be connected along the perimeter [49]. This makes optimizing for laser power consumption a first-order design constraint in photonic networks.

- **Waveguide layouts need to avoid crossings:** An important consideration in the design of silicon photonic networks is to avoid waveguide crossings in the SOI routing

substrate as they introduce significant crosstalk and power loss [49]. Two popular approaches explored in literature include:

- Employing a multi-layer routing substrate where the horizontal and vertical waveguides are incorporated on separate layers [50]. Inter-layers couplers are then used to communicate between the layers.
- Using a single-layer routing substrate but relying on non-minimal channels to avoid crossings [49]. In this case, channels of the network originate at the sender and are routed to the destination as part of a loop.

All photonic network designs considered in this dissertation employ the latter approach for avoiding waveguide crossings, i.e. they use non-minimal channels laid out as part a loop.

2.3.3 System-level Considerations

The following constraint must be carefully considered when deploying silicon photonic technology in a system.

- **Photonic technology is static power dominated:** Silicon photonic networks based on ring-resonators are static power dominated, specifically, laser power and ring-resonator tuning power. This means that the bulk of the power consumption in these networks is independent of the network load. To justify these high activity-agnostic power costs, photonic networks should be deployed in high utilization scenarios.

2.4 Technology Opportunities

Although photonic technology is facing many challenges, it is also providing some unique opportunities. These advantages are leveraged heavily in the network designs proposed in

this dissertation. The most pertinent technology opportunities are highlighted below.

- **'Speed of light' communication leads to low message latencies:** The propagation speed of optical signals is significantly higher compared to traditional electrical communication. This results in lower message latencies compared to networks built with electrical channels. This latency advantage is critical for latency-sensitive traffic such as that encountered in cache-coherent shared memory systems.

- **WDM enables high bandwidth densities in the network:** The ability to multiplex many independent parallel streams of information in a single waveguide or fiber enables tremendous bandwidth densities in the network. Such abundance in bandwidth enables designers to explore richly-connected topologies, even those that are considered too prohibitive for traditional electrical networks, such as a fully-connected topology. Furthermore, such plentiful bandwidth facilitates system scaling to higher chip (node) counts as discussed in chapter 5.

- **Low propagation losses mitigate channel length considerations:** Light traveling in optical links suffer from low propagation losses $\approx 0.05\mathrm{dB/cm}$ [60]. Thus, light can be carried to long distances without the need for repeaters as is the case in traditional electrical cables thereby mitigating channel length considerations. This is especially advantageous in high-radix topologies that have been shown to provide higher performance albeit at the cost of longer inter-router channels [42]. Furthermore, the ability to cheaply incorporate long channels also mitigates many layout and packaging considerations such as avoiding waveguide crossings as discussed earlier.

2.5 Summary

This chapter presented a brief overview of nanophotonic components involved in building optical networks. These include: lasers, modulators, waveguides, drop-filters, receiver circuits etc. The chapter then described how these components come together to build a simple photonic link. Modulators are employed to impart bit information on laser light which is carried by waveguides to the receiver where it is converted back into electrical (bit) form. Since silicon photonics is a young technology, it is facing many challenges. Some important considerations that must be carefully weighed when designing photonic networks were presented. These constraints were broken down into three categories: link-level, network-level and system-level considerations. Most of the constraints revolve around optical device losses, layout of waveguides, and in generating as well as delivering laser light into the system. Finally, the chapter concludes by highlighting some important opportunities afforded by silicon photonic technology and how they come into play in network design.

3 RELATED WORK

This dissertation makes contributions in three aspects of photonic network design: network topology, router microarchitecture and quality-of-service (QoS) guarantees. The goal of this chapter is to survey the most pertinent prior work related to these aspects. As such, this chapter is structured as follows. Section 3.1 covers the various optical network topologies proposed in prior work. Next, related work that addresses the scalability challenges of router microarchitectures is presented in section 3.2. Mechanisms to incorporate QoS guarantees in photonic networks proposed in literature are surveyed in section 3.3. In all these sections, the contributions made by this dissertation are briefly highlighted to convey to the reader where these ideas fit into the landscape of prior work.

3.1 Silicon Photonic Topology Approaches

In recent years, many silicon photonic network designs have appeared in literature ranging from the simplest unshared fully-connected network [50, 49, 51] to numerous sharing based designs [95, 96, 69, 76, 100, 78]. In most cases, the assumed device losses have varied greatly. Some designs have made aggressive (low) loss assumptions and, as a result, have been able to show significant performance gains. Yet, other papers have assumed more conservative device loss parameters and have argued for simpler designs and topologies to keep the laser power consumption low. Broadly speaking, the photonic topology designs proposed in literature can be placed into two categories: channel sharing (optical crossbars) and switched (path sharing) networks. These categories are surveyed below.

3.1.1 Channel Sharing Topologies (Optical Crossbars)

Channel sharing networks are also called 'optical crossbars' or 'all-optical' designs. The key defining feature of these networks is that they are 1-*hop*. Specifically, in these networks, messages originate at the sender and are conveyed to the destination in just 1-hop. This minimizes the dynamic energy expended in converting the information between electrical and optical form. Specifically, these networks minimize the number of E/O and O/E conversions required to convey a message. However, these savings in dynamic energy typically come at the cost of increased laser power consumption per link as explained in chapter 4.

In channel sharing, the wavelengths of the channels are shared by multiple senders and/ or receivers. This is accomplished by placing microring resonators belonging to different nodes along the channel waveguides. Different flavors of channel sharing have been proposed in literature: single-writer single-reader (SWSR) [50, 49], multiple-writer single-reader (MWSR) [95, 96, 69, 109], single-writer multiple-reader (SWMR) [78], and multiple-writer multiple-reader (MWMR) [76, 100].

3.1.1.1 Single-Writer, Single-Reader (SWSR)

In SWSR, each communication channel has only one source and one destination. This category represents a degenerate case where there is no sharing between the senders or receivers on the network links. The only architecture that fits this category is a statically allocated fully-connected, point-to-point (P2P) topology. A computer architect unfamiliar with photonic literature may be astounded by the inclusion of a fully-connected network as a possible design point because of its packaging complexity. However, owing to the high bandwidth densities of nanophotonics (see chapter 2), some recent photonic papers [50, 49] have proposed efficient layouts of the fully-connected topology for the macrochip fabric.

Koka *et al*. in [50] present a dual-layer layout of the fully-connected topology for the

macrochip system. Specifically, the vertical and horizontal waveguides are fabricated on separate routing layers to avoid waveguide crossings. Inter-layer OPxC couplers are then employed to transmit optical signals between these routing layers. Since, each sender has a dedicated channel to every other destination and the routing algorithm employed is 1-hop, this network does not suffer from any switching or arbitration overheads. In a follow on paper, Koka *et al*. [49] describe a single-layer layout of the fully-connected topology for the macrochip fabric. This layout employs non-minimal channels to avoid waveguide crossings and approximately halves the number of inter-layer couplers employed compared to the two-layer layout. Since, the P2P topology employs the least number of optical devices along a link, it incurs the least complexity and link loss compared to the channel sharing designs presented below.

3.1.1.2 Single-Writer, Multiple-Reader (SWMR)

Multiple reader channels are typically implemented using optical broadcast or tunable microrings to selectively divert all the optical energy to one destination. For the latter case, SWMR networks require an broadcast-based mechanism to notify the target destination to tune in and the other destinations on the channel to tune out.

Kirman *et al*. [47] implement a SWMR based bus interconnect. In this architecture, each node uses a dedicated channel to transmit its information which can be listened on by the other nodes in the network. Specifically, if there are N nodes in the network, then there are N channels. Each channel will have one sender and $N-1$ receivers. The sender node broadcasts its messages to all the receiver nodes on its channel. Implementing this broadcast mechanism requires $(N-1)\times$ laser power to be sourced per link compared to a unicast link. Since generating laser light is expensive (see chapter 2), the high laser power consumption of this network limits its scalability. The Firefly architecture proposed in [78] mitigates some of the scalability concerns of [47] by restricting the expensive broadcast to

just a small header flit that notifies the intended destination to turn on its receiver. This notification mechanism however adds latency to message communication.

Due to active and pass-through losses of ring-resonator devices, multiple reader channels have significant link losses that increase with sharing degree leading to high laser power consumption in these networks.

3.1.1.3 Multiple-Writer, Single-Reader (MWSR)

Network architectures in this category require multiple ring modulators per channel to enable shared access to a waveguide or selective wavelengths in a waveguide from multiple sources. This category of interconnects has similar link loss as SWMR networks and require an arbitration mechanism at the source to resolve access conflicts on the shared channel.

The Corona network [95] implements a MWSR architecture which has N channels for an N node network. Each channel has one receiver and $N-1$ writers on it. To resolve conflicts between the senders, each channel employs a special control wavelength that carries tokens. A token conveys the right to use the channel. Before a sender node can write a message on the channel, it has to acquire a token. After this sender is done communicating its information, it injects the token back on the arbitration wavelength. This scheme fairly allocates the channels in a round-robin manner. This leads to high channel utilization when the contention between the senders is high. If however, only a small subset of the senders have messages to send, then the channel utilization suffers as a sender may need to wait for the full token rotation latency to acquire an uncontested token. Vantrease *et al*. [96] build on [95] and propose two new token-based schemes called 'token channel' and 'token slot' that improve the channel utilization by foregoing the round-robin constraint of the earlier design. In token channel, only one token is circulated on the token wavelength. Alternatively, the token slot architecture employs a continuous stream of tokens to improve the channel utilization even further. To provide fairness, these schemes use some special

control wavelengths in addition to the token wavelength. These extra wavelengths increase the control overheads incurred per channel and result in higher laser power consumption in the network. Morris *et al*. [69] propose a network-on-chip interconnect architecture that employs MWSR channels. This design uses the token slot scheme proposed in [96] for arbitration between the sender nodes that share a network channel.

3.1.1.4 Multiple-Writer, Multiple-Reader (MWMR)

MWMR networks require microring resonators both at the source and the destination leading to the highest link loss compared to the other three categories. These networks require both an arbitration mechanism at the source as well as a mechanism to select the appropriate destination.

MWMR channels were first considered by Pan *et al*. [76] where both senders and receivers share the channels. This design proposes a 'two-pass' continuous token arbitration scheme between the senders and requires receiver side arbitration as well on its MWMR channels. In the two-pass token arbitration scheme, only a single node is serviced on the first pass and if unused, any node can be serviced by the token on the second pass. By prioritizing different nodes in the first pass, this scheme ensures a minimum level of fairness in the network. The 'Channel Borrowing' scheme [100] argues for simplifying the two-pass token arbitration proposed in [76] by restricting the number of senders on shared channels to two. One of the two senders on a channel is called the 'owner' while the other one is called the 'borrower'. The owner node has higher priority for sending data compared to the borrower. These priorities are enforced using token-arbitration. Furthermore, the owner and borrower priorities are distributed across the network channels in such a way that a minimum level of bandwidth is guaranteed to all nodes in the network.

3.1.2 Path Sharing Topologies (Switched Networks)

Path sharing topologies include switched networks. In switched networks, messages can take multiple hops to reach their intended destination. That is, the communication in these networks experience $\geqslant$ 1-*hops*. Higher hop counts in conveying a message result in more E/O and O/E conversions. This leads to higher dynamic energy consumption compared to channel sharing designs (discussed in section 3.1.1) where messages are relayed in just 1-hop. However, silicon photonic networks are static power dominated, specifically laser power and tuning power of ring-resonators. Thus, even though switched networks incur higher dynamic energy consumption, this has negligible impact on the total power consumption of photonic networks as explained in chapter 5. Therefore, switched network designs represent viable options for deployment in silicon photonic systems.

Switched networks are ubiquitous in computer systems. In a silicon photonic network, the switching elements (routers) can be optical [50, 49, 84, 21, 46] or electrical [40, 50].

3.1.2.1 Optical Switching

In optically-switched photonic networks, both the switching elements (routers) as well as the network links are optical. The shared-source-row (SSR) architecture proposed by Koka *et al*. [50] falls into this category of optical networks. In this architecture, the N nodes in an N $\times$ N array share a data channel to each destination. This shared channel is implemented using MZI broadband switches (see chapter 2). This design proposes a two-phase arbitration mechanism that employs special control wavelengths to setup the optical switches. In their follow on paper, Koka *et al*. [49] propose a broadcast-based arbitration mechanism that incurs lower complexity than the two-pass arbitration scheme but gives similar performance. In addition, this paper presents an implementation of the butterfly topology on the macrochip platform that employs optical switches. To avoid waveguide crossings, the optical switches in this topology are implemented using two

waveguide routing layers. Inter-layer couplers are used to communicate between these layers.

Shacham *et al*. [84] have proposed a circuit-switched 2D torus network that uses an electrical control plane with an optical data plane. The electrical control plane sets up the optical switches in the data plane before data transmission takes place and tears down the network paths thereafter. This network must transmit large amount of data to amortize the high latency of electrical path setup and tear down making this design unsuitable for cache coherent shared memory traffic where the biggest unit of transfer is a cache line.

Cianchetti *et al*. [21] have proposed a hybrid opto-electrical network-on-chip architecture called phastlane. The optical components of the phastlane network are integrated on a separate chip and packaged with the processor die via 3D integration. Microring resonators are employed in the optical switches to perform the switching functionality. Control signals that setup the ring resonators are sent optically along with the data. If a conflict arises between two incoming packets intending to go out of the same output port, then only one of the packets is forwarded through the optical switch. The other packet is converted into electrical form and buffered. This packet is then retried at a later time. The optical switches in the phastlane architecture employ a large amount of waveguide crossings which can lead to significant optical loss (see chapter 2).

Current state of the art optical switches incur a loss of about 3.0dB [25]. Furthermore, each waveguide crossing in the network causes a loss of about 0.2dB [49]. Such high losses lead to considerable laser power consumption in optically-switched networks.

3.1.2.2 Electrical Switching

In electrically-switched photonic networks, the switching elements (routers) are electrical while the links are optical. Electrically-switched silicon photonic networks have so far received the least attention in literature. This is because the conventional wisdom behind

initial photonic designs was that electrical-to-optical (E/O) and optical-to-electrical (O/E) conversions along a communication-path would incur significant dynamic energy costs. However, as recent papers have demonstrated [49, 109], the energy consumption in a photonic network is dominated by the static components, specifically laser power and ring resonator tuning power. Therefore, electrically-switched photonic networks represent a viable option for deployment in the near-term. In fact, one of the key contributions of this dissertation is an in-depth evaluation of this category of silicon photonic networks (see chapter 5).

Joshi *et al*. [40] have proposed a 3-stage clos network design that uses electrical routers and photonic channels. This architecture employs a U-shaped layout for the network waveguides to avoid waveguide crossings. This paper only compares their clos architecture against a mesh network and ignores popular topologies such as fat-tree [57] or flattened-butterfly [43] that offer higher path diversity in the network leading to superior throughput performance. An exhaustive evaluation of these popular topologies within the constraints of silicon photonic technology is provided in this dissertation in chapter 5.

3.1.3 Thesis Contributions

Silicon photonic technology offers integration of multiple chips to provide high performance, improved yields and lower costs (energy-per-bit). This has inspired architects to explore different networking strategies for adoption in silicon photonic networks. Popular approaches in literature include optical crossbars (channel sharing) and switched networks as discussed earlier. This dissertation makes contributions in both these categories of optical networks. A brief description of these contributions is provided below.

3.1.3.1 Contributions in Channel Sharing Approach to Photonic Networks

Analytical model to quantify the limits and gains of channel sharing: The fully-connected, point-to-point (P2P) network falls into the SWSR category of channel sharing topologies. The P2P topology employs the fewest number of optical devices along a network link. Hence, this topology suffers from the lowest laser power consumption cost per link. The other categories of channel sharing topologies (SWMR, MWSR, and MWMR) provide higher performance but also exacerbate the laser power consumption in the photonic network by requiring higher number of optical devices along a wavelength path. To explore this trade-off, an analytical model is developed in this dissertation that quantifies the limits and potential gains of channel sharing techniques given a laser power budget. This analytical model is described in detail in chapter 4.

Novel arbitration-free channel sharing architecture called 'wavelength stealing': All channel sharing topologies that incorporate multiple senders and/or receivers per link (SWMR, MWSR, and MWMR) require some form of arbitration at the source and/or at the destination depending on how the sharing is implemented. As surveyed above, a number of arbitration techniques in shared networks [78, 76, 96, 95, 100] have been proposed. Each of these techniques suit different topologies and differ in complexity and latency overheads. This dissertation introduces the 'wavelength stealing' architecture which implements an MWSR-type sharing over a fully-connected point-to-point (P2P) topology and avoids arbitration completely by using a novel aggressive channel-stealing mechanism with graceful recovery from collisions. The wavelength stealing architecture is presented in chapter 4. Evaluation results show that this arbitration-free architecture exhibits lower latency and better throughput performance compared to traditional arbitration-based architectures.

3.1.3.2 Contributions in Switched Photonic Network Topologies

Thorough evaluation of electrical switching within the purview of photonic technology: Switched networks are pervasive in computer systems. In silicon photonic networks, the switching elements (routers) can be optical or electrical. In the current technology generation, optical switches incur a significant optical loss as discussed earlier. Koka *et al*. in a recent ISCA paper [49] show that significant breakthroughs are required in device technology to make optical switching viable. This leaves electrical switching as an alternative approach to adopt in switched photonic networks. Electrically-switched silicon photonic networks do not exacerbate the laser power consumption by incorporating many ring resonators along a waveguide nor do they require high loss components such as optical switches. Thus, on the static (laser and ring tuning) power front, these networks offer an inherent advantage over optical switching. However, as discussed earlier, electrically-switched photonic designs have so far received minimal attention in literature. The goal of this thesis is to fill this gap. In this vein, this dissertation presents an in depth evaluation of electrically-switched network designs within the constraints imposed by silicon photonic technology in chapter 5. Many popular topologies, both low- and high- radix, are investigated and it is shown that a fully-connected network capable of adaptive routing provides the highest performance.

3.2 Microarchitecture of High-Radix Routers

The topology of a network has a profound impact on system performance. The trend in recent years indicates a push towards high-radix networks [44]. In high-radix topologies, the routers of the network are built using many narrow ports as opposed to low-radix routers which have fewer ports which are wider. Moving to a high-radix topology increases the overall 'connectedness' of the network. That is, in a high-radix network, packets are able

to reach their destinations with fewer hops resulting in higher performance compared to a low-radix network. Furthermore, increasing the radix reduces switch count in the network leading to lower power costs [44]. This increased performance together with lower costs is driving the scaling trend of routers towards higher radix [44]. Examples of high-radix topologies include butterfly [23], fat-tree [57] and flattened-butterfly [43] topology. The Thinking Machine CM-5 [58] and the Cray BlackWidow vector multiprocessor [82] both use the high-radix fat-tree topology for their network.

Increasing the port count (radix) leads to a significant increase in the complexity of the router microarchitecture. Traditional router designs employ three main *logic* structures: two allocators and a crossbar switch. The complexity of these logic structures scales *quadratically* with the number of ports making high-radix router design particularly challenging. In recent years, there have been a variety of proposals [44, 2, 16] that address the scalability concerns of high-radix routers. These prior proposals are surveyed below.

3.2.1 Prior Work

The progression of a packet through a traditional router pipeline involves two allocations followed by traversal through a crossbar switch (more details are provided in chapter 5). The two allocations are performed in the following order: 'virtual-channel allocation' (VA) followed by 'switch allocation' (SA). Two allocator circuits are employed to perform these allocations. The complexity of a traditional switch allocator scales as $O(k^2)$ where k is the number of ports of the router. To address the quadratic increase in complexity with router radix, Kim *et al*. [44] have proposed a distributed switch allocator design that breaks down the allocation process into three stages. For the first two stages, the arbitration decisions are made locally over just a small subset of inputs such that each stage can fit into the router clock cycle. Then the distributed request signals are collected via global wiring to perform the final stage of allocation. The winning requesters are notified by propagating the grant

signals back after the final allocation stage.

The complexity of a traditional virtual-channel allocator circuit poses even a bigger problem then the switch allocator. This is because the virtual-channel allocator logic scales as $O(k^2v^2)$, where v is the number of buffer queues (virtual channels) per router port. This new multiplicative term, v^2, exacerbates the complexity concerns and makes scaling the virtual channel allocator prohibitively expensive. The authors in [44] adopt slightly modified versions of their distributed switch-allocator circuit for performing virtual-channel allocation. For better scalability, they incorporate speculation as well where virtual channel allocation is performed deeper in the pipeline.

To facilitate distributed allocation, the authors in [44] propose incorporating intermediate buffering in the crossbar switch. However, incorporating buffering at each crosspoint granularity becomes prohibitively expensive in terms of storage. To reduce this complexity, the authors in [44] propose a hierarchical switch organization which partitions the crossbar into many smaller subcrossbars with buffering applied to only the inputs and outputs of these subswitches leading to a significant reduction in the storage overheads.

In a recent paper, Ahn *et al*. [2] argue that the hierarchical switch organization proposed by [44] has limited scalability due to the power and area overheads of the wires and intermediate buffering. Instead the authors of this paper propose a network within a switch microarchitecture design approach for building high-radix routers. In other words, this work proposes replacing the crossbar switch employed inside the routers with a network itself. They call this intra-router network as the *local* network and the inter-router topology as the *global* network. The authors evaluate different topologies for the local (intra-router) network such as fattree [57], 2D torus [23], and 2D HyperX [3]. Their evaluation results show that the fattree local topology provides the lowest area and power consumption. However, employing a network topology as a replacement for the crossbar switch incurs high performance cost in terms of latency.

Binkert *et al*. in their paper [16] also highlight the scalability concerns of the crossbar switch within a traditional high-radix router. To improve scalability, the authors in this work propose an optical switch architecture that employs MWSR photonic channels (see section 3.1.1.3) to build a flat crossbar switch. The main components of their high-radix router are input and output buffers along with an optical crossbar. Packets arriving into the router via input fibers are immediately converted into electrical form and are buffered at the input. There they undergo optical arbitration to gain access to the optical crossbar switch. This work borrows the *token channel* scheme proposed in [96] for performing the optical arbitration. The optical crossbar employs a separate MWSR channel for each output port and all the input ports are allowed to write on it (after winning arbitration). Once packets traverse the crossbar and arrive at an output port, they are put in buffers until they can leave the router. One of the biggest concerns about this optical crossbar design is the high laser power consumption of MWSR channels as discussed in section 3.1.1.

3.2.2 Thesis Contributions

Novel router architectures that do not employ allocators or a full crossbar switch: This dissertation provides an in-depth evaluation of electrically-switched photonic network designs and demonstrates in chapter 5 that the fully-connected, point-to-point (P2P) topology provides the highest throughput performance . However, designing the router presents a significant challenge for the fully-connected topology as it is the highest-radix network that can be constructed to interconnect a set of nodes. Specifically, for an N node network, the router in a fully-connected topology has N ports. As discussed earlier, the complexity of the main logic elements (allocators and a crossbar switch) employed in a traditional router scales quadratically with the number of ports. Hence, employing a traditional router design for the fully-connected topology becomes prohibitively expensive as its area and power consumption scales as $O(N^2)$, i.e quadratically with the network size. Instead, this

dissertation adopts a 'topology-aware' router design approach and demonstrates that low cost routers can be designed that employ simple arbiters instead of allocators and do not require full crossbar switches. These innovative router designs are presented in chapter 5.

3.3 Quality-of-Service (QoS) guarantees

Broadly speaking, quality-of-service (QoS) refers to mechanisms or techniques for providing service guarantees in the network. QoS guarantees become essential when many entities (e.g. nodes or VMs) share the network and the available resources (e.g. bandwidth) are limited. In this case, it is the job of the QoS mechanism to allocate resources to the clients or entities according to some fairness criterion or service level agreement (SLA). Popular uses of QoS include providing *performance isolation* and *differentiated service*. Performance isolation guarantee ensures that the resource utilization of one entity should not impact the service promised to another entity sharing the network. Differentiated QoS enables a resource to be allocated in varied proportions.

Two recent papers [75, 77] have proposed mechanisms for incorporating QoS guarantees in photonic networks. This prior work is surveyed below.

3.3.1 Prior Work

Ouyang *et al*. [75] have proposed a QoS mechanism for their photonic network-on-chip (NoC) architecture that provides bandwidth differentiation in the network. This work adopts the 'token slot' scheme proposed by [96] (discussed in section 3.1.1.3) along with MWSR channels as their baseline architecture. On top of this baseline, they incorporate two additional optical rings per channel: a 'completion' ring and a 'frame-switching' ring. These additional optical components are used to implement 'frame-based' arbitration that enables differentiated QoS support in the network. In frame-based arbitration, a

frame is composed of multiple packets. Different shares of this frame are allocated to the source nodes according to the desired differentiation levels. A frame number is associated with every frame. A source node keeps pushing its packets into a frame as long as it does not exceed its assigned share. Once this node reaches its share limit, it pushes any subsequent packets to the next frame. Frame-based arbitration has two requirements: 1) packets belonging to the same frame must be delivered before the next frame can start; and, 2) frames are delivered in the order of increasing frame numbers. By enforcing these requirements, it is ensured that the share guarantees of the source nodes are upheld. The biggest drawback of this architecture is that the frame-switching ring implements optical broadcast that consumes significant laser power consumption.

The design proposed by Pan *et al*. [77] does not require optical broadcast to incorporate differentiated QoS guarantees. Instead, this paper proposes an adaptive feedback mechanism that throttles the source nodes to maintain the desired bandwidth guarantees. This design also adopts the token slot scheme proposed by [96] together with MWSR channels. A QoS controller is integrated with every MWSR channel. Each source node that shares a MWSR channel maintains token consumption information which is communicated to the QoS controller at the end of a communication epoch. Using the received token consumption information, the QoS controller calculates the token quota required by each source node for the next communication epoch to maintain the desired bandwidth guarantees. The QoS controller then conveys the assigned quotas to the source nodes. The key insight employed by this architecture is the use of data channels to exchange QoS information instead of incorporating separate control wavelengths just for this purpose. This reduces implementation overheads provided the data channels are *wide*. If the data channels are narrow, then periodically exchanging QoS information can significantly hurt the network throughput. Another disadvantage of this architecture is that it can only be applied to arbitration based networks.

3.3.2 Thesis Contributions

Novel mechanisms for incorporating QoS guarantees in multichip photonic networks: This dissertation provides novel QoS mechanisms for providing performance isolation (see chapter 4) and service differentiation (see chapter 5) in multichip photonic networks. The proposed techniques differ from prior work in several ways. First, since this dissertation focuses on multichip systems, the scale of the network is much larger compared to the NoC designs proposed in prior papers. Thus, QoS guarantees are enforced at the granularity of virtual machines (VMs) instead of individual nodes in the network. Furthermore, the techniques proposed in this thesis rely on the hypervisor to configure the network resources and map VMs to the appropriate network regions to ensure that the desired QoS guarantees are enforced. Finally, prior work on QoS techniques in photonic networks employ arbitration based network designs. Since, the network architectures proposed in this dissertation are either arbitration-free or use electrical switching, these earlier approaches are not applicable to the designs presented in this thesis. The novel QoS mechanisms proposed in this dissertation are discussed in detail in chapters 4 and 5.

3.4 Summary

This chapter surveyed the most relevant prior work related to this thesis. The chapter started by discussing different topology architectures in photonic networks. Channel sharing topologies were the first category of networks considered. Channel sharing designs are 1-hop networks. It was shown that prior channel sharing architectures that incorporate sharing either at the sender side and/ or receiver side incorporate some sort of arbitration mechanism. Furthermore, it was highlighted that the high laser power consumption per link in these architectures pose a significant challenge in terms of network design. Next, prior work on switched photonic networks was surveyed. It was highlighted that the high

loss of optical switches in the current technology generation raises concerns about the laser power consumption in optically-switched photonic networks. The chapter then delved into the complexities of router design for high-radix networks. A brief overview of the design approaches proposed in prior papers for alleviating these complexity concerns was presented. Finally, a brief description of mechanisms for incorporating QoS guarantees in photonic networks proposed in literature was presented. Throughout this chapter, the contributions made by this dissertation were briefly highlighted to illustrate where they fit in the landscape of previous work.

4 CHANNEL SHARING IN PHOTONIC NETWORKS

The topology of the interconnect has a profound impact on network performance. This chapter investigates channel sharing topologies in photonic networks. Channel sharing designs, or optical crossbars as they are sometimes called, are 1-hop networks. That is, channel sharing architectures minimize the number of E/O and O/E conversions required to communicate a message to a destination. The simplest channel sharing network is a fully-connected, point-to-point (P2P) topology. A P2P network statically partitions the total network bandwidth (wavelengths) between the sender-destination pairs leading to relatively low bandwidth (narrow) node-to-node channels. On the other hand, a network that enables sharing combines wavelengths to form a single logical high bandwidth (wide) shared channel. Thus, sharing-based networks can potentially provide higher node-to-node bandwidths compared to a P2P network, albeit at the cost of arbitration delays in accessing the shared channel. The peak node-to-node bandwidth of a channel is proportional to the following terms:

$$\text{Node-to-Node bandwidth} \propto s \times \mathit{Eff}(s) \times \frac{\mathit{Total\ wavelengths}}{N^2} \tag{4.1}$$

where, N is the total network nodes, s is the sharing degree ($s \geqslant 1$), and $\mathit{Eff}(s)$ signifies the efficiency of sharing ($\mathit{Eff}(s) \in [0,1]$). This fractional term $\mathit{Eff}(s)$ captures the costs associated with sharing, e.g. overheads of arbitration, fairness, etc. $\mathit{Eff}(s)$ is inversely proportional to the sharing degree s due to higher overheads (e.g. contention). Sharing ($s > 1$) can provide higher bandwidths compared to a P2P network ($s = 1$) as long as the costs do not outweigh the benefits i.e. $s \times \mathit{Eff}(s) > 1$. In addition to the efficiency penalty however, there is anothor significent cost associated with sharing in photonic networks: *high static power consumption*.

Photonic networks based on ring resonators are static power dominated because of laser

power and ring tuning power [76, 49, 48]. A higher degree of sharing requires more devices along a wavelength, thereby increasing the required input laser power (optical) and the device tuning power (electrical). Efficiencies of commercially available WDM lasers are $1-5\%$, and may be expected to exceed 10% in the next decade [108, 14, 54]. When laser efficiency is considered, laser power becomes the dominant contributer to static power dissipation. Thus, optimizing for laser power must be considered a first-order design constraint. The laser power consumption of a photonic network is proportional to the following terms:

$$\text{Laser Power Consumption} \propto \text{Total}\,wavelengths \times \overbrace{\underbrace{\frac{\#\ devices}{wavelength}}_{\text{Increases with sharing s}} \times \frac{\text{loss}}{\text{device}}}^{\text{Avg. loss per wavelength}} \tag{4.2}$$

This thesis uses the power-constrained design approach described in [49] which assumes a *fixed* input laser power budget for all designs under consideration. This constraint ensures that any performance gains that arise from sharing do not come with the costs of increased laser power consumption. Equating the laser power consumption of a sharing design to a P2P network using Eq.(4.2) leads to the observation that:

$$\text{Total}\ wavelengths_{\text{sharing}} < \text{Total}\ wavelengths_{\text{P2P}}$$

Thus it is clear that the total peak bandwidth of a network with wavelength sharing will be lower than that of an energy-equivalent point-to-point network. If this sharing design can still provide higher node-to-node bandwidth (Eq.(4.1)) even with fewer total wavelengths, then it may be the preferred design choice over a P2P network depending on the target applications. Thus, a sharing design can win on performance (bandwidth) *and* power

(laser) only when:

$$\underbrace{s \times Eff(s)}_{(>1)} \times \underbrace{\frac{Total\ wavelengths_{sharing}}{Total\ wavelengths_{P2P}}}_{(<1)} > 1 \tag{4.3}$$

Most prior sharing-based proposals have assumed very *aggressive* values for device losses. This has led to designs in which sharing has had negligible impact on the average loss per wavelength in Eq.(4.2) leading to $\frac{Total\ wavelengths_{sharing}}{Total\ wavelengths_{P2P}} \approx 1$. With no penalty from this ratio term in Eq.(4.3), these designs have pushed sharing to very high levels (e.g. $s = 64$) and have shown significant performance gains with minimal impact on laser power consumption.

This dissertation models the impacts of *conservative* loss assumptions on photonic network design and makes the following contributions in this chapter:

- An analytical model to determine the limits and gains of sharing,
- The design of a novel arbitration-free, energy efficient shared channel network architecture, called *wavelength stealing*,
- Detailed performance evaluation of the wavelength stealing architecture implementation on a single-layer wafer-scale multichip system, and
- Application of the wavelength stealing architecture to improve the network throughput of a partitioned multichip cluster using a smart hypervisor.

The rest of the chapter is organized as follows. Section 4.1 discusses the additional losses that arise due to sharing and quantifies the performance gains achievable by a sharing design. Section 4.2 presents a novel sharing-based design called wavelength stealing. Implementation of wavelength stealing on a multichip system is discussed in section 4.3, and the application of wavelength stealing architecture to support multiple virtual machines is presented in section 4.4. Section 4.5 discusses the evaluation methodology and

results, and section 4.6 concludes this chapter. Most of the work presented in this chapter also appears in my MICRO paper [109].

4.1 Sharing in Photonic Networks

4.1.1 Ring Modulator Losses

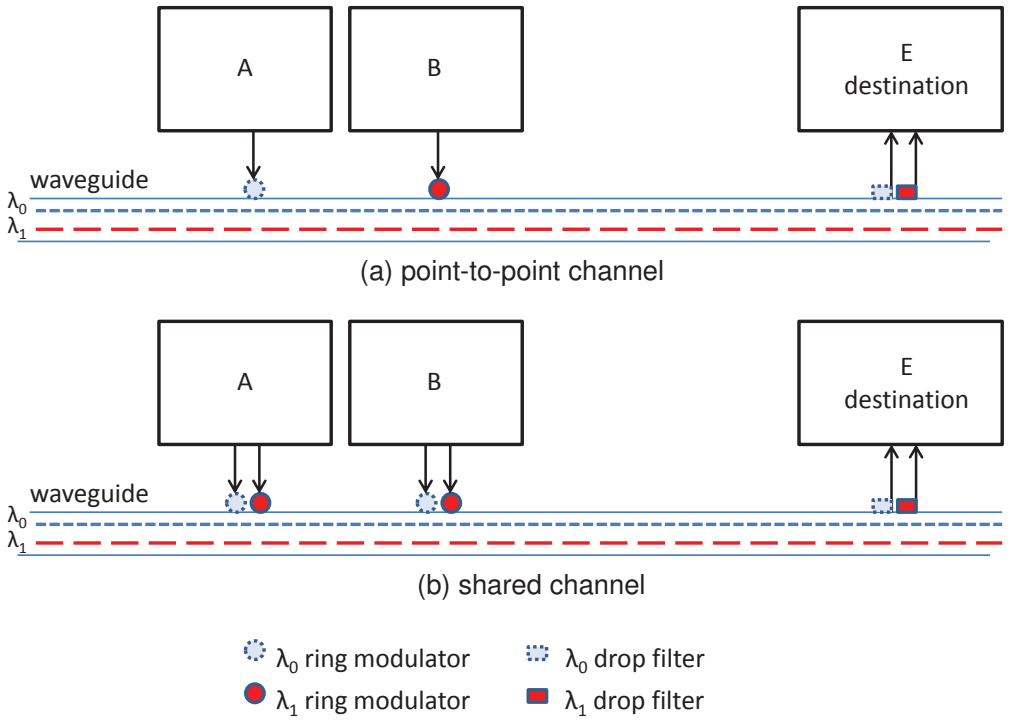


Figure 4.1: A point-to-point (P2P) versus shared channel. Due to extra modulator rings, light on a shared wavelength suffers from higher losses.

Figure 4.1a shows a waveguide carrying two wavelengths in a point-to-point topology where source nodes 'A' and 'B' modulate different wavelengths to destination node 'E'. Each modulator ring placed along the waveguide is tuned to a specific wavelength and modulates light on that wavelength. Modulation is controlled by electrically biasing the ring using the data stream to either pass light (transmit a '1') or absorb light (transmit a '0'). An active ring (that modulates a wavelength) causes a significant insertion loss of 4.0dB

to the wavelength. As shown in the figure, the wavelength of light also passes by rings that are tuned to other wavelengths of a waveguide. These rings cause a smaller passive through-loss of 0.05dB per ring.

4.1.2 Wavelength Sharing

Figure 4.1b shows a waveguide carrying two wavelengths that are shared by two senders 'A' and 'B' to a destination node 'E'. Each node sharing a wavelength has a ring along the waveguide tuned to that wavelength. Thus in figure 4.1b, each wavelength passes by twice as many rings compared to a wavelength in the point-to-point channel. Multiple active rings on a wavelength will significantly increase the loss even though only one of them would be transmitting data. To achieve lower loss, a ring can be detuned dynamically away from the target wavelength as long as it is not transmitting data. However, due to the fast response times required, it is not feasible to detune a ring far enough from the target wavelength to make the loss negligible. Even with aggressive device techniques, we can expect a loss of 0.5dB per detuned (inactive) ring. In this work, it is assumed that tuning or de-tuning the microrings will occur in one bit time. This is an *aggressive* device technology goal and is under investigation.

4.1.3 Sharing Gains

From figure 4.1b, it is evident that wavelength sharing increases the link loss. This section explores the limits on sharing imposed by these additional losses. By extending the topology shown in figure 4.1b to sharing degree s and WDM factor w, the additional

optical power loss of a shared wavelength compared to the P2P wavelength becomes:

$$\begin{aligned} \triangle L_{dB}(\lambda) &= Loss_{sharing} - Loss_{P2P} \\ &= (s-1)[\underbrace{0.5dB}_{\text{inactive rings}} + \underbrace{(w-1)0.05dB}_{\text{other } \lambda \text{ rings}}] \end{aligned} \tag{4.4}$$

Now, the amount of laser power consumed by $W_{sharing}$ wavelengths in a shared design and W_{P2P} unshared wavelengths in the P2P design is given by:

$$P_{sharing} = W_{sharing} \times 10^{(P_{rx} + \triangle L_{dB}(\lambda) + Loss_{P2P})/10}$$

$$P_{P2P} = W_{P2P} \times 10^{(P_{rx} + Loss_{P2P})/10}$$

By equating these two equations, the number of unshared wavelengths that consume *equivalent laser power* to a given number of shared wavelengths can be expressed as:

$$W_{P2P} = W_{sharing} \times 10^{\triangle L_{dB}(\lambda)/10} \tag{4.5}$$

This equation clearly shows that under the equivalent laser power constraint, the unshared P2P network can support higher number of wavelengths and hence offers higher total bandwidth (capacity) than a shared design. However, sharing can lead to higher node-to-node bandwidths over the P2P network provided there is no contention on the shared channel. These node-to-node bandwidth gains are quantified below.

Let us define $Speedup_{ideal}$ to be the ratio of time taken by a message of size $messagesize$ to be delivered to a destination on a P2P (unshared) channel versus time taken on a shared channel. It can be computed as:

$$Speedup_{ideal} = \frac{\left[\frac{messagesize}{W_{P2P}} + T_{prop}\right]}{\left[\frac{messagesize}{s \times W_{sharing}} + T_{prop}\right]} \tag{4.6}$$

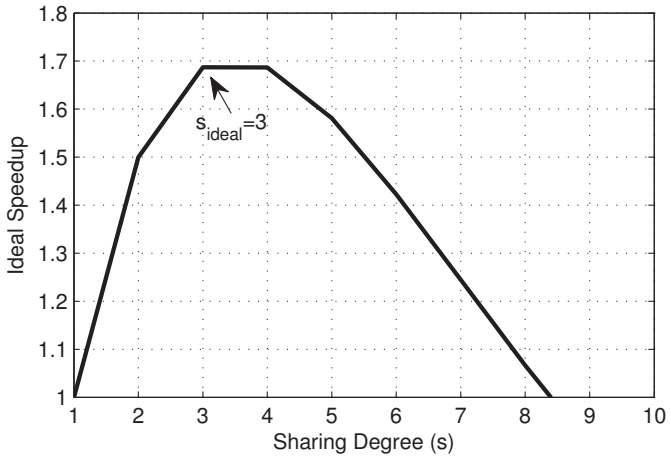


Figure 4.2: Ideal speedup versus sharing degree s assuming $w = W_{sharing} = 16$ and $T_{prop} = 0$.

where T_{prop} is the propagation time between the sender and destination. This definition of speedup is called 'ideal' because it does not associate any overheads (in terms of time or wavelengths) with sharing.

Figure 4.2 shows the ideal sharing gains achievable as a function of sharing s assuming 16-way WDM waveguides. From Figure 4.2 and Eq.(4.6), the following observations can be made:

- The ideal achievable speedup is independent of message size assuming $T_{prop} = 0$. This is because the $messagesize$ term in the numerator and denominator simply cancel each other out in Eq.(4.6).

- Wavelength sharing is only effective at low sharing degrees. In fact, ignoring all overheads of sharing, the optimal[1] sharing degree is just 3 (s_{ideal}).

- Beyond the optimal point, the number of wavelengths in the shared channels decreases significantly leading to a drop in the achievable speedup.

[1] defined as the lowest sharing degree with the highest speedup value.

4.2 Wavelength Stealing Architecture

This section presents a novel interconnection architecture for multichip systems called *wavelength stealing*.

4.2.1 Design Overview

The topology of the wavelength stealing interconnect is similar to that of a point-to-point (P2P) network. Each node in the system has a dedicated channel (one or more waveguides) to every other node in the system and is called the 'owner' of that channel. The owner has non-blocking access to send information to a destination using its dedicated channel and is always guaranteed service on that channel. In addition to its dedicated channel, the sender can also steal access to channels owned by other senders to that destination. However, access to this additional (stolen) bandwidth is not guaranteed. Figure 4.3 shows an example where node 'B' has a dedicated (owned) channel to destination 'E' and can also steal on the channel owned by node 'A'. Similarly node B's dedicated channel to E can be stolen by another node 'C'. Hence every channel in the system is owned by one node and can be stolen by one or more other nodes. Stealing is performed arbitration-free (without notification to the owner or other stealers). Any errors (collisions) that arise from stealing are corrected at the destination using mechanisms described in later sections. Stealing is a form of wavelength sharing and is accomplished by placing additional modulator rings along the shared waveguide as shown in figure 4.3. These additional rings cause higher wavelength losses (as described in section 4.1). Hence to match the laser power budget of a P2P network, the wavelength stealing architecture will have to use fewer wavelengths per channel than the P2P network. However, it can still provide higher node-to-node bandwidths than the P2P network provided stealing access on other channels is successful.

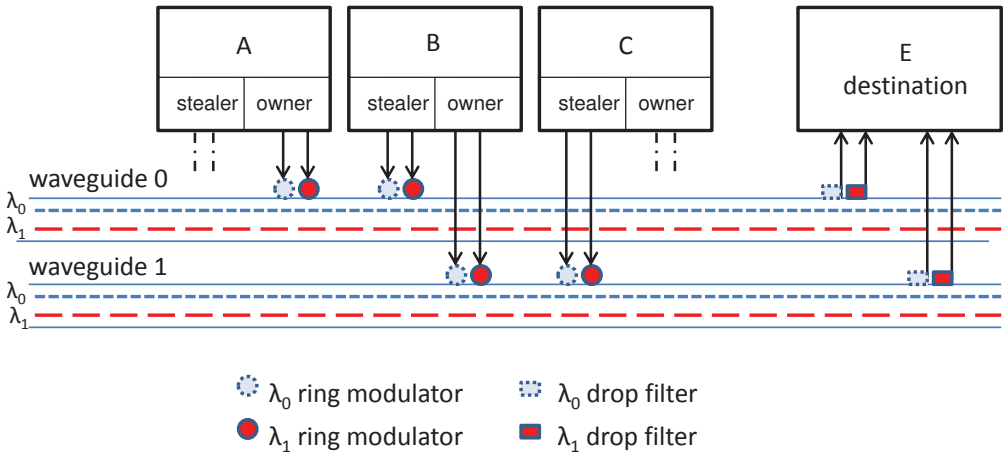


Figure 4.3: A 2-way wavelength stealing design example showing sender B's channels to destination E. Sender B can send 2bits/cycle guaranteed on its (owned) channel to E, and can opportunistically steal bandwidth on A's channel to send 2 extra bits/cycle provided A is not using its channel. Note that this figure does not show the stealing channel of sender A and the owned channel of sender C to destination E.

4.2.2 Implementation Details

For correct operation, an implementation of the wavelength stealing design should satisfy some strict requirements:

1. The owner must be guaranteed non-blocking access without any arbitration delays.

2. A stealer can steal bandwidth without arbitration (no prior notification to the owner or other stealers) and should be notified if it needs to stop stealing.

3. The destination must be notified if a received phit is corrupted due to collision and must be able to correct the bit errors. On receiving a valid phit, the destination must be able to identify the sender of the phit.

To meet the above requirements, the wavelength stealing architecture employs erasure coding [91] and special control wavelengths per channel. For simplicity, the rest of this section assumes only one stealer per channel.

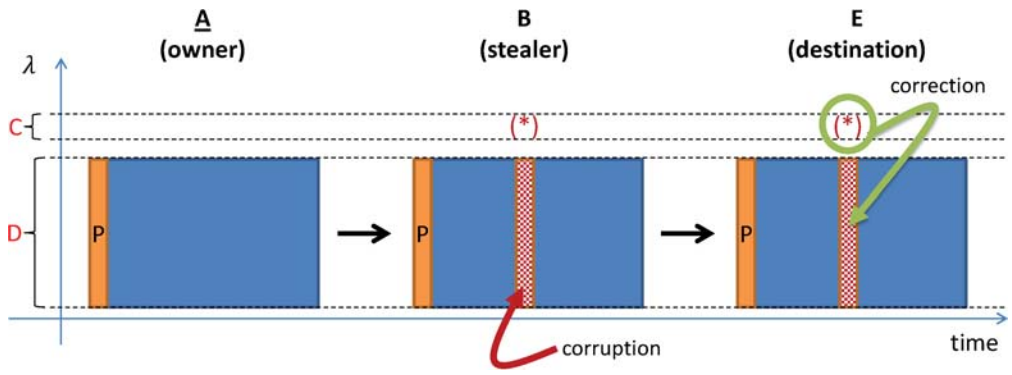


Figure 4.4: Erasure coding example. Corruption in A's message due to a collision from B gets marked (∗) in the control wavelengths. This location information is used to perform erasure correction at the destination.

4.2.2.1 Erasure Coding

In the wavelength stealing architecture, a stealer is allowed to steal (use wavelengths) on a channel without prior notification to the owner (i.e., it is arbitration-free). In this case, whenever a stealer steals on a channel on which the owner is actively sending data, a collision occurs, causing errors in the owner's message. These errors are corrected at the destination using erasure coding. When a collision occurs, a stealer is notified by the control wavelengths to stop stealing, preventing further errors in the owner's message. This ensures that an owner's message is never corrupted beyond the point of recovery. Erasure codes rely on location information of potential errors to provide better correction capability than codes that correct random bit errors [31]. For example, with location information, a parity code is capable of *correcting* a single bit error. In the case of multi-bit errors, stronger erasure codes can be employed [31].

Figure 4.4 shows a channel in the wavelength stealing architecture with associated data wavelengths (indicated by D on the y-axis) and control wavelengths (indicated by C on the y-axis). The owner's message (A) has a parity column appended to it. As this message goes past the stealer (B), B steals on the owner's channel leading to an error. This error is automatically marked (discussed later) in the control wavelengths (∗). A stealer detects

collisions with the help of the control wavelengths and stops stealing to prevent further errors. The corrupted message arrives at the destination (E) where the computed parities are compared with the parity column in the message. If there is a parity mismatch, the corresponding bits at the marked location are inverted to correct the bits in error.

It is important to emphasize that if no errors are marked in the control wavelengths then a received message is completely error-free and the destination doesn't need to wait for the subsequently arriving parity bits. Thus, in the absence of contention at low loads, the latency overhead of accessing the shared channel is completely hidden and messages only experience minimal latencies[2] which is not possible in a design based on arbitration.

4.2.2.2 Control Wavelengths - Two Designs

The control mechanism for wavelength stealing can be implemented using one of two designs, called abort and sense. These designs exhibit different trade-offs but provide the following functionality:

1. Mark the location of corrupted bits for erasure correction at the destination.
2. Inform stealer to stop stealing when the owner becomes active to limit the corruption to a single bit collision.
3. Inform destination of the ID (owner's, stealer's, or corrupted) of the received communication (phit).

Abort Design: Figure 4.5 shows a channel consisting of one waveguide to destination 'E' owned by sender 'A' with a stealer 'B'. $\overline{Owner}$ and $\overline{Stealer}$ are the control wavelengths and $D0 - D13$ are the data wavelengths in the waveguide. The behavior of the control wavelengths in the abort design is given in table 4.1. When the owner (A) is not using

[2]There are no latency overheads beyond message serialization delay and propagation delay.

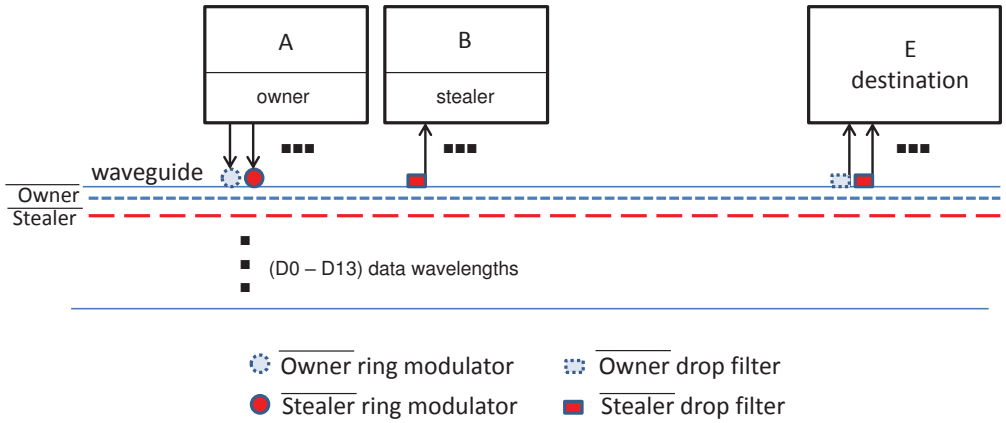


Figure 4.5: Abort control wavelengths.

Active Sender	A		B	E		Received
	$\overline{Own.}$	$\overline{St.}$	$\overline{St.}$	$\overline{Own.}$	$\overline{St.}$	
A	0	1	–	0	1	A
B	1	0	0	1	0	B
A, B	0	1	1	0	0	Collision
(Invalid)	1	1	–	1	1	(Invalid)

Table 4.1: Abort design functionality for owner (A), stealer (B) and destination (E). (The values 11 should not arise during normal system operation.)

the channel, it transmits a continuous 10 on the control wavelengths $\overline{Owner}$ and $\overline{Stealer}$ respectively. If the owner (A) uses the channel, it transmits a continuous 01 on the two control wavelengths. When the stealer (B) needs to transmit data to E it begins data transmission on its dedicated channel to E and steals the channel owned by A. Sender B also turns on the drop filter on the $\overline{Stealer}$ wavelength. The drop filter pulls out all light (bits) traveling on the control wavelength. If a value of 0 is read by the drop filter, then the stealer (B) knows that there has not been a collision with the owner. If the drop filter reads a value of 1, then the stealer (B) knows that a collision has just occurred. It then suspends stealing, but continues to use its dedicated channel to E. At the destination side, a 01 indicates owner's (A) phit, a 10 indicates stealer's (B) phit and a 00 represents a corrupted (collided) phit. The destination tracks the control wavelength information to

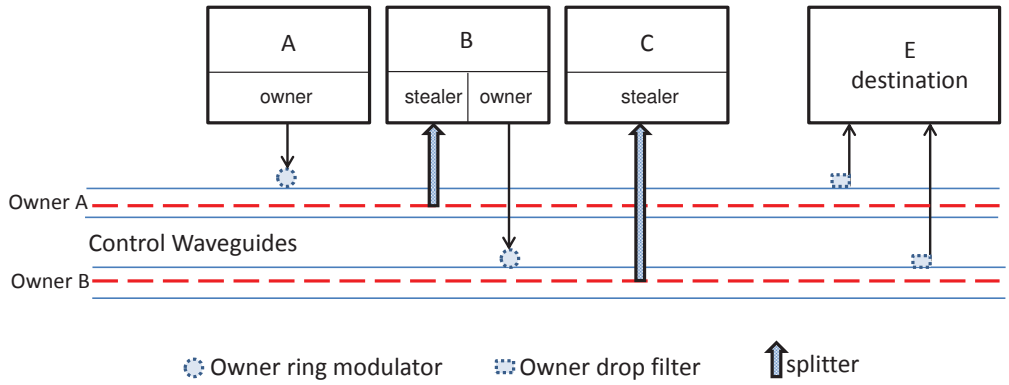


Figure 4.6: Sense control waveguides.

perform the protocol steps discussed in section 4.2.2.3.

Sense Design: The sense design requires separate waveguides for control and data. The control waveguides of two owner channels 'A' and 'B' are as shown in figure 4.6. The need for separate waveguides arises because this design uses optical splitters which are fabricated as broadband devices that split all wavelengths in a waveguide. Since the splitting functionality is only required for the control wavelengths, they are placed in waveguides that are separate from the data wavelengths. There is only one control wavelength per control waveguide, called *Owner* and abbreviated as '*OW*' in the rest of the discussion. The control wavelengths for the owner A's channel and owner B's channel are denoted by $OW(A)$ and $OW(B)$ respectively. Some useful terminology is defined in figure 4.7.

In the sense design, the control functionality of the owner (A), stealer (B) and destination (E) depends on both the current and previous values (state) of the control wavelengths (*OW*) as shown in the state machine diagrams in figure 4.7. The state machine diagram for owner (A) shows that whenever A uses its channel, it puts a continuous 1 on $OW(A)$. The operation of the stealer (B) then depends on the value of $OW(A)$. From the stealer's (B) state machine, it is clear that it can be in one of two states when it has a message to send: STEAL or SENSE. In the STEAL state, the stealer (B) can actively steal on the owner's (A) channel. Now, if the owner becomes active ($OW(A) == 1$), then the stealer (B) transitions

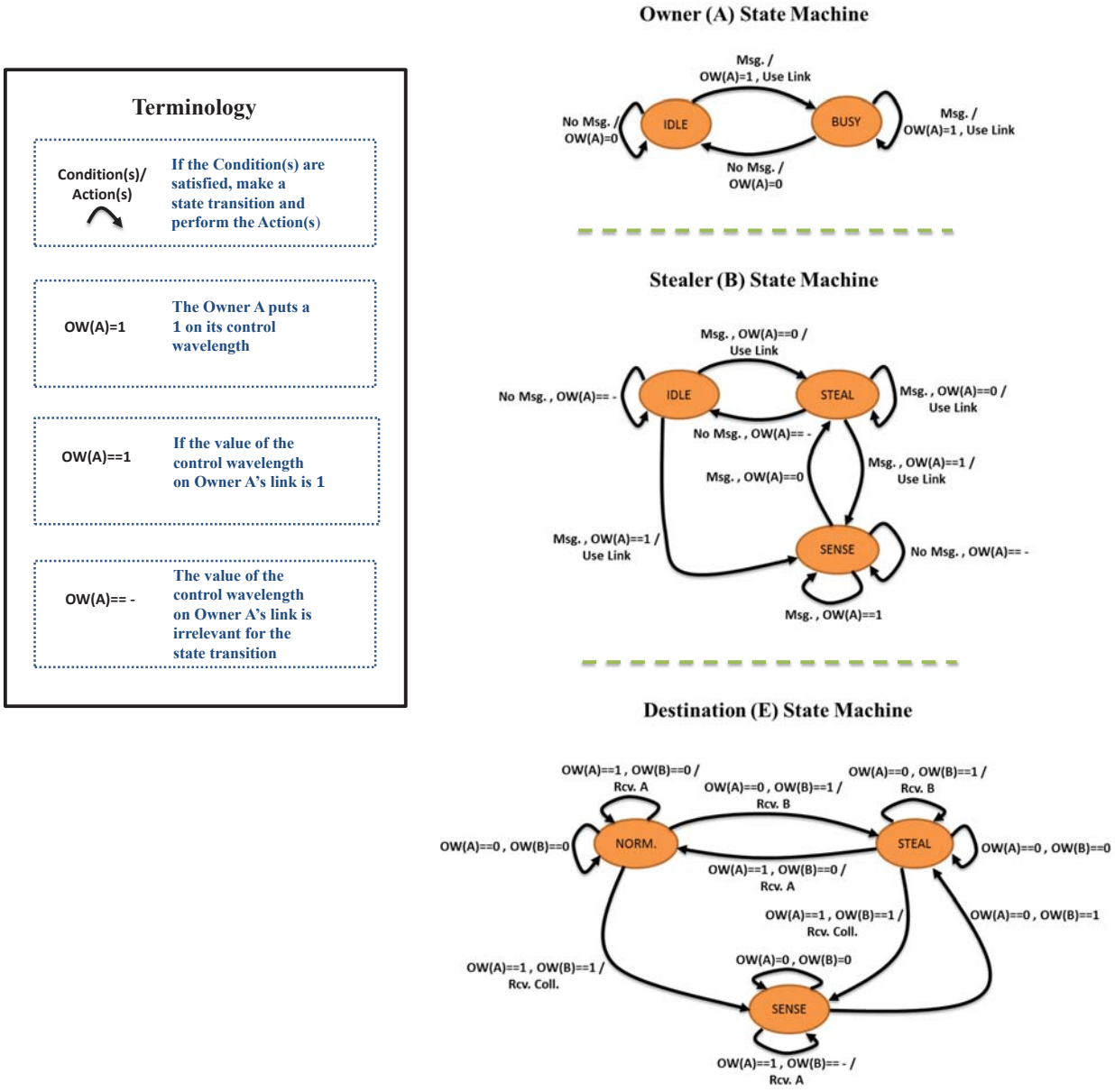


Figure 4.7: Sense design functionality for owner (A), stealer (B) and destination (E).

to the SENSE state. While in this state, the stealer does not steal and simply waits for an opening on the owner's (A) channel so that it can revert to stealing.

Note that the destination state machine needs to monitor the control wavelength of both the owner (A) and the stealer (B) to function properly. If the destination observes both $OW(A) == 1$ and $OW(B) == 1$, it knows that a collision has occurred. The destination

then transitions into the SENSE state. While in the SENSE state, the only valid phit that is received is from the owner (A). The rest of the functionality in these state machines (figure 4.7) is self explanatory.

Abort vs. Sense Trade-offs: The two control wavelength designs discussed above exhibit the following trade-offs.

Device-Level Trade-offs: The control wavelengths of the abort design can be accommodated with the data wavelengths of a channel in a single waveguide. The sense design requires separate waveguides for the control wavelengths. However, the sense design requires fewer modulator rings than the abort design, and hence is more energy-efficient.

Performance Trade-offs: The sense design can potentially provide better performance gains than the abort design because of its 'sensing' capability. That is, the sense design does not require the stealer to abort stealing upon collision of its message; instead it temporarily halts stealing and waits for an opening to revert to stealing. The abort design does not have the sense capability and thus has to operate more conservatively.

4.2.2.3 Protocol Operation

When a sender node needs to transmit a flit, it performs several steps. These steps are explained according to the example channels shown in figure 4.3 where the sender 'B' has a flit to send to destination 'E':

1. B's flit has T phits (value of T is known at design time).
2. Split the flit occupying T cycles into two chunks each of length T/2 phits: 'owner chunk' and 'stealer chunk'.
3. Parity protect the owner chunk and send it on B's channel.
4. Send the stealer chunk on A's channel.

5. If a collision occurs:

    - Abort design: Terminate stealing. The unsent phits are parity protected and sent on B's channel after the owner's chunk is sent.
    - Sense design: Halt stealing. Resume if an opening is sensed. If the owner chunk completes before the stealer chunk, then send the remaining stealer chunk phits (with parity protection) on B's owned channel.

6. The destination uses the information on the control wavelengths to perform erasure correction and correctly reassemble the received phits into the original flit.

The protocol operation described above assumes a basic flit size of T phits. It can be extended to support flits of multiple sizes (number of phits). For example, to support flits of two sizes - data flits and control flits - just two control bits are needed at special locations in the flit to identify its size at the destination. For the data flit, the sender can set the first bit of the first two phits in the owner chunk to 0. Now, even if one of the phits gets corrupted, the destination can look at the duplicated value to know which size flit this is. For control flits, the sender can use the value 1. For large messages, these two bits will amount to negligible overhead.

4.2.3 Wavelength Stealing Gains

Section 4.1 analyzed the ideal case benefits and limits of a wavelength sharing network[3]. This section extends the analysis to the wavelength stealing architecture taking into account the overheads of control wavelengths and erasure coding.

[3]The speedup discussion presented in this section assumes the abort design. The sense design will experience similar speedups because it uses the same erasure coding technique and its single broadcast control wavelength consumes (approximately) the same laser power as the abort design's two control wavelengths.

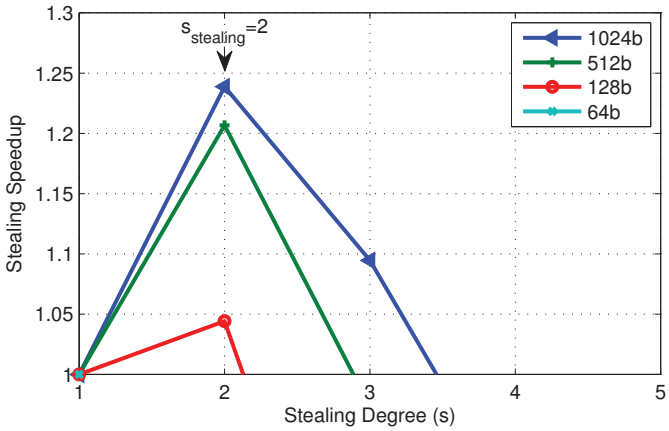


Figure 4.8: Wavelength stealing gains versus stealing degree s for different message sizes assuming $w = W_{sharing} = 16$ and $T_{prop} = 0$. The small 64b message does not exhibit a speedup.

The achievable speedup of the wavelength stealing architecture as a function of the stealing degree s can be expressed as:

$$Speedup_{stealing} = \frac{\left[\frac{messagesize}{W_{P2P}} + T_{prop}\right]}{\left[\frac{messagesize}{s\times\{W_{sharing}-c(s)\}} + e(s) + T_{prop}\right]}; (s \geqslant 2) \tag{4.7}$$

where, s: stealing degree, $c(s)$: control wavelength overheads; and, $e(s)$: erasure coding overheads. For 2-way ($s = 2$) stealing, $c(2) = 2$ (two control wavelengths per channel), and $e(2) = 1$ (single parity bit). For any arbitrary $s \geqslant 3$, the number of stealers on a channel is $(s-1)$. This requires control overheads $c(s)$ that scale linearly with s. In addition, the minimum number of check bits $e(s)$ required to correct up to $(s-1)$ erasures can be estimated from the Hamming bound [61].

Figure 4.8 plots the speedup gains of the wavelength stealing architecture as a function of the stealing degree s. From this figure, the following observations can be made:

- Ignoring the overheads of sharing, the ideal sharing degree is $s_{ideal} = 3$ (shown in section 4.1.3). However, due to overheads, the wavelength stealing architecture yields

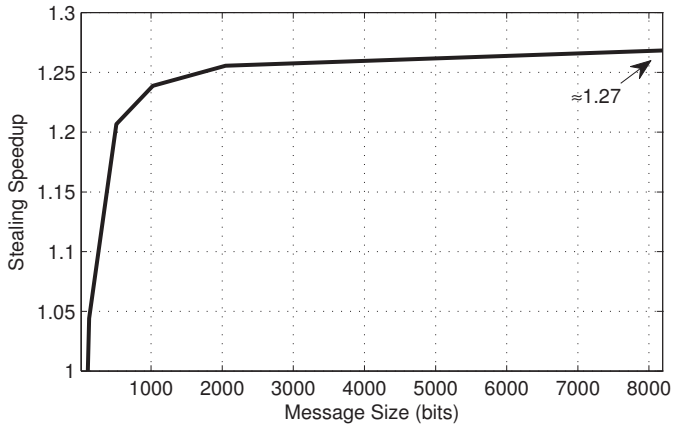


Figure 4.9: Wavelength stealing speedup as a function of message sizes for $s = 2$.

maximum speedup at a stealing degree of $s = 2$ (2-way stealing).

- Contrary to an ideal wavelength shared network, the speedup in the wavelength stealing architecture is dependent on the message (flit) size. This dependency is due to the overheads of erasure correction coding which get amortized better at larger message sizes. Figure 4.9 plots Eq.(4.7) as a function of message sizes for 2-way stealing. This figure clearly shows higher speedups for large messages with saturation at a speedup of 1.27.

The wavelength stealing architecture implements dedicated all-to-all connectivity similar to a P2P but is able to achieve higher node-to-node bandwidth in the presence of idle channels (for stealing to be successful) while consuming equivalent optical power. From the speedup analysis, it is also clear that the performance gains of the wavelength stealing architecture are more pronounced for larger messages. This makes the architecture more suitable to message passing applications that exhibit large-messages and low 'fan-out' communication patterns [49, 104].

4.3 'Macrochip' - A Message-Passing Multi-Chip System

This section dives into the details of the macrochip architecture on which the wavelength stealing techniques are evaluated. The macrochip architecture consists of an array of sites (also called nodes). These sites are interconnected using a high-bandwidth silicon photonic communication substrate. Sites in the macrochip can be processor chips (with multiple cores), memory chips or some other components. This chapter however uses a configuration in which all sites have processors and memory that generate messages directed to other sites in the array.

4.3.1 System Layout

The layout of a 64-site macrochip system is shown in figure 4.10. Each site has an optical bridge chip on the top layer and communicates with the other sites using data waveguides in the bottom substrate layer. The optical bridges house the optical devices and circuitry to support them. Optical (laser) power is generated by external lasers and delivered to the macrochip using edge connected fibers. This laser light is then forwarded to the sites using power waveguides (shown in red) for modulation. The data waveguides carry modulated light for inter-site communication.

Figure 4.10 shows a fully connected point-to-point layout composed of data waveguides shown as a blue loop. When implemented, the data path is composed of multiple waveguide segments where each segment begins at a sender site and terminates at a destination site and does not form a loop. Between any two nodes, there are two possible paths for laying out a channel between them: a clockwise and a counter-clockwise path. With these two choices, channels in this layout are designed such that the propagation distance between a sender and its destination is minimized. Thus, sender 7 has a counter-clockwise channel to destination 0 and a clock-wise channel to destination 44 as shown in figure 4.10.

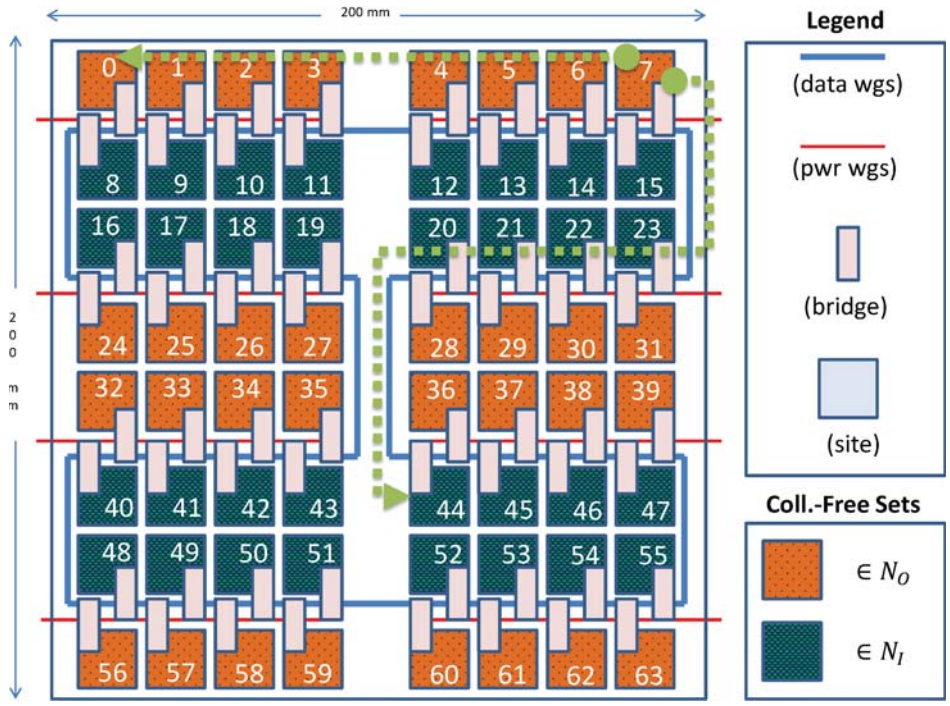


Figure 4.10: 8×8 single-layer (planar) macrochip layout.

Consequently, for a given destination, half of its senders will have channels that go in the clockwise direction towards that destination, and the other half will have channels that travel in the counter-clockwise direction.

Implementing the wavelength stealing interconnect on the macrochip requires placement of some modulator rings (for the stealers) in the bottom communication (waveguide) substrate. This can make the fabrication process more complex compared to a simple point-to-point interconnect. Interlayer couplers can be used to avoid having rings in the substrate layer; the trade-off however is higher link losses[4].

4.3.2 Stealing Pattern and Collision-Free Subsets

In the wavelength stealing architecture, a sender node uses its dedicated point-to-point channel to communicate with a destination but can also steal access on a channel to the

[4]The device loss of an inter-layer coupler is $\approx 2 - 3$dB [49].

same destination owned by another node. For a given destination, the static mapping between a sender and the node it steals from specifies the 'stealing pattern' of a wavelength stealing topology. The wavelength stealing architecture for the macrochip uses a stealing pattern in which *a sender steals on channels owned by its two immediate bridge chip neighbors along the waveguide loop*. Thus, in figure 4.10, sender 7 steals from 15 and 23 because they are its immediate two neighbors along the blue waveguide loop. To communicate with destination 0, sender 7 steals on node 23's channel to node 0. Similarly, to communicate with destination 44, sender 7 steals on node 15's channel to node 44. Thus, for a given destination, a sender steals from its immediate 'upstream' neighbor along the waveguide loop. This upstream neighbor stealing pattern leads to a partitioning of the macrochip into two node sets, N_O and N_I, with the property that all nodes in one set steal only from nodes in the other set. These two sets are highlighted in figure 4.10 and are called *collision-free sets*. They are called collision-free because as long as nodes in one set do not communicate with destinations in the other set and vice versa, collisions never occur. This is because under this scenario, there is never a case where both the owner and the stealer of a channel talk to the same destination. Restating this more formally: *the two sets* N_I *and* N_O *are collision-free because when members of a set restrict their communication to nodes within the set, there are no collisions*. The collision-free property is valid regardless of the communication pattern and the number of active senders within the sets, as long as the restriction on *no* inter-set communication is observed. This property is used extensively in the next section. For a N-node layout, the maximum number of nodes in each of the collision-free sets is $N/2$.

Pairing a node with an upstream neighbor to form an owner-stealer relationship leads to the farthest two senders of a destination being devoid of any channels to steal on. Only $\approx 3\%$ of the total source-destination pairs in the network fall into this category [5]. In order to maintain bandwidth symmetry, these few sender-destination pairs are provisioned with

[5] For an N-node layout, this fraction is $2N/(N \times (N-1))$.

some additional wavelengths. The energy required for these is accounted for in the power budget.

4.4 Guaranteed Gains on Virtual Machines

An architectural implication of the collision-free sets is that the cluster of nodes on the macrochip can be partitioned into multiple virtual machines (VMs) such that nodes within a VM always steal from nodes outside a VM. With no inter-VM communication, this architecture provides higher node-to-node bandwidth (because stealing is guaranteed to be successful with no collisions), and lower message latencies compared to a P2P network. To realize these VM gains, a hypervisor scheduling layer can be designed that schedules the VM jobs on the appropriate sites of the macrochip to ensure a collision-free operation.

To explain further, denote a VM job as $VM(np)$ where np is the number of processor chips (network nodes) required by this virtual machine for execution. In the 64-node macrochip system shown in figure 4.10, the two collision-free sets N_O and N_I contain 32 nodes each. This means that, two independent 32-processor virtual machines each hosting a multi-process or multi-threaded application can be scheduled *concurrently* and the intra application communication will not suffer *any* collisions in the network. With this scheduling, the wavelength stealing architecture will guarantee a $1.27\times$ higher node-node bandwidth over the P2P network. Since any subset of a collision-free set (N_O or N_I) is also collision-free, multiple VMs that require fewer processors than 32 can be scheduled together on a single collision-free set and take advantage of the guaranteed bandwidth gains over the P2P network. Thus, a set of VMs $\{VM_0(16), VM_1(16)\}$ can be scheduled on N_I and an independent set $\{VM_2(16), VM_3(16)\}$ can be scheduled on N_O so that all of them execute concurrently without collisions.

In general, suppose there are $(m + n)$ VM that need to be scheduled on an N-node

macrochip. A hypervisor scheduling layer can be constructed that maps each of these $(m+n)$ VMs to the appropriate collision-free subsets. This hypervisor simply partitions the total VMs into two sets such that each of them can be scheduled on an N/2 sized collision-free set. Formally put, the hypervisor scheduling layer is able to schedule the $(m+n)$ VMs if it can separate them into two sets, $S_m = \{VM_0(np_0), \ldots, VM_{m-1}(np_{m-1})\}$ and $S_n = \{VM_0(np_0), \ldots, VM_{n-1}(np_{n-1})\}$ such that they satisfy the following conditions:

$$\underbrace{S_m : \sum_{i=0}^{m-1} np_i \leqslant \frac{N}{2}}_{m \text{ VMs}} \quad ; \quad \underbrace{S_n : \sum_{j=0}^{n-1} np_j \leqslant \frac{N}{2}}_{n \text{ VMs}} \tag{4.8}$$

4.5 Results and Discussion

4.5.1 Evaluation Methodology

The performance of the wavelength stealing architecture was evaluated against two baseline designs: the unshared P2P network and the classic token-ring arbitration scheme [38] that has inspired many recent photonic network implementations [96, 95]. A detailed cycle-accurate network simulator was developed that models the complete functionality of these interconnect architectures.

All designs were evaluated on the 64-node macrochip layout shown in figure 4.10. Both synthetic and application-derived traffic from message-passing applications was used to evaluate the networks. These workloads are summarized in table 4.2. Performance of the applications running in single cluster and partitioned cluster configurations was also analyzed.

Pattern		Description
Synthetic	High-Radix	Uniform Random
	Low-Radix	Permutation/ Asymmetric
Application	NAS BT	Block Tridiagonal Solver
	NAS CG	Conjugate Gradient Kernel
	NAS DT WH	Data Intensive "White Hole" Graph Analysis
	NAS DT BH	Data Intensive "Black Hole" Graph Analysis
	NAS DT SH	Data Intensive "Shuffle" Graph Analysis

Table 4.2: Workload descriptions.

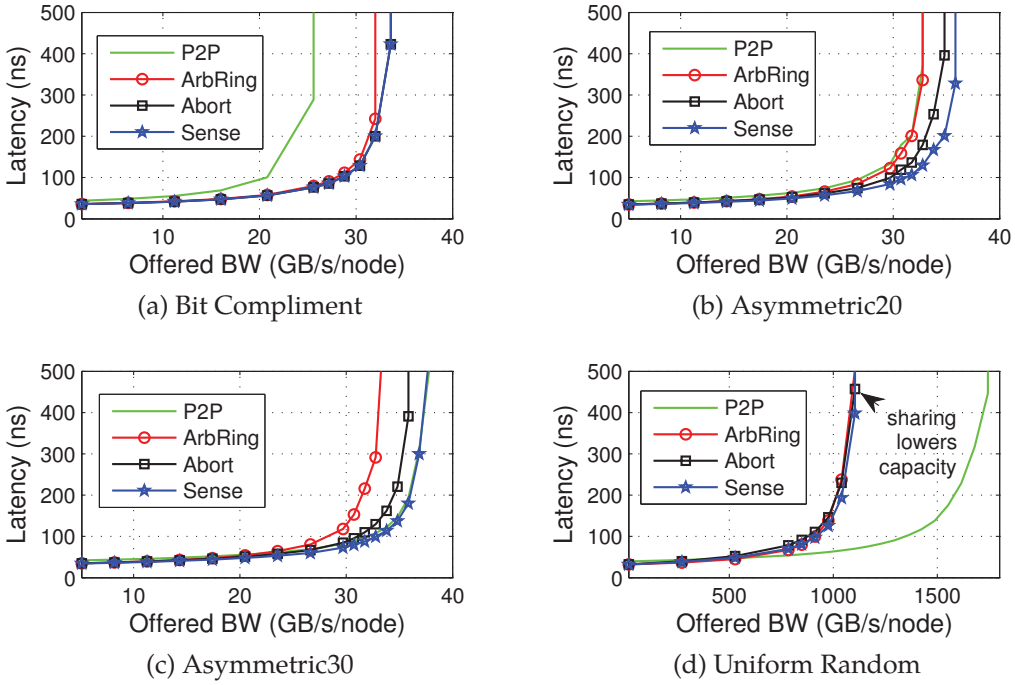
4.5.2 Synthetic Workload Evaluation

For synthetic workload evaluation, two categories of traffic patterns were simulated: high-radix and low-radix. A traffic pattern is characterized as having high-radix (low-radix) if a sender node communicates with a large (small) number of destination nodes. All synthetic patterns use a fixed message size of 1KB. Synthetic simulation results are shown in figure 4.11.

4.5.2.1 Wavelength Stealing vs. Arbitration

In token-ring arbitration, a single token is circulated per shared channel. This token represents the exclusive right of a sender to use the shared channel. It is well-known that the token-ring design does not scale well to highly-shared channels owing to high token-rotation latencies [50]. However current device loss constraints restrict sharing to just two senders per shared channel. With limited sharing, the rotation latency of the token-ring design is very small making this scheme a competitive point of comparison for the proposed wavelength stealing designs.

Three types of traffic patterns are used to compare the performance of the interconnect architectures. The bit-complement [23] (low-radix) traffic pattern causes no contention in the shared networks. To evaluate performance under various levels of contention, a new permutation pattern called Asymmetric k was devised. In this traffic pattern, given an



(a) Bit Compliment (b) Asymmetric20

(c) Asymmetric30 (d) Uniform Random

Figure 4.11: Synthetic traffic simulations depicting latency versus offered load for the three network architectures: wavelength stealing (Abort/Sense), token-ring arbitration (ArbRing) and point-to-point (P2P).

offered load, one of the two senders on the shared channel is active (on-average) k% of the time while the other is active $100 - k\%$ of the time (note that bit-complement traffic represents $k = 100\%$). Finally, the uniform-random [23] traffic pattern represents all-to-all (high-radix) communication that causes uniform contention on the shared channels.

From figure 4.11 it can be seen that as contention on the shared channel is increased, the throughput of the arbitration design drops significantly compared to the proposed stealing approaches. In addition the latency of the wavelength stealing designs is lower than the arbitration network, making it a good design-choice for latency-sensitive applications as well. Since the wavelength-stealing architecture performs either as well or better than a classical arbitration-based network, the rest of the evaluation only focuses on the arbitration-free stealing architecture.

4.5.2.2 Wavelength Stealing vs. Point-to-Point (P2P)

As discussed in section 4.1, sharing based networks have fewer wavelengths per channel and hence lower total bandwidth (capacity) compared to the P2P network. The effect of this can be observed from the uniform random ('all-to-all') traffic pattern in figure 4.11. The P2P network has higher total bandwidth and hence exhibits higher sustained throughput on this pattern.

From figure 4.11, it can be observed that the wavelength stealing schemes yield 1.27× higher throughput than the P2P network on the contention-free bit complement traffic pattern as quantified in section 4.2. As contention in the traffic increases (see asymmetric patterns in figure 4.11) the performance of the P2P network increases due to better utilization of the channels. In addition, the sense design gives better performance than the abort design at higher contention (see section 4.2).

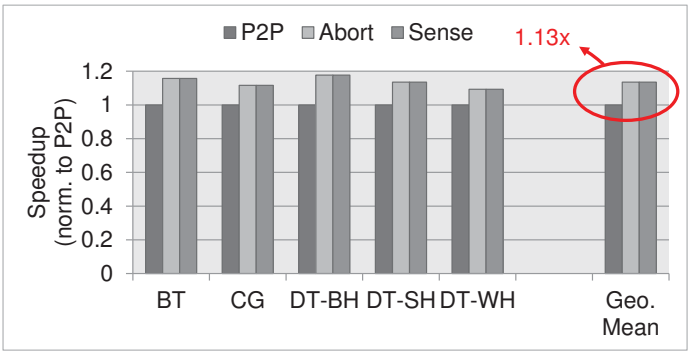
These simulations clearly show that the P2P network is ideally suited for high-contention traffic patterns while the sharing-based wavelength stealing architecture gives excellent performance under low-contention traffic. This fundamental design trade-off should be carefully considered when choosing a network implementation for a target application.

4.5.3 Application Workload Evaluation

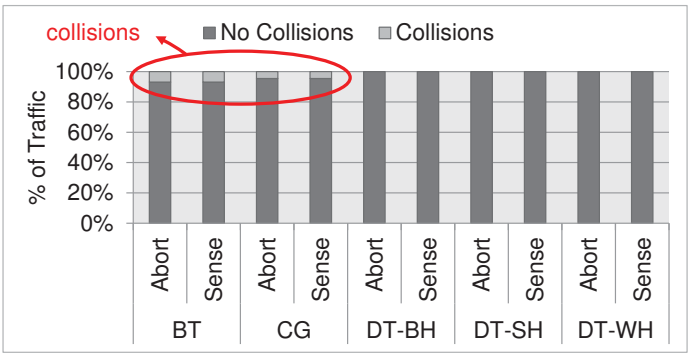
For application-traffic simulations, five benchmarks listed in table 4.2 were chosen from the NAS parallel benchmark suite [11]. Traces collected from the MPI versions of these benchmarks using Scalasca [99] were used to drive the network simulator.

4.5.3.1 Performance Analysis

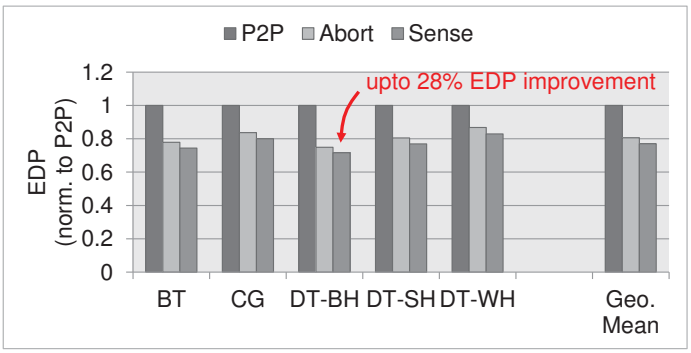
To evaluate benchmark performance, the execution time of the application traces was measured on the P2P topology as well as the abort/sense designs of the wavelength stealing architecture. Figure 4.12a shows the speedup as the execution time of the wavelength



(a) Execution time speedup



(b) Traffic breakdown (Collision vs. No Collision)



(c) Energy × Delay (EDP)

Figure 4.12: Application benchmark simulations.

stealing designs relative to that of the P2P network. The wavelength stealing designs achieve up to 1.17× speedup on some benchmarks and a geometric-mean speedup of 1.13× over the P2P network. These benefits come from the low-contention traffic behavior of these applications. Figure 4.12b shows that over 90% of the traffic in these applications does not suffer from any collisions and is able to utilize higher site-to-site bandwidths

Parameter	Assumption
Mod. (Insertion) Ring Loss	4dB
Inactive Mod. Ring Loss	0.5dB
Active Drop-Filter Ring Loss	1dB
Passive Ring Loss	0.05dB
Waveguide Loss	0.05dB/cm
Bridge Chip Waveguide Loss	1dB
Coupler Loss	2dB
Receiver Sensitivity Margin	4dB
Receiver Sensitivity Level	−21dBm
Ring Tuning Power	0.3mW/ring
Mod. Driver	35fJ/bit
Detector Driver	65fJ/bit
Max. Fiber WDM-Factor	32
Max. Waveguide WDM-Factor	16
Max. Port Fibers	2500
Power per Fiber	32mW

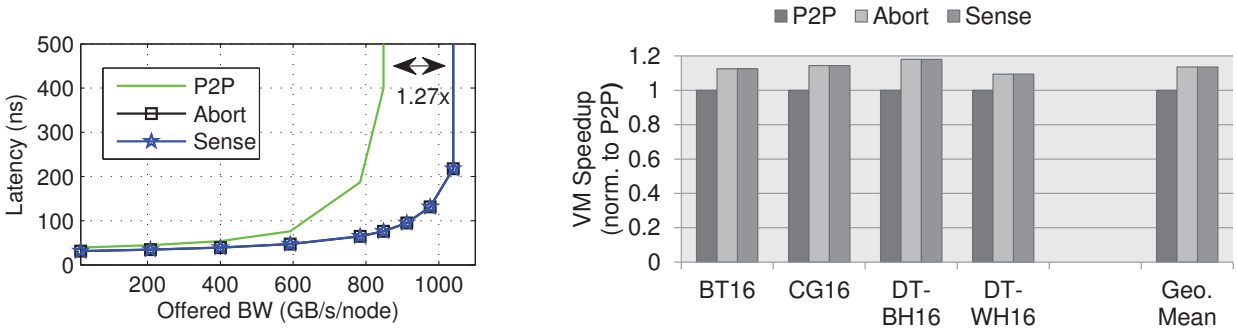
Table 4.3: Optical device parameters.

by successfully stealing idle channels. The variations in the achieved speedups between benchmarks arise due to the differences in their traffic patterns (collisions), message sizes and frequency of messages. Since much of the stealing is performed without contention, the conservative abort design performs on par with the sense design.

4.5.3.2 Energy-Delay Analysis

This section discusses the performance and energy trade-offs of the simulated networks[6]. The key metric used to compare the different network architectures is $Energy \times Delay$ (EDP). The static and dynamic energy for the networks were calculated using device parameters given in table 4.3. The energy calculation for the wavelength stealing architecture takes into account the additional dynamic energy expended on the parity bits. However, this has negligible impact on the total energy because the dynamic energy consumed is just a small fraction compared to the static energy consumption of these networks.

[6]The power estimates presented here only show the delivered laser power to the macrochip, not the wall-socket power.



(a) Domain uniform synthetic traffic pattern

(b) Application execution time speedup

Figure 4.13: Virtual machine performance gains. (a) Domain uniform synthetic traffic pattern depicting the collision-free subset property of the wavelength stealing architecture. (b) Four VMs are mapped into collision-free subsets to realize speedup gains.

Figure 4.12c shows the EDP of the networks for each workload. This graph is normalized to the P2P network. The wavelength stealing architectures achieve up to 28% lower EDP than the point-to-point network in the best case. The abort and sense designs achieve on average (geometric mean) 20% and 23% lower EDP respectively over the P2P network. The sense design uses fewer rings than the abort design leading to a slight reduction in the static tuning power and hence a marginally better EDP.

4.5.4 Virtual Machine (VM) Evaluation

Section 4.4 discussed leveraging the collision-free subset property of the wavelength stealing architecture to partition the macrochip into multiple VMs where each VM can execute an application and realize guaranteed bandwidth gains over the P2P network irrespective of the traffic pattern. To highlight the collision-free property of the subsets, a variant of the uniform random traffic pattern called 'domain uniform random' was devised. This communication pattern is the same as the uniform random pattern in table 4.2 except that senders belonging to a collision-free set only pick other nodes within the set as their random destinations. Figure 4.13a shows the latency curve for this synthetic pattern. Because no collisions are encountered in the system and stealing is successful 100% of the

time, the wavelength stealing architecture is able to achieve the theoretical 1.27× bandwidth advantage over the P2P network.

To explore the VM scheduling gains on applications, 16-node traces were collected for four NAS benchmarks listed in table 4.2. The macrochip was partitioned into four clusters and the four applications were scheduled concurrently using the algorithm presented in section 4.4. Figure 4.13b shows the execution speedups observed on these four application derived traffic patterns. All four applications achieved positive speedups and experienced no collisions. These results show the potential applicability of the wavelength stealing interconnect on a wide range of cluster configurations.

4.6 Summary

Interconnects with shared optical channels overcome the low node-to-node bandwidth limitation of a simple P2P network but suffer from high optical losses. In this chapter, analytical models were developed to quantify the limits on shared channels and it was determined that channel sharing with realistic photonics device losses does not scale beyond a sharing degree of three.

Based on this analysis, a novel interconnect architecture called *wavelength stealing* was proposed that enables arbitration-free optimistic access to shared optical channels and uses simple erasure coding to recover from collisions. Analytically, it was shown that the maximum performance benefits of this architecture occurs with two sharers on every channel. The design and implementation of such an architecture was presented using the same input optical power budget as that of a P2P network.

The P2P network and the wavelength stealing interconnect were simulated using both synthetic and application derived traffic patterns in the context of a 64-node multichip system. Using detailed performance and power analysis, it was demonstrated that the

wavelength stealing architecture exhibits up to 28% better EDP than the P2P network on applications with low-radix traffic. Furthermore it was shown that the wavelength stealing architecture can be leveraged to partition a multichip cluster into multiple VMs with guaranteed bandwidth gains over a P2P network under certain constraints.

5 SWITCHING IN PHOTONIC NETWORKS

Silicon photonic technology offers high-speed communication (up to 20Gb/sec per wavelength) at high bandwidth densities enabled by wavelength division multiplexing (WDM) which allows many wavelengths of light to be supported in a single waveguide or fiber. Light at these wavelengths can be encoded with information, routed to the destination and decoded at the receiver using various silicon photonic components.

Type	Implementation
Channel sharing (1 hop)	Single-writer single-reader (SWSR)
	Singlr-writer multiple-reader (SWMR)
	Multiple-writer single-reader (MWSR)
	Multiple-writer multiple-reader (MWMR)
Path sharing ($\geqslant$ 1 hops)	Optical switching
	Electrical switching

Table 5.1: Categories of silicon photonic networks.

In recent years, a variety of nanophotonic architectures have been proposed in literature (see chapter 3). These prior designs can be broadly placed into one of two categories highlighted in table 5.1. As silicon photonic technology advances and prototypes for various optical components are demonstrated [107, 102, 105], it becomes imperative that the efficacy of the proposed photonic architectures be evaluated under achievable loss characteristics.

The previous chapter explored channel sharing topologies for photonic networks under realistic loss assumptions. Channel sharing designs are quite popular in literature as they are 1-hop. Thus, communication in these networks require the fewest number of O/E and E/O conversions resulting in low dynamic energy consumption. However, as explained in the previous chapter, photonic networks based on ring-resonators are static power dominated. Specifically, laser power together with the static energy expended in tuning the ring-resonator devices constitute the bulk of the power consumption in these

networks. Thus, in the current technology generation, optimizing for dynamic energy consumption has minimal impact on a network's power cost. Therefore, this chapter relaxes the 1-hop constraint and explores switched (path sharing) photonic network designs (see table 5.1) where packets can experience multiple ($\geqslant 1$) hops. Depending on whether the routing elements are optical or electrical, switched photonic networks can be classified into two sub-categories, optically-switched and electrically-switched (see table 5.1).

Current state of the art optical switches incur a loss of about 3dB [25]. A recent ISCA paper [49] has analyzed optical switching in photonic networks. They show that a device loss of 0.75dB or lower must be acheived to make optical switching viable from a performance-power standpoint. They conclude that achieving this optical loss goal requires a major breakthrough in device development. This leaves electrical switching as an alternative design approach to explore in silicon photonic networks.

Electrically-switched photonic networks do not exacerbate the laser power consumption by incorporating many ring-resonators along a waveguide as is the case in channel sharing, nor do they require high loss optical components as in optical switching. Thus, on the static power front, these networks offer an inherent advantages over the other approaches. However no prior work has provided a comprehensive analysis of electrically-switched photonic network designs. In this vein, this dissertation makes the following contributions in this chapter:

- An in-depth comparison of popular electrically-switched networks within the constraints of silicon photonic technology is presented. It is demonstrated that silicon photonic technology imposes a *fixed* cost on all network channels. This is in stark contrast to traditional large-scale electrical networks where the network channels are assigned *different* costs depending on the length of the channel. Both low-radix and high-radix topologies as well as direct/ indirect networks are evaluated and it is demonstrated that the fully-connected topology offers significantly higher perfor-

mance than the other topologies.

- It is shown that a traditional input-queued crossbar router becomes prohibitively expensive in terms of area and power when scaled up for the fully-connected topology. It is demonstrated that by adopting a 'topology-aware' router design approach, low cost routers can be designed that employ simple arbiters instead of expensive allocators and do not require full crossbars. Three novel router architectures are presented that offer different performance characteristics.

- A novel mechanism for enabling differentiated quality-of-service (QoS) guarantees in a photonic network employing a fully-connected topology is developed. Using the proposed mechanism, varying levels of bandwidth can be realized in different regions of the network. This enables a hypervisor to map virtual machines (VMs) with different bandwidth demands to the appropriate bandwidth regions in the network.

The rest of the chapter is organized as follows. Section 5.1 presents the architecture of the baseline macrochip system. Section 5.2 highlights some important design considerations for nanophotonic systems. A quantitative analysis of electrically-switched photonic networks is presented in section 5.3. Section 5.4 describes the topology-aware router design approach. A novel mechanism for enabling differentiated QoS in presented is section 5.5, and section 5.6 concludes this chapter.

5.1 Macrochip - A Kilo-core Architecture

This section revisits the macrochip architecture that was first introduced in chapter 1. Recall that the central idea behind the macrochip concept is to employ a fast high-bandwidth interconnection network that overcomes the die size limits imposed by technology yields

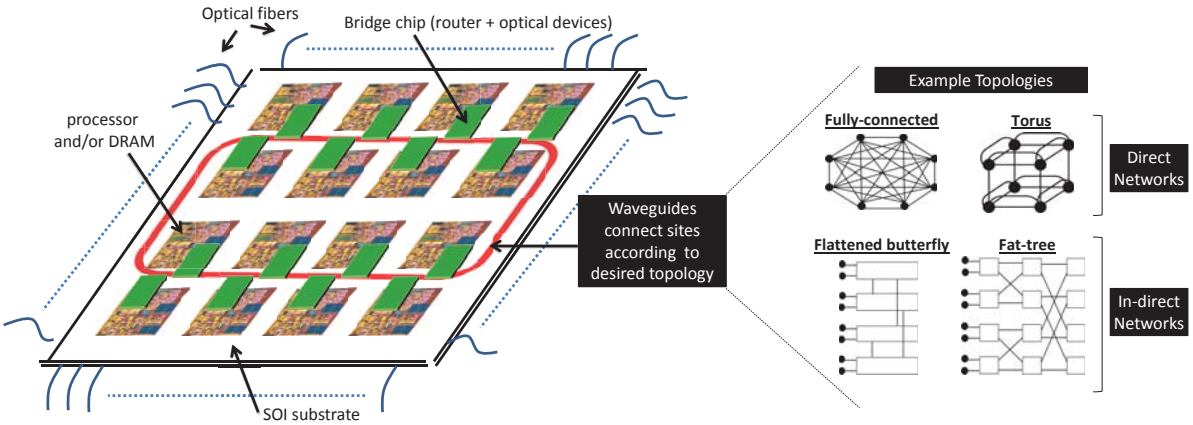


Figure 5.1: A 16-site macrochip system [51, 50]. The waveguides are the "wires" that connect the sites (nodes) together. Different topologies can be realized on the macrochip platform. The routers in these topologies are incorporated in the bridge chips.

while achieving the performance of a single large monolithic 'virtual' chip. For the purposes of this chapter however, it serves as a baseline to highlight how the opportunities and constraints afforded by silicon photonic technology differ from those in traditional large-scale electrical networks. Understanding these technology considerations is crucial in designing and evaluating different photonic networks.

The macrochip architecture is depicted in figure 5.1. Laser light is generated by external lasers (not shown) and sourced to the macrochip via edge-connected optical fibers. The amount of laser power that can be brought into the system is limited by the number of fibers than can be connected along the system perimeter. The laser light couples into the waveguides that are fabricated on the SOI routing substrate. The waveguides connect different sites of the macrochip and carry information as modulated light. Optical devices that perform O/E and E/O conversions are integrated in photonic bridge chips that are mounted on the sites. Router logic used to perform electrical switching can be incorporated in the bridge chips as well.

The sites in the macrochip can be processor chips, memory chips or both. A 64-site configuration where each site contains a multi-core processor chip ($125 mm^2$) die-stacked

on top of a memory die ($225mm^2$) is presented in [50]. The macrochip system can either support a shared memory model [50] or target a message-passing paradigm [49].

5.2 Optical Technology Considerations - Why Traditional Solutions Don't Apply?

This section discusses some important characteristics that differentiate silicon photonic networks from their electrical counterparts.

Long distance communication is "cheap": The propagation losses in silicon photonic links are low ($0.05dB/cm$) [59]. Thus, *optical technology largely obviates "cable" length considerations and enables topologies with long non-minimal length channels*[1]. This is in stark contrast to traditional electrical networks where an important design consideration is to avoid long channels [42].

Every link has (approximately) the same cost: All links in an electrically-switched photonic network employ the same number of optical devices. Hence, they suffer from the same optical losses *except* for propagation loss (this is due to the differences in channel lengths). However, as discussed earlier, the propagation losses in optical links is *low*. Thus, the minute differences in optical losses due to channel lengths can be ignored leading to the implication that *every link in a photonic network has (approximately) the same cost*.

Total bandwidth is the key design constraint: The amount of laser light that can be delivered to a photonic substrate is limited either due to considerations of power, packaging, or both. For example, the total input laser power delivered to the macrochip is limited by

[1]Non-minimal length channel layouts are used to avoid waveguide crossings as they introduce significant crosstalk and power loss [49].

the number of fibers that can be connected along the perimeter (see figure 5.1). Even if this packaging limitation could be overcome, generating laser light is expensive. This is because efficiencies of commercial WDM lasers currently fall within the paltry range of $1-5\%$ [108, 14, 54]. This is why recent papers in the area have argued for imposing a laser power budget constraint on all network designs when evaluating them for performance [50, 49]. Now,

$$\text{Total laser power} \propto \frac{\text{loss}}{\text{wavelength}} \times \text{total wavelengths}$$

Since, the loss per wavelength (link) is approximately the same for all wavelengths, the input laser power sets the total available bandwidth (links) in the network. Thus, *when evaluating different silicon-photonic network designs for performance, it is important to hold the total bandwidth constant*. This approach differs from the "equivalent bisection bandwidth" methodology employed in electrical networks [43].

Can build richly-connected topologies: Although photonic technology is facing many challenges, it is also providing unique opportunities that are new to our area. Using WDM, many parallel streams of communication can be established in a single waveguide enabling *system designers to explore richly-connected topologies such as a fully-connected network* [50, 49].

Throughput is the key performance metric: The cost of an optical channel is largely defined by the optical losses incurred by light when passing by various optical components (modulators, drop-filters, couplers etc.) along a light path and the number of ring resonator devices used by the channel. This channel cost is rather significant (17dB loss per link [51]) and is independent of the activity on the channel. To justify these high upfront power costs, silicon-photonic networks should be designed and employed in high utilization scenarios. In other words, *an important design requirement of an efficient photonic design is to sustain high operating throughputs under a variety of traffic conditions.*

5.3 Photonic Topology Exploration

This section presents an in-depth comparison of popular topologies within the purview of silicon photonic technology. The **goal** of this analysis is to identify the target topology that yields the *highest sustainable throughput*.

5.3.1 Candidate Topologies

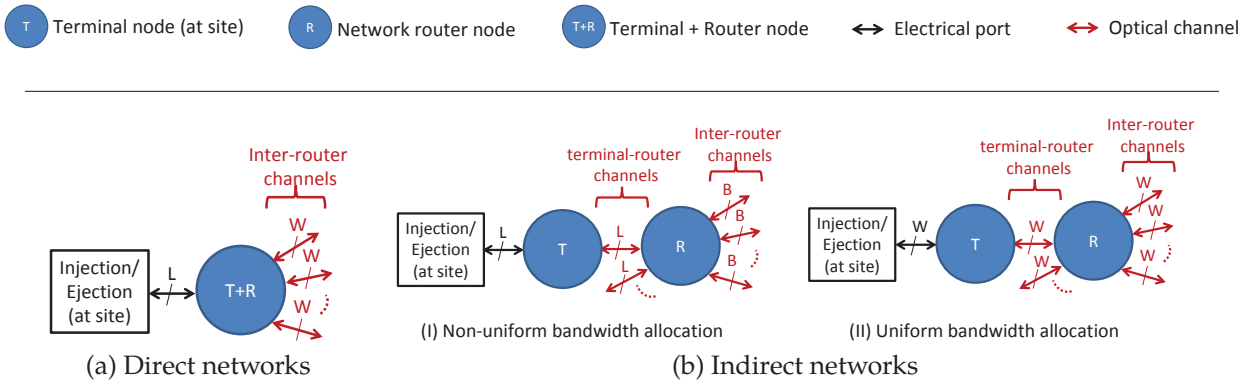
The topology of an interconnection network has a profound impact on performance. These network topologies can be placed into one of two categories: direct and indirect.

5.3.1.1 Direct networks

In a direct network, the terminal (site) and the network router are incorporated into the same node as depicted in figure 5.2a. In this evaluation, two direct networks are chosen for comparison: a torus (k-ary n-cube) and a fully-connected topology. Although a fully-connected topology is considered too prohibitive for electrical networks; it can be efficiently packaged in a silicon photonic fabric as discussed in section 5.2. Note that a fully-connected topology represents an 'extreme' radix design having a network diameter of one where each network node can directly communicate with any other node in the system. Although a mesh topology is quite popular in electrical interconnection networks, it is ignored in this evaluation as it yields lower throughput performance than a torus network [23].

5.3.1.2 Indirect networks

In an indirect network, the terminal (site) nodes are distinct from the network router nodes as shown in figure 5.2b. Two well-known indirect topologies are explored: fat-tree (k-ary n-fat) [57] and flattened-butterfly (k-ary n-flat) [43]. Both these topologies provide high path diversity leading to good throughput performance even under adversarial traffic



(a) Direct networks (b) Indirect networks

Figure 5.2: Network nodes can be either terminal (T), router (R) or both (T+R) (a) Direct network node (b) Indirect network nodes

conditions. Recent studies have indicated a departure from low-radix designs towards high-radix topologies in traditional interconnection networks [44]. By including the fat-tree and the flattened-butterfly topologies in the evaluation, the performance gains offered by traditional high-radix designs in silicon photonics can be explored.

5.3.1.3 Routing Algorithms

Achieving good throughput performance requires an efficient routing algorithm that can load-balance the network channels under different traffic conditions. Universal-globally-adaptive-load-balanced (UGAL) routing [87] is employed in three networks: fully-connected, torus and flattened-butterfly. UGAL adaptively switches between minimal and non-minimal routing on a packet-by-packet basis depending on the congestion conditions. The decision to use minimal or non-minimal routing is made at the source node. Packets that are sent non-minimally in UGAL employ Valiant routing i.e. they are first routed to a random intermediate node which then forwards the packet to the intended destination. Hence, with UGAL, the throughput performance obtained varies between minimal and valiant routing performance depending on the traffic conditions.

The fat-tree network provides high path diversity and all of its paths are minimal. Therefore it does not need to rely on non-minimal routing to achieve load balance. Instead,

it uses adaptive nearest-common-ancestor routing [23] where the packet is first adaptively 'uprouted' to one of the common ancestors of the source and destination node. From there, the packet is 'downrouted' to the destination using a deterministic path. Because of the high path diversity that exists in the uprouting phase, the fat-tree network provides high throughput performance even on difficult traffic patterns.

5.3.2 Performance Evaluation

The performance of the networks is evaluated to find the design that can sustain the highest bandwidth offered by optics. Following the discussion presented in section 5.2, all network designs are **constrained** to use a *fixed total optical bandwidth budget*, henceforth denoted by T.

Denote N as number of terminal (site) nodes; L (bits) as the unit of injection (henceforth referred to as packet); and, $W_{\mathtt{full}}$ (bits/cycle) as the optical channel bandwidth of a fully-connected network. Without loss of generality, the total optical bandwidth budget is set as $T = N \times (N-1) \times W_{\mathtt{full}}$. Given this budget value, the optical channel widths W of the different networks can be easily computed. For example, in the torus network, the optical channel bandwidth $W_{\mathtt{torus}}$ can be solved from the following equation:

$$\overbrace{W_{\mathtt{torus}}}^{\text{channel BW}} \times \overbrace{2 \times n \times N}^{\text{\# inter-router channels}} = N \times (N-1) \times W_{\mathtt{full}}$$

where the left-hand side expression is the total router-to-router bandwidth in a (k-ary n-cube) torus network. The value for $W_{\mathtt{torus}}$ is given in table 5.2.

For the indirect networks, there are two options on how to divide up the bandwidth budget amongst the optical network channels: (I) non-uniform bandwidth allocation, and (II) uniform bandwidth allocation (see figure 5.2b). In the former case, the optical channel bandwidth between a site and router is L-wide and the router-to-router channels are sized

Topology	Optical Channel BW (W)
Fully-connected	W_{full}
Torus (k-ary n-cube)	$\frac{(N-1)W_{full}}{2n}$
Fat-tree (folded-clos) (k-ary n-fat)	$\frac{(N-1)W_{full}}{2n}$
Flattened-butterfly (k-ary n-flat)	$\frac{k(N-1)W_{full}}{(k-1)(n-1)+2k}$

Table 5.2: Optical channel bandwidth of the topologies.

to be B bits wide. Alternatively, in the uniform allocation strategy, all optical channels are sized to be the same bandwidth W. Both bandwidth allocation strategies were evaluated on the fat-tree and flattened butterfly networks and it was always observed that the uniform bandwidth allocation design yields better throughput performance. Hence for the sake of brevity, the rest of this section only focuses on the uniform bandwidth allocation design approach for indirect networks. The optical channel bandwidths for the fat-tree W_{fat} and flattened-butterfly W_{flat} networks can be solved from the following equations:

$$W_{fat} \times [\overbrace{2 \times N \times (n-1)}^{\text{\# inter-router channels}} + \overbrace{2 \times N}^{\text{\# term-router channels}}] = N(N-1)W_{full}$$

$$W_{flat} \times [\frac{N}{k}(k-1)(n-1) + 2 \times N] = N(N-1)W_{full}$$

The values of the optical channel bandwidths W_{fat} and W_{flat} are given in table 5.2.

The performance of the different topologies is compared on a $N = 64$ site macrochip system. Packets are sized to be $L = 256bits$ and the total optical bandwidth budget is computed using $W_{full} = 4bits/cycle$; that is, each optical channel in the fully-connected network is 4 bits wide. Using this bandwidth budget, the channel widths of the other photonic networks can be derived from table 5.2. Depending on the channel bandwidth, packets in the network suffer from different serialization delays (phits) as highlighted in table 5.3.

Topology	$(\mathbf{k}, \mathbf{n})$	phits/packet
Fully-connected	$(64, 1)$	64
Torus	$(8, 2)$	4
	$(4, 3)$	6
	$(2, 6)$	13
Fat-tree	$(8, 2)$	4
	$(4, 3)$	6
	$(2, 6)$	13
Flattened-butterfly	$(8, 2)$	3
	$(4, 3)$	4
	$(2, 6)$	5

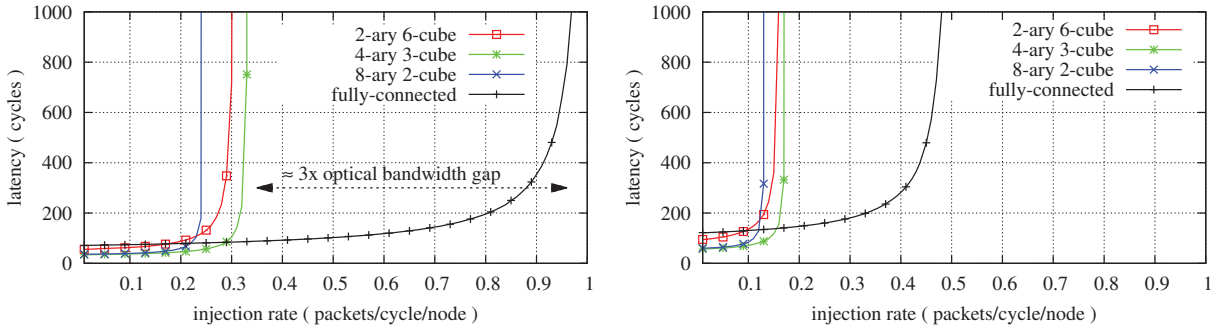
Table 5.3: Serialization delay of packets on a channel.

Cycle-accurate simulations are used to evaluate the throughput performance of the networks. These simulations are conducted under the following **assumptions:**

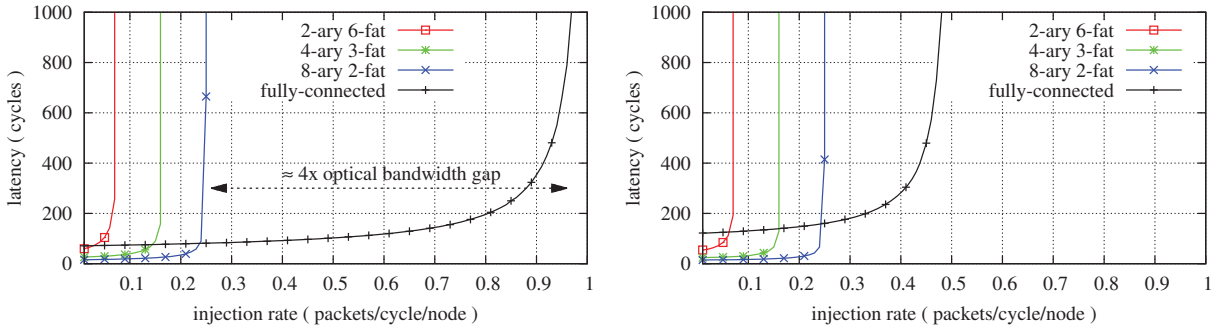
- **Router model:** A single-cycle input-queued crossbar router model is used for all networks. The router is provisioned with infinite virtual-channels (VCs), infinite buffers per VC, and infinite output buffering. The term infinite here implies a value large enough to ensure that that the resource does not become the performance bottleneck.

- **Flow control mechanism:** A credit-based flow control mechanism is employed in the networks and is implemented using dedicated credit lines. It is ideally assumed that this credit bandwidth is completely free and does not use up any of the optical budget.

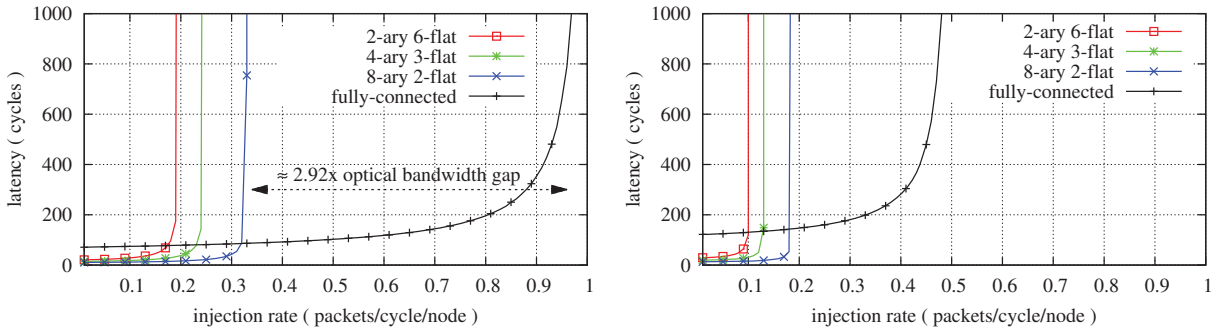
These assumptions enable us to obtain upper-bound estimates of the sustainable throughput performance offered by these network topologies. The performance results of the networks are given in figure 5.3.



(a) Torus



(b) Fat-tree



(c) Flattened-butterfly

Figure 5.3: Performance of networks under a fixed optical bandwidth budget on both (Left) favorable traffic and (Right) adversarial traffic patterns. If the optical bandwidth budget constraint is removed, then the figure also highlights the factor increase in total wavelengths required by the networks to match the capacity performance of a fully-connected topology (see section 5.3.3).

5.3.2.1 Favorable traffic performance

The uniform random traffic pattern is used to evaluate the performance of the networks under benign traffic conditions. The capacity of a network (or peak throughput under uniform random traffic) is given by $C = 2B/N$ where B is the bisection bandwidth and N is the number of nodes in the network [23]. A N-node fully-connected network has $N^2/2$ bisection channels. With each channel sized as L/N, the bisection bandwidth of the fully-connected network becomes $LN/2$ leading to an achievable capacity of 1 packet/cycle/node. Hence as figure 5.3 shows, the fully-connected network can sustain the peak injection-rate of 1 packet/cycle/node under uniform random traffic. The other photonic networks however demonstrate significantly lower sustainable throughput performance as shown in figure 5.3. These results directly follow from the factor difference between the channel bandwidth requirements of these networks to achieve capacity versus what is available to them due to the total optical budget constraint. For example, to operate at capacity, the inter-router channels of a k-ary n-cube torus network must be sized as $k/8$ with respect to the injection bandwidth[2] [23]. Therefore, for a 8-ary 2-cube network, in order to operate at the peak injection rate of 1 packet/cycle/node, each channel must be sized to be the same width as the injection (packet) bandwidth. However, looking at table 5.3, it can be seen that the optical bandwidth budget constraint restricts the inter-router channels to only 1/4th of the required size. This reduces the sustainable throughput by a factor of 4 compared to peak value (see figure 5.3a-Left). Similarly, both fat-tree and flattened-butterfly networks require the channels to be matched in width compared to the injection bandwidth [43]. However, as table 5.3 shows, the total optical bandwidth budget restricts the channels to be lower width than the required value leading to a significant drop in peak throughput (see figure 5.3b,5.3c-Left).

[2] when k is even.

5.3.2.2 Adversarial traffic performance

The three networks that employ UGAL (fully-connected, torus and flattened-butterfly) revert to Valiant routing when faced with adversarial traffic conditions. Valiant routing doubles the load on the network channels regardless of the traffic pattern leading to only half the throughput performance compared to the uniform random traffic case. Thus, the peak sustainable throughput of the fully-connected, torus and flattened-butterfly networks under an adversarial traffic pattern (see figure 5.3a,5.3c-Right) is only half of what they achieve in the uniform random case (figure 5.3a,5.3c-Left). The fat-tree network incorporates load balancing regardless of the traffic pattern when uprouting to a middle stage switch (see section 5.3.1.3). Therefore, the fat-tree network saturates at the same throughput under both adversarial and benign traffic conditions. The networks were simulated against some other adversarial traffic patterns[3] and similar performance trends were observed.

5.3.2.3 Performance Analysis takeaway

The analysis presented in this section shows that the fully-connected topology provides more than 2$\times$ higher throughput performance under both benign and adversarial traffic conditions compared to the other networks for the *same total available bandwidth*. However, the fully-connected network also suffers from the highest packet serialization delay as shown in table 5.3 leading to higher zero-load latency as shown in figure 5.3. This higher latency can adversely impact the performance of latency-critical applications such as websearch. Fortunately, given the link-rates offered by photonics ($\geqslant 10\mathrm{Gbps}$), the latency of a cache-line sized packet (256b) is very competitive (10s of nanoseconds) *regardless* of topology. However, as shown in this section, the factor differences in accepted loads of these topologies is significant and can become the system bottleneck. That is, if the utilization of the network is sufficiently high, then the other topologies besides fully-connected would

[3] bit-complement, bit-rotation, tornado etc.

suffer from large queuing delays in message communication as shown in figure 5.3 leading to inferior performance.

5.3.3 Optical Power Discussion

Imposing a fixed optical bandwidth budget ensures that all networks use the same number of: 1) ring resonators and 2) total wavelengths. Observations 1 and 2 restrict all topology designs to consume equivalent ring resonator tuning power and laser power[4] respectively. Therefore *the optical power costs of the different topology designs under an optical bandwidth budget are the same*. The energy efficiency of optics is currently *projected* to be significantly less than $1pJ/bit$ [50]. Under this efficiency target, all the simulated networks consume less than $80W$ optical power assuming $5GHz$ routers and $10Gbps$ link-rate.

An interested reader at this point may pose the question: if the fixed optical bandwidth condition is relaxed, then how many more wavelengths are required by the different topologies to meet the throughput performance of the fully-connected network? The answer to this question is labeled as the *optical bandwidth gap* in figure 5.3-Left. For example, the best performing torus network configuration (4-ary, 3-cube) requires $3\times$ more wavelengths than the fully-connected network to provide the same performance. This translates to requiring $3\times$ more laser power which is not practical as discussed in section 5.2.

5.3.4 Scalability Discussion

This section analyzes the scalability of the fully-connected network for the macrochip system. A single layer 64-site layout of the fully-connected topology is demonstrated in [49] for the macrochip. This layout employs 512 data waveguides and is packaged in a $20cm \times 20cm$ SOI substrate. Scaling the fully-connected topology to higher node counts requires band-

[4]The exact laser power consumption of the different topologies may differ slightly due to differences in channel (waveguide) lengths. However, as section 5.2 points out, the propagation losses in silicon photonics is low. Thus, these small differences can be ignored.

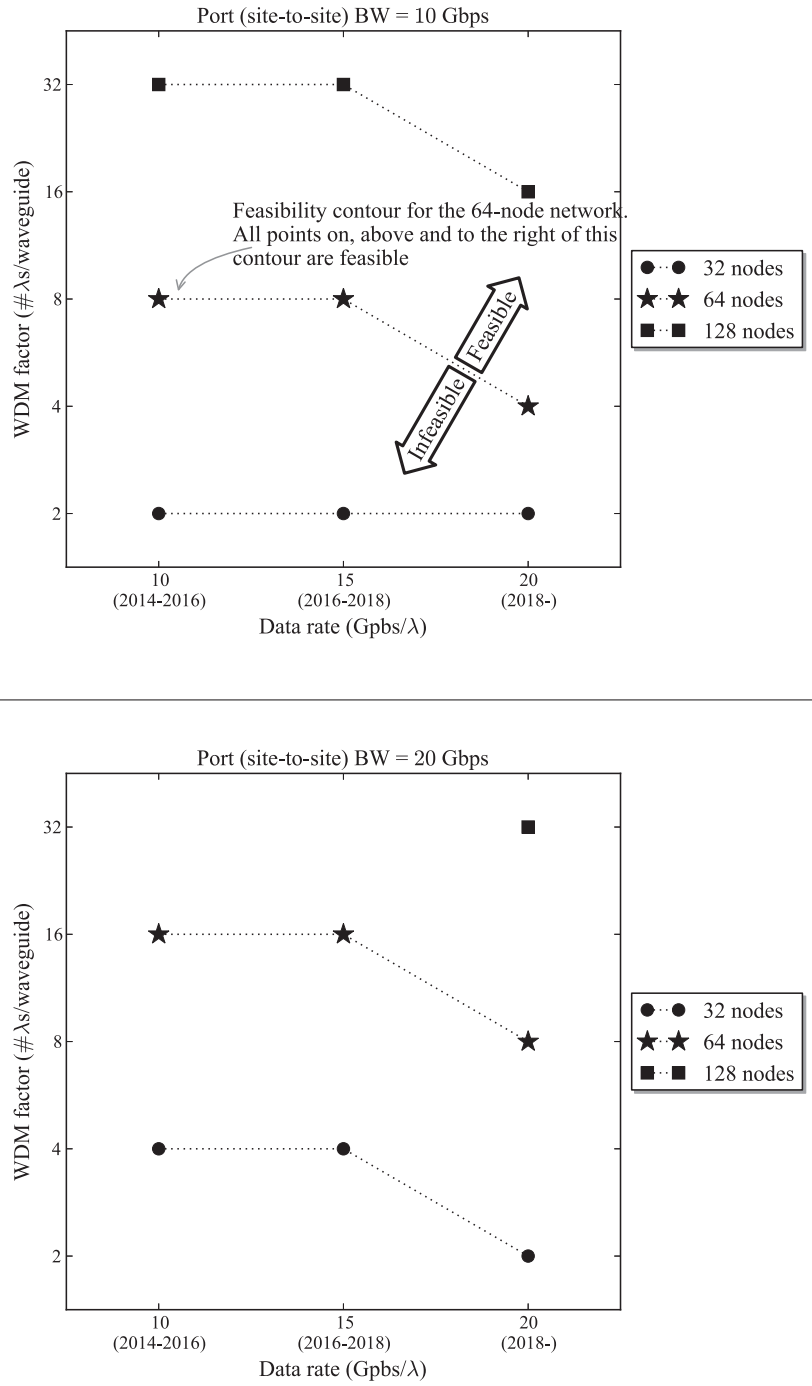


Figure 5.4: Feasibility contours for different network sizes assuming site-to-site bandwidths of 10Gbps (Top) and 20Gbps (Bottom). All points on, above and to the right of a contour are feasible for that network size.

Pattern		Description
Application	NAS BT	Block Tridiagonal Solver
	NAS CG	Conjugate Gradient Kernel
	NAS DT WH	Data Intensive "White Hole" Graph Analysis
	NAS DT BH	Data Intensive "Black Hole" Graph Analysis
	NAS DT SH	Data Intensive "Shuffle" Graph Analysis

Table 5.4: Workload descriptions.

width to scale quadratically with the number of nodes. Such bandwidth scaling can be achieved by increasing any combination of the following parameters: number of waveguides, WDM factor per waveguide, and data rate per wavelength. This work conservatively assumes that the number of waveguides that can be packaged on the macrochip will not increase in the near term due to device scaling. This leaves us two options to explore when scaling the macrochip system to higher chip counts: data rate and WDM factor. The scaling trends of these two parameters are estimated from the macrochip projections provided in [51, 49].

Figure 5.4 plots the 'feasibility contours' for different network sizes as a function of technology parameters (data rate and WDM). For example, a 128-site macrochip system with a site-to-site bandwidth requirement of 20Gbps can be fabricated using a data rate of 20Gbps and a WDM factor of 32. If these technology parameters cannot be realized in a particular technology generation, than a 'multi-macrochip' approach can be adopted. However, exploration of multi-macrochip designs is beyond the scope of this dissertation and is left for future work (see chapter 6).

5.3.5 Application Workload Evaluation

For application-traffic simulations, five benchmarks listed in table 5.4 were chosen from the NAS parallel benchmark suite [11]. Traces collected from the MPI versions of these benchmarks using Scalasca [99] were used to drive the network simulator. To evaluate

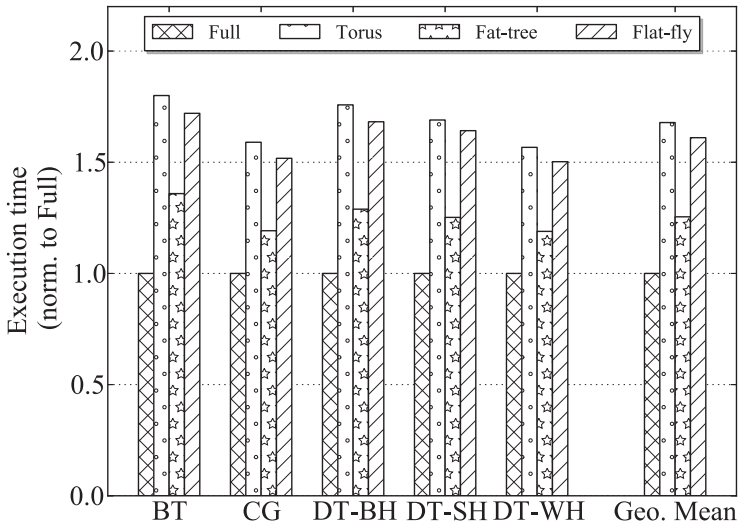


Figure 5.5: Application benchmark simulations.

benchmark performance, the execution time of the application traces on the different network topologies was measured. These results are shown in figure 5.5. This figure only shows results for the best performing configuration of each network topology.

As this figure shows, the fully-connected topology exhibits the lowest execution time on all the application workloads. This is due to the ample bandwidth advantage of this topology over the other networks (see figure 5.3). The variations in the execution time between benchmarks arise due to differences in their traffic patterns, message sizes and frequency.

5.4 Router Design for the Fully-connected Topology

The goal of this section is to investigate whether 'affordable' routers can be designed for a fully-connected network that can operate at the link speeds offered by photonics. This section starts by exploring the traditional input-queued virtual channel router and shows that this design becomes too complex for implementation in a fully-connected network. It

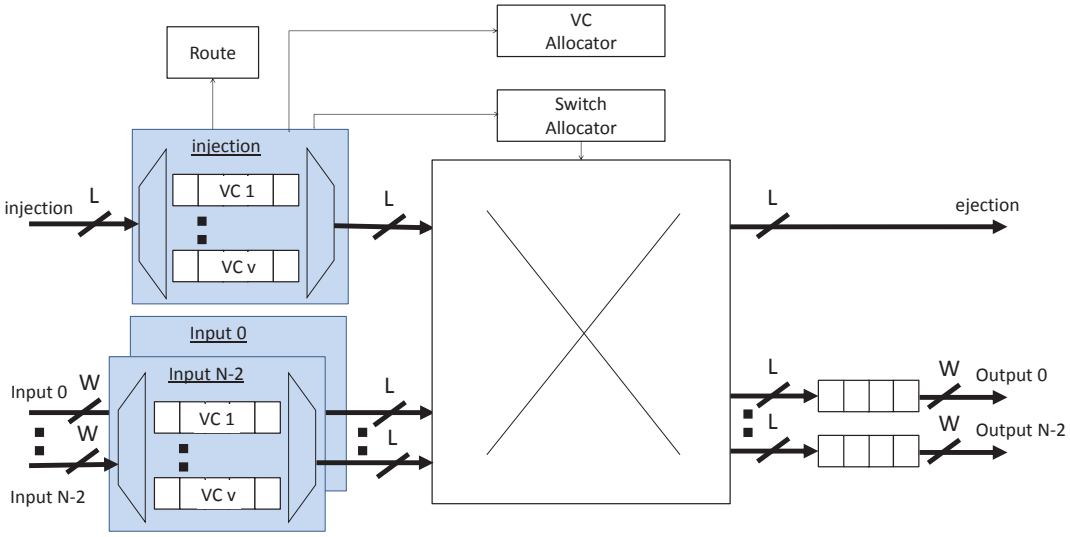


Figure 5.6: Microarchitecture of the input-queued router (IQR).

then shows that instead of using 'generic' topology-agnostic routers – be it input-queued [71, 83] or its high-radix variants such as [44] – a *topology-aware* router design philosophy should be adopted for the fully-connected network. Following this approach, novel routers designs are proposed that do not employ expensive components and give huge savings in energy and area compared to a generic router.

5.4.1 Baseline Input-Queued Router (IQR)

Figure 5.6 shows the block diagram of a traditional input-queued router for an N-node fully-connected network. Packets arrive via the input ports and are queued at the input virtual channel (VC) buffers. Route computation (RC) is performed on the packet to determine the output port. Next, virtual channel allocation (VA) is performed on the packet to reserve a virtual channel downstream from the output port determined in the RC stage. Now if buffer space is available in the downstream virtual channel, then the packet takes part in switch allocation (SA). Upon winning switch allocation, the packet traverses the crossbar switch (ST) and is queued at the output port buffer. Note that output buffers are required in a fully-connected network because of the bandwidth mismatch between the network

ports and the internal router data-path.

The logic complexity of the IQR router lies in three structures: crossbar, switch allocator and virtual-channel allocator. For an N-node fully-connected network, the complexity of all three structures scales quadratically with N. Hence, even for modest values of N, the area and power consumption of these structures reach prohibitive levels (shown in section 5.4.3). Note that the scalability issues of the IQR router in high-radix networks have been highlighted in prior papers [44]. To reduce this complexity, Kim *et al*. [44] have proposed techniques such as: aggressive speculation, distributed allocators (resulting in deep pipelines) and a hierarchical crossbar structure with intermediate buffering that incorporates its own intra-crossbar credit-flow mechanism. In the next section, it is shown that these aggressive techniques/ structures are simply *not required* by a router designed for a fully-connected network.

5.4.2 Topology-Aware Router Design

This section presents the topology-aware router design approach for a fully-connected network. The section starts by analyzing the affects of network radix on the router structures (crossbar and allocators). The goal is to develop insights that can be leveraged to simplify router design.

Insight #1 **- A full crossbar is not required in a fully-connected network router:** Consider a low-radix 8-ary 2-cube (64-node) torus network. To achieve capacity, the network channel widths (W) of this network must equal the injection bandwidth (L), i.e. $W == L$ [23]. Now, to ensure that the router does not become the bandwidth bottleneck, it must be capable of forwarding all the L sized packets arriving at the various input ports to any of the output ports. This internal forwarding bandwidth can only be achieved in the router by employing a full crossbar. Now consider a 64-node fully-connected network. In order to

achieve capacity, the network channel widths, W, need to be only 1/64th of the injection bandwidth (L), i.e. $W << L$. Hence, packets arrive at much lower rates in a fully-connected router. This implies that on average only a single L-sized packet is fully received at an input port on any given cycle ready to be moved to an output port. Therefore, the forwarding bandwidth required in a fully-connected network router is much less compared to a torus router, mitigating the need for a full crossbar.

Insight #2 **- A switch allocator (SA) is not required in a fully-connected network router:** Traditional routers employ switch allocators to schedule packets across the full crossbar. Since a full crossbar is not required in a fully-connected router (insight #1), one can do away with the switch allocator as well and replace it with simple arbitration.

Insight #3 **- A virtual channel allocator (VA) is not required in a fully-connected network router:** Virtual channels (VCs) are employed in routers to avoid head-of-line (HOL) blocking and prevent (protocol/ routing) deadlocks. Head-of-line blocking can severely degrade the throughput performance of low-radix networks because packets going to different destinations share intersecting paths in the network. Hence, any packet that fails to make forward progress (due to congestion) can block subsequently enqueued packets intending to go out via different output ports. This blocking issue disappears in a fully-connected network because every node has the ability to communicate directly with any other node in the network using dedicated unshared channels. Thus, the only reason a fully-connected network requires VCs is to prevent deadlocks. In this case, packets can be assigned VCs *statically* during the route computation stage removing the need for 'on-the-fly' allocation using a virtual channel allocator.

Insight #4 **- Dedicated credit-lines only provide marginal performance gains in a fully-connected network:** Compared to low-radix networks, packets in a fully-connected net-

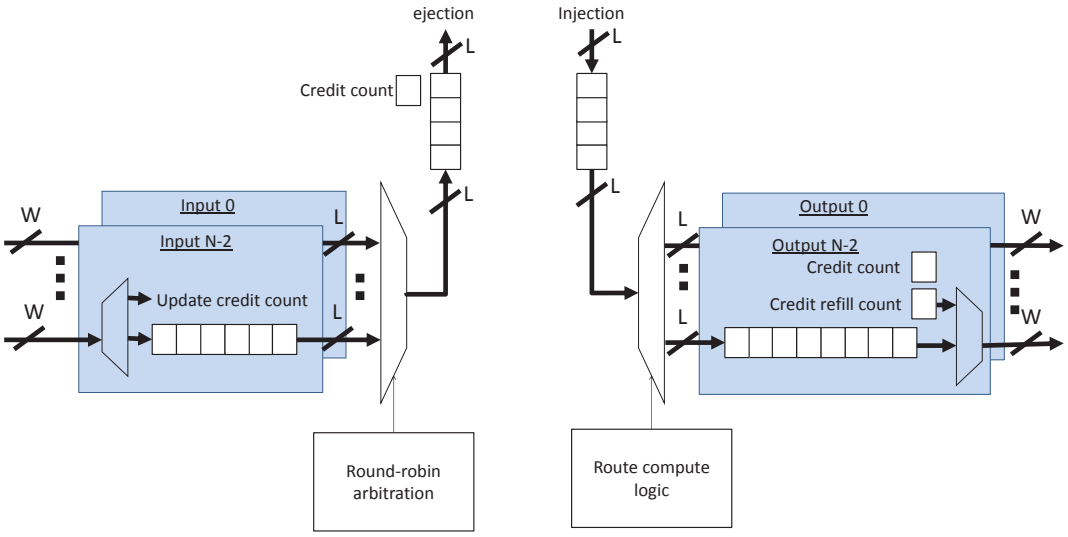


Figure 5.7: Microarchitecture of the minimal router (MR).

work arrive at much lower rates (see insight #1 above). In an N-node fully-connected network, a packet is received into an input buffer after N cycles. Once the packet is received, it can leave the queue resulting in the generation of a credit. Thus a credit is generated at an input port once every N cycles. Even for moderate network sizes, N, this credit generation frequency is so low that credits can be easily conveyed in the data channels leading to minimal loss in throughput performance compared to dedicated credit-lines.

Based on these insights, novel router designs are devised for a fully-connected network that are considerably more cost-efficient than traditional routers. These designs are presented below.

5.4.2.1 Minimal Router (MR)

The minimal router (MR) is a 1-hop design that only supports minimal routing. The microarchitecture of the MR design is shown in figure 5.7. L-bit packets enter the injection port and are enqueued into a FIFO buffer. Route computation is performed on the packet at the head of this queue to determine the output port it must take to go to its intended destination. Once route computation is complete, a $N - 1$ sized demultiplexer moves

this packet to the appropriate output port queue to be sent out on the channel. Outgoing transmissions from an output port are flow-controlled using a credit count. Packets arriving at an input port are buffered in a queue and await their turn to be forwarded to the ejection port via a $N-1$ sized multiplexer. This packet forwarding is controlled by a round-robin arbiter of size $N-1$. Note that by leveraging insight #1, the MR design completely forgoes the need for a full crossbar switch. Instead, it relies on simple multiplexing/ de-multiplexing to move packets into and out of the network ports. Furthermore, insight #2 enables the use of a simple arbiter instead of a complex switch allocator.

When a packet leaves a queue belonging to an input port, the 'credit refill count' on the corresponding output port is incremented to reflect that a credit needs to be conveyed to the upstream router using the outgoing channel. A credit is conveyed as a single phit unit to the upstream router, and increments the 'credit count' field by one. Credits and outgoing packets are time-multiplexed on the channel using a round-robin arbiter. Specifically, a single credit is conveyed after every *full* packet transmission. At the upstream router, the first bit of a received phit indicates whether it is a credit or the *start* of a new packet. Finally, since local (injection/ejection) bandwidth is cheap, these ports use dedicated credit lines.

5.4.2.2 Forwarding Router (FR)

To enable UGAL (see section 5.3.1.3), a router must be able to support both minimal (1-hop) and non-minimal (2-hop) routing. The forwarding router (FR) discussed in this section is capable of just that. Figure 5.8 shows the microarchitecture of the FR router. Based on the routing decision of UGAL, a packet may proceed directly to the destination (1-hop) or be sent to an intermediate node which then forwards the packet to the intended destination (2-hops). When the next hop of a packet is an intermediate node, then this packet is said to be in 'phase-1' of its routing. Alternatively, a packet whose next hop is the final destination is considered to be in 'phase-2'. To avoid routing deadlock, these two phases of

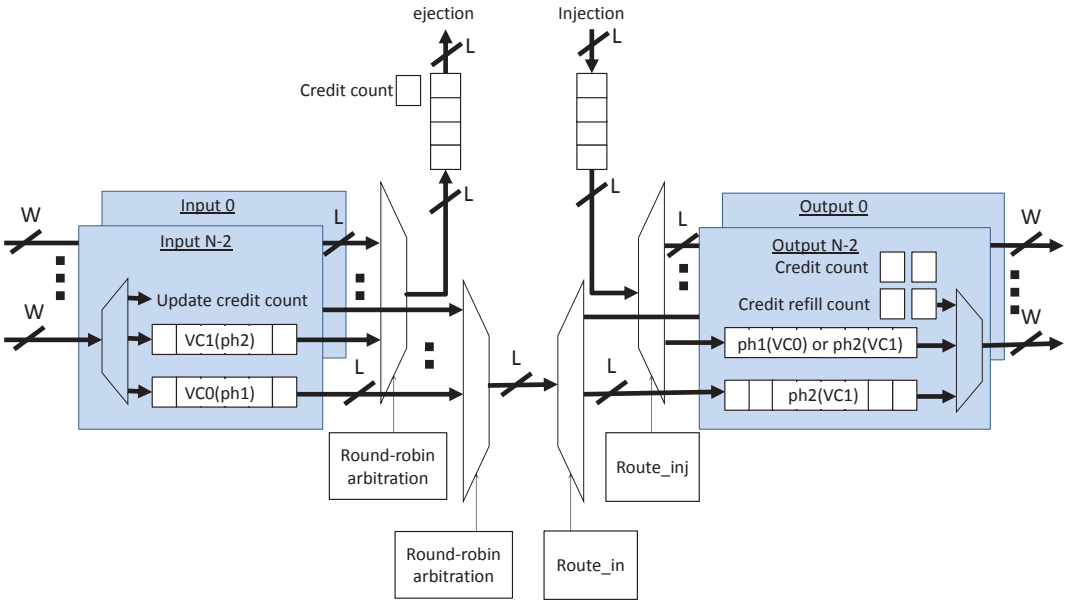


Figure 5.8: Microarchitecture of the forwarding router (FR).

communication need to kept in separate VCs at the input ports as shown in figure 5.8.

In the FR router, packets are assigned VCs *statically* at the route computation stage. Route computation is performed on packets at the head of the injection queue and queues belonging to VC0 (see figure 5.8). After route computation, a packet that is in phase-1 (next hop is an intermediate node) gets assigned VC0. Similarly, a packet in phase-2 (next hop is destination) gets assigned VC1. To explain how a packet progresses through the FR routers, it is instructive to consider an example at this point. Suppose route computation is performed on a packet at the head of the injection queue and UGAL routing decides that it is to be routed non-minimally (i.e. using two phases). Since, the packet is in phase-1 (heading to an intermediate node), it is assigned VC0. The packet arrives at the intermediate node and is enqueued in VC0. Route computation is performed on this packet once it reaches the head of the VC0 queue. The route computation transitions the packet into phase-2 (heading to destination now) assigning it VC1 in the process. Finally, the packet arrives in VC1 at the destination router and is ejected out.

Since the FR router employs two VCs, the 'credit count' and 'credit refill count' entries

contain two fields each (see figure 5.8). Credits are still conveyed as single phit units and two-bit binary encoding is employed to identify the VCs whose credit counts get incremented[5].

It is important to emphasize that, unlike a traditional input-queued router, the FR design does *not* require a full crossbar to forward packets from the input ports to the output ports. Instead, it is able to perform the forwarding functionality using a simple multiplexer-demultiplexer pair as shown in figure 5.8. This in turn, mitigates the need for a switch allocator leading to the use of simple arbitration to control the port-to-port forwarding.

5.4.3 Evaluation of the Router Designs

This section presents evaluation results that compare the performance, power and area consumption of the proposed routers (MR and FR) with a traditional input-queued router (IQR) design. The simulation parameters employed in our evaluation are provided in table 5.5. By choosing this configuration, the objective is to simulate a network design that can be deployed in a silicon photonic multichip system in the near-term.

Parameter	**Value**
Number of nodes (N)	64
Packet size (L)	256bits
Router frequency	5GHz
Photonic link frequency	10Gbps
Channel width ($W_{\mathtt{full}}$)	4bits/router cycle
Peak router bandwidth ($\mathtt{in} + \mathtt{out}$)	320GBps
Process technology	22nm LVT process ($\mathrm{Vdd} = 0.8\mathrm{V}$)

Table 5.5: Simulation parameters.

The 'booksim' cycle accurate simulator [23] is used to evaluate the performance of the routers. Packets are injected using a Bernoulli process. The simulator is warmed up until

[5] $0 \implies$ increment VC0 credit count; $1 \implies$ increment VC1 credit count; and, $2 \implies$ increment both VC0 and VC1 credit counts.

convergence is achieved before actual measurements (latency/ throughput) are made on the injected packets.

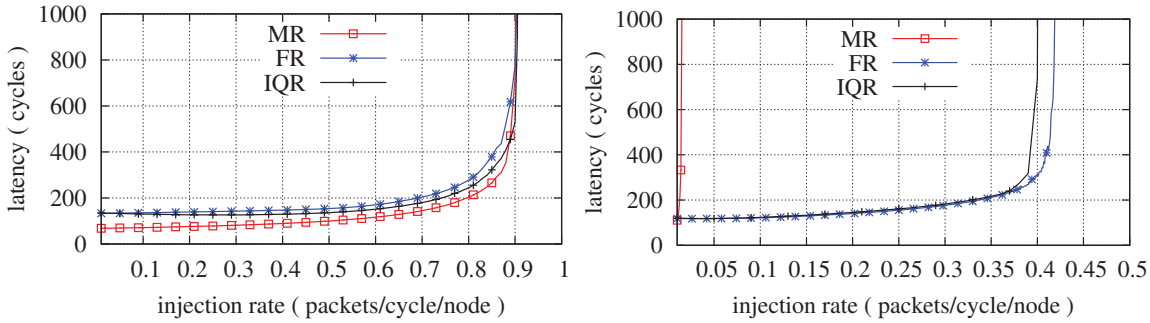
In order to do a fair comparison, a performance target is set for all the router designs. Then for each router model, the amount of resources it requires to achieve this performance goal is estimated. This resource usage is quantified using power and area analysis. The rest of the evaluation uses the following **target performance goal**: *Each router should be able to sustain* 90% *of peak throughput (capacity).*

5.4.3.1 Performance Evaluation

Figure 5.3 in section 5.3 showed the performance of the baseline input-queued router (IQR) on uniform random traffic when infinite [6] virtual channels (VCs), infinite buffers per VC and infinite output buffers are used. In this section, their values are swept to determine what configuration of the IQR enables one to meet the performance goal: 90% of peak sustainable throughput on uniform random traffic. To avoid routing deadlock in UGAL, the number of VCs at each input port is set to two. Then the number of buffers per VC are varied while keeping the output buffering as infinite. Through simulations, it is found that 16 buffers per VC are required to meet the target performance. Finally, it is determined that 16 output buffers are needed at each router output port. In addition to these buffering structures, the IQR design also requires: a 64×64 switch allocator, a 128×128 virtual-channel allocator and a 64×64 crossbar with ports of size 256b. These expensive structures lead to considerable power and area consumption in the IQR design (see section 5.4.3.2).

The proposed routers (MR and FR) are configured with sufficient buffering to meet the throughput goal and are simulated on both uniform random as well as permutation traffic patterns. The buffering requirements in the MR design are: 2 buffers per injection/ ejection queue, 3 buffers per input port queue and 32 buffers for the output port queues. Similarly

[6] By infinite, an extremely large value is implied.



(a) Favorable traffic (b) Adversarial traffic

Figure 5.9: Performance of the router designs.

the FR design requires 3 buffers per input VC, and two 16 buffer queues at the output ports (see figure 5.8). The injection/ ejection queues are the same size as the MR design.

Discussion: Figure 5.9 shows the performance of the routers under both benign and adversarial traffic conditions. From figure 5.9a, it can be observed that all router designs achieve the target throughput when configured with the buffering resources mentioned earlier. However in terms of latency, the MR design outperforms the FR and IQR routers. This discrepancy is due to the routing algorithms employed by the routers. The MR design only uses minimal routing while the FR and IQR routers employs UGAL which can send packets both minimally and non-minimally. Although no packets should be sent non-minimally under favorable conditions, our simulations shows that UGAL makes incorrect decisions on about 8% of total packets and sends them non-minimally leading to higher average latencies compared to the MR router. Permutation traffic patterns such as transpose traffic represent adversarial traffic conditions for the fully connected network. Under these conditions (see figure 5.9b), the performance of the MR design is limited to only 1/64 packets/cycle/node. However, by using non-minimal routing, the FR and IQR design can load-balance the network channels and achieve much higher throughput performance ($\approx$ 0.42 packets/cycle/node for FR). Similar performance results were observed on other well-known permutation patterns.

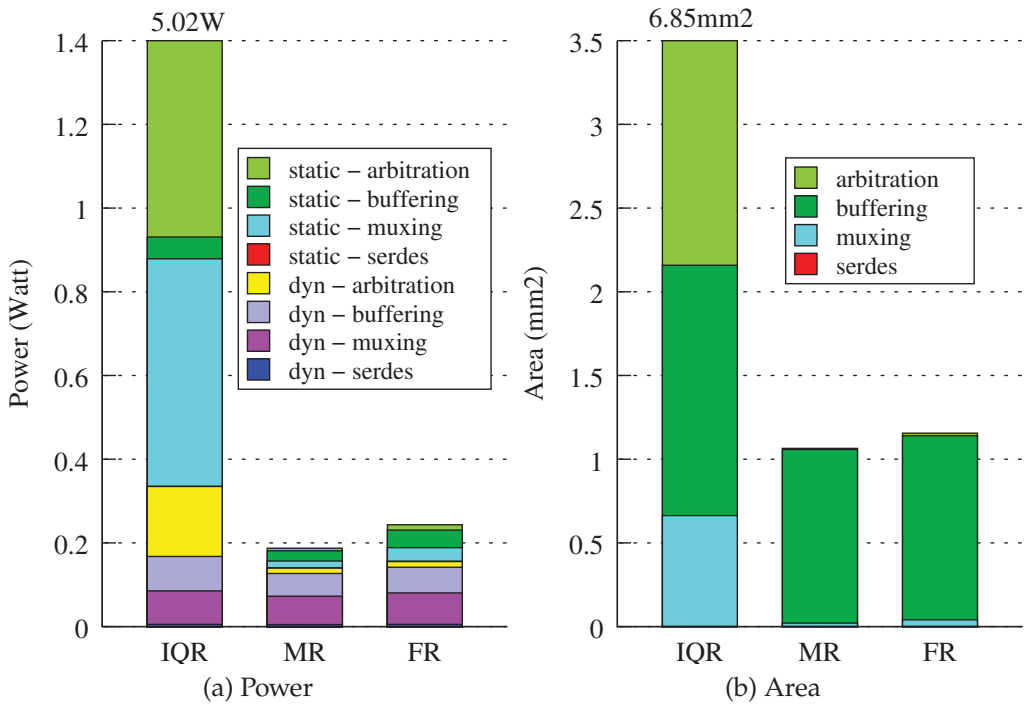


Figure 5.10: Power and area comparison of the designs.

5.4.3.2 Power and Area Evaluation

This section discusses the power and area consumption of the router designs. A 22nm technology process is targeted for the evaluation as shown in table 5.5. For each router component, its static power, dynamic energy and area consumption is estimated using the DSENT [89] modeling tool. SRAM-based buffers are assumed for the router queues and CACTI [98] is used to estimate the power, energy and area of these structures. The dynamic power consumption of a router model is estimated by collecting activity factors for its various components from the cycle-accurate simulator while it is operating at the target throughput (90% of capacity). For ease of presentation, the power/area numbers are lumped into four categories: *arbitration* (allocators, arbiters), *buffering* (queues, pipeline registers), *muxing* (crossbars, multiplexers, demultiplexers) and *SERDES* (serializer, de-serializer circuits).

Discussion: From figure 5.10, it can be seen that the traditional input-queued router (IQR) consumes about 20× more power and about 6× more area than the proposed routers (MR and FR). The bulk of the power/area consumption in the IQR design (see figures 5.10a, 5.10b) is due to the allocators and the crossbar switch. This is because the complexity of these structures scale quadratically with the number of ports [44]. Therefore they become prohibitively expensive in a fully-connected network. These expensive structures are replaced by simple arbitration and multiplexers/de-multiplexers in our designs. Hence, by adopting a topology-aware design approach, the expensive structures found in traditional routers are completely avoided leading to considerable savings in power and area. Figure 5.10 shows that the IQR design employs more buffering than the proposed routers. This is because the topology-agnostic IQR design puts down the same amount of buffering at each router port regardless of its utilization requirements. In a fully-connected network, since packets arrive at different rates at the router ports (see insights in section 5.4.2), provisioning the same amount of buffering for all ports causes underutilization leading to higher power and area consumption.

5.5 Quality-of-Service (QoS) Guarantees

Virtualization is important for server consolidation and enables better utilization of system resources. A basic requirement of virtualization is *isolation*. Ideally, isolation has to be provided both at the software and hardware level to minimize interference between the virtual machines (VMs). Providing isolation at the hardware level ensures fairness in resource usage and prevents any one virtual machine from hogging up the system resources either inadvertently or maliciously. For hardware isolation to be effective, it has to be provided at all levels, i.e. processor, memory and network. A system employing a fully-connected point-to-point topology is ensured fairness at the network level. Every node

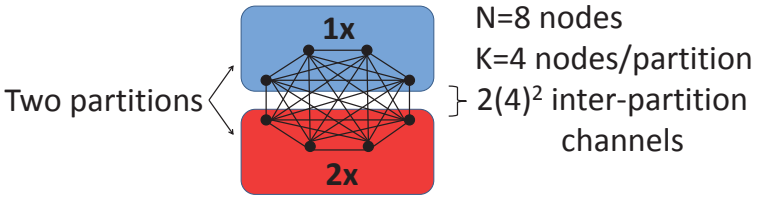


Figure 5.11: An 8-node fully-connected network is partitioned into two halves. Each line in this topology figure represents two unidirectional channels. In the absence of any inter-partition communication, the 32 channels that cross the partitioning boundary can be used to double the bandwidth in the bottom partition of the network.

has dedicated channels to other nodes in the network. Hence, a fair uniform level of bandwidth is guaranteed to each node of a virtual machine (VM). However, this section shows that a fully-connected network can be easily reconfigured to provide *differentiated quality-of-service (QoS)* to the virtual machines. Differentiated QoS enables *varied* levels of bandwidth guarantees in the network.

A possible use scenario for differentiated QoS is in the cloud-computing space where the infrastructure is offered to customers as a service such as Amazon's Elastic Compute Cloud (EC2) [5]. Amazon offers different pricing options for different levels of CPU/memory service. Extending this further, if differentiated QoS is available at the network level, then different pricing options can be offered to customers (VMs) for different levels of guaranteed bandwidth service.

5.5.1 Differentiated QoS

To see how differentiated QoS can be provided on a fully-connected network let us consider an N-node system. There are N^2 channels in the network[7]. Suppose this system is partitioned into two portions, a $N - K$ node portion and a K node portion then it can be

[7]The exact number is $N(N-1)$. However, this discrepancy does not affect the results of this section.

seen that:

$$
\begin{aligned}
N^2 &= (N-K+K)^2 \\
&= \underbrace{(N-K)^2}_{N-K \text{ partition channels}} + \underbrace{K^2}_{K \text{ partition channels}} + \underbrace{2(N-K)(K)}_{\text{inter-partition channels}}
\end{aligned}
$$

Now, if there is *no inter-partition communication* then the channels that cross the partitioning boundary are not utilized and lead to wasted bandwidth in the network. These unused channels can be reconfigured to bolster bandwidth in different portions of the network. To illustrate, consider the network partitioning example shown in figure 5.11. In this example, the network is partitioned into two halves i.e. $K = N/2$. With this partitioning, the $2(N/2)^2$ channels that cross the partitioning boundary go unused. However, these inter-partition channels can be used to *double* the bandwidth in one half of the network (the bottom half in figure). To explain how this is accomplished, suppose a virtual machine is mapped to the bottom half of the network and a sender node s wants to communicate with a destination node d in this VM. This sender can double its throughput to d by using its dedicated channel to d and by forwarding packets to d via an intermediate node in the top partition. Since, each forwarding requires two channels and $(N/2)^2$ forwardings are needed to double the bandwidth of the bottom half, the unused $2(N/2)^2$ inter-partition channels can be employed to accomplish this task. This establishes an important result:

Observation 5.1. *Every equal sized partitioning of the fully-connected network results in enough unused inter-partition channels to provide a* 100% *increase in bandwidth in any one of the partitions*

Using observation 5.1, many partitioning scenarios can be constructed for the fully-connected network as shown in figure 5.12. A hypervisor can be used to setup (program) the forwarding paths and map virtual machines to the appropriate bandwidth regions.

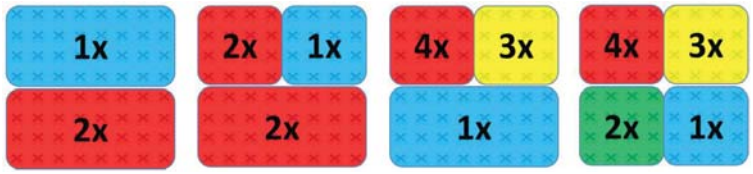


Figure 5.12: Different bandwidth regions can be realized in a partitioned fully-connected network.

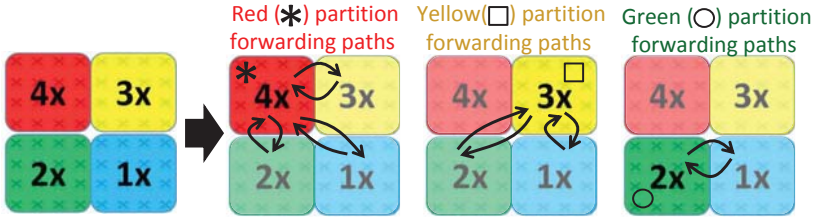


Figure 5.13: Forwarding paths for a 4-way partitioning example.

Illustrative Example: Figure 5.13 shows the forwarding paths for one example QoS partitioning. Consider the nodes in the red VM. These nodes can communicate with each other at 4× the bandwidth. This is because in addition to their direct channels, nodes in the red VM can forward traffic via intermediate nodes in three other partitions (yellow, blue and green) resulting in a node-to-node bandwidth gain of 4×. Note that the yellow VM cannot use the inter-partition channels between itself and the red VM because they have been allocated to the red VM. Hence, nodes in the yellow VM can only achieve a bandwidth gain of 3×. Similarly, nodes in the green partition can forward traffic via blue partition nodes leading to a 2× bandwidth gain.

5.5.2 Programmable Router (PR)

This section presents a programmable router (PR) that can support both minimal (1-hop) routing as well as 'circuit-switched' style (2-hop) forwarding. The microarchitecture of the PR design is shown in figure 5.14. The PR router employs a phit-sized crossbar to provide the forwarding functionality. This crossbar can be configured (programmed) by a hypervisor to setup the forwarding paths in a circuit-switched fashion such that all

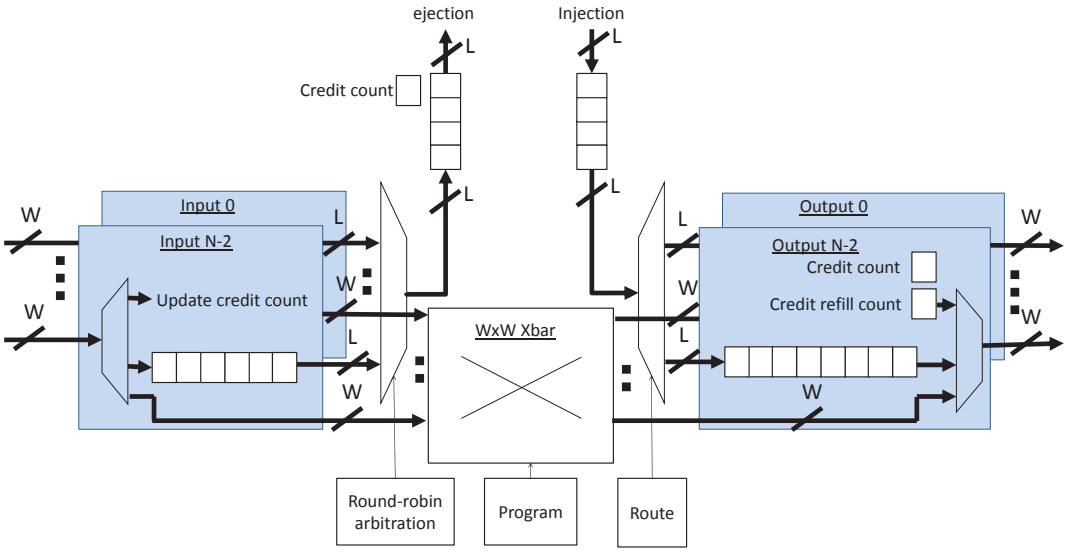


Figure 5.14: Microarchitecture of the programmable router (PR).

buffering and arbitration components of the intermediate router are completely bypassed. Thus, this design exhibits lower latencies compared to the FR router of section 5.4.2.2. Another opportunity afforded by the PR router is that the buffers at the intermediate nodes can be power-gated (since they are bypassed) leading to energy-savings in the network.

Note that the forwarding between any two nodes in a network employing PR routers has to be *symmetric*. That is, if a node s communicates with d using an intermediate node i ($s \rightarrow i \rightarrow d$), then d must communicate with s using the *same* intermediate node i ($d \rightarrow i \rightarrow s$). This symmetry is required to maintain proper credit flow between the two nodes s and d. In the differentiated QoS application, this symmetry condition can be easily met. However, if there is some usage scenario where the symmetry condition proves to be problematic, then map tables can be employed that relate data ports with the appropriate credit ports.

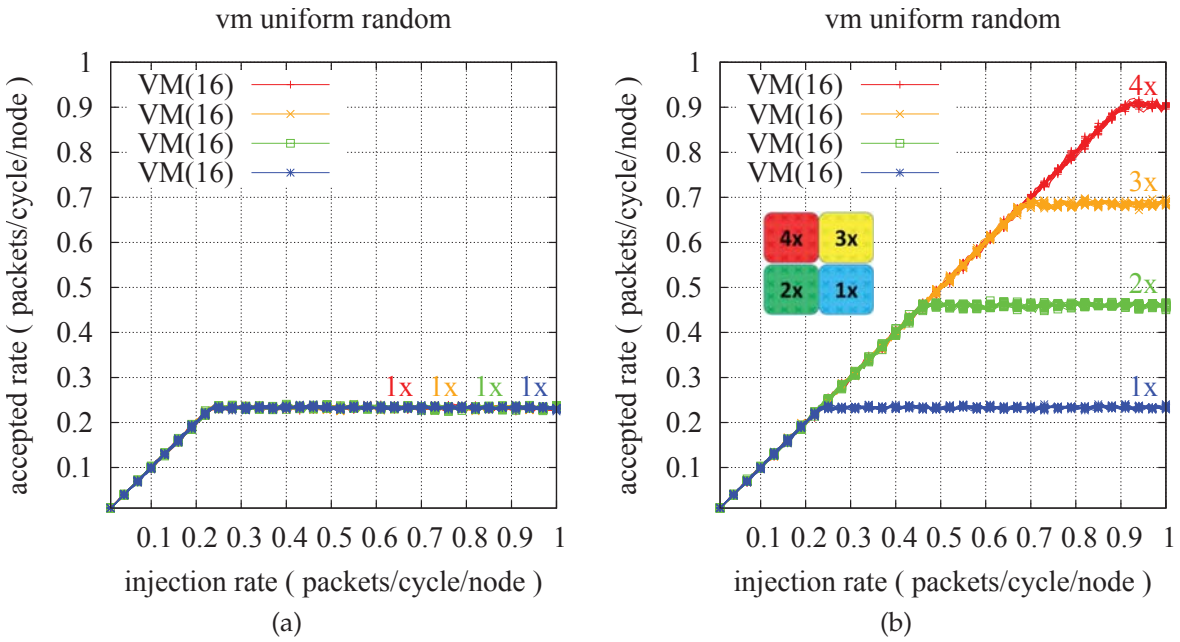


Figure 5.15: Throughput performance (a) without and (b) with differentiated QoS guarantees.

5.5.3 Evaluation of Differentiated QoS

In section 5.5.1, it was shown that a multi-node cluster employing a fully-connected network can be reconfigured to provide differentiated QoS when partitioned into multiple virtual machines (VMs). Differentiated QoS enables multiple levels of bandwidth to be supported in the network. To highlight these QoS gains, a variant of the uniform random traffic pattern called 'virtual machine uniform random (VM-UR)' was devised. In this traffic pattern, senders belonging to a VM only pick other nodes in the VM as their random destinations such that there is no inter-VM communication in the network. The proposed programmable router (PR) design is used to reconfigure the bandwidths in the fully-connected network.

A 64-node cluster is partitioned into four VMs with 16 nodes each and the VM-UR synthetic pattern is used to generate traffic inside a VM. Figure 5.15a shows the throughput performance of the VMs without bandwidth reconfiguration. It can be seen that all VM nodes exhibit similar bandwidth performance and saturate at an offered load of

0.25packets/cycle/node. Now using the proposed network partitioning methodology developed in section 5.5.1, the 64-node cluster is reconfigured into four bandwidth regions: $1\times$, $2\times$, $3\times$ and $4\times$. Figure 5.15b shows the throughput performance of the VMs when they are mapped into these bandwidth regions. In this case, it can be seen that each VM exhibits throughput performance proportional to the bandwidth level guaranteed to it. Finally, as highlighted in section 5.5.2, power-gating buffers at the forwarding ports in the PR design leads to $\approx 5\%$ savings in router power for this cluster partitioning.

5.6 Summary

The topology of an interconnection network has a profound impact on performance. Silicon photonic technology enables designers to explore rich topologies that are considered too complex (from a packaging standpoint) in traditional high-performance networks. However, it imposes new constraints as well which must be considered during network design. This chapter presented a thorough analysis of electrical switching within the constraints of silicon photonic technology. A wide range of network designs were evaluated and it was shown that a fully-connected network provides the best performance under both benign and adversarial traffic conditions. Traditional routers employ allocators and crossbars whose complexity scales quadratically with the radix making them too prohibitive for a fully-connected network. To overcome the scalability issues of a generic router, this work adopted a 'topology-aware' design approach and proposed novel router designs that do not use allocators or crossbars. Through detailed analysis, it was demonstrated that the proposed routers provide greater than 80% savings in area and power. Finally, a mechanism to incorporate differentiated QoS guarantees on a partitioned fully-connected network was presented which can easily be deployed in a cloud-computing cluster.

6 CONCLUSION

Multicore processor chips are commonplace. Processor vendors continue to add new chips to their product portfolios that offer higher core counts in newer generations targeting market segments as varied as mobile hand held devices to large-scale server systems. Putting logic units on the same die has two main advantages. First, since the logic units are co-located on the same die, the communication distances are small leading to low message latencies. Second, the high-bandwidth densities of on-chip wires coupled with multiple metal layers provide ample bandwidth necessary to sustain large amounts of communication. Unfortunately, as discussed in chapter 1, there is growing evidence that scaling a single chip to very high core counts will lead to: increasing fabrication costs, low process yields, power delivery and heat removal limitations. Due to these cost and yield concerns, a system designer would ideally like to package together multiple smaller die that are easy to fabricate using a communication technology that enables the same performance level as a single large monolithic piece of silicon. Unfortunately, electrical interconnects are too slow and energy inefficient for communicating over distances longer than a few centimeters.

Silicon photonics is an emerging technology that *offers* seamless integration of multiple chips with high bandwidth density, 'speed-of-light' communication at lower energy-per-bit consumption compared to electrical interconnects. Leveraging these opportunities, this thesis makes several contributions in different aspects of photonic network design for multichip systems. A summary of these contributions is provided in section 6.1. Future research directions are discussed in section 6.2. Personal thoughts and reflections on this research are presented in section 6.3. Finally, some concluding thoughts and remarks are provided in section 6.4.

6.1 Summary

When designing an interconnect, the topology chosen by the architect ultimately determines its performance – generally measured in terms of latency and throughput – as well as the cost, e.g. the number of channels and switches required to build the network. As explained in this dissertation, the topology in a photonic network also determines its laser power consumption requirements. Considering the optical losses of photonic components and efficiencies of laser sources in the current technology generation, optimizing for laser power consumption is a first-order design constraint in photonic networks. Topologies in photonic networks can be classified into two categories depending on the maximum number of hops required to convey a message. Optical crossbars or channel sharing networks are 1-hop whereas communication in path sharing or switched network designs can take multiple ($\geqslant$ 1) hops. This thesis investigated both classes of photonic networks and developed several new insights and architectures.

Channel sharing designs were the first category of networks considered in this thesis. In channel sharing architectures, multiple senders and/ or receivers share access on the network channels. The simplest network that fits into this category is a minimally (1-hop) routed fully-connected, point-to-point topology. In a fully-connected topology, each network channel has one sender and one receiver. Thus, a point-to-point topology represents a *degenerate* case in which there is no actual sharing on the network channels. A fully-connected network provides arbitration-free, non-blocking access to channels but suffers from low node-to-node (port) bandwidth. Alternatively, sharing between nodes on the network channels can be increased to improve inter-node bandwidth but this leads to higher laser power consumption. To investigate this performance-power trade-off, this dissertation developed an analytical model to demonstrate the limits of channel sharing under a fixed laser (optical) power budget, and quantify its performance benefits over a point-to-point network. Using this model, it was demonstrated that under realistic device

loss characteristics, sharing on the network channels should be restricted to three or fewer senders per channel. Based on this analysis, a novel channel sharing architecture called wavelength stealing was proposed. The topology of the wavelength stealing architecture is similar to the point-to-point network except for the fact that multiple senders share a network channel. One of the senders is called the 'owner' while the other senders are called the 'stealers'. In the wavelength stealing architecture, the owner is guaranteed access on the channel in a non-blocking manner and the stealers are allowed access only when the owner is not active. Another feature of the wavelength stealing architecture is that the stealer nodes access the channel opportunistically in an arbitration-free manner. That is, these nodes do not participate in any arbitration or wait for permissions to access the channel. This enables the wavelength stealing architecture to achieve lower latencies and higher throughput compared to arbitration-based designs. Evaluation of the wavelength stealing architecture on a 64-chip macrochip system revealed that this design provides up to 28% better energy-delay performance compared to the point-to-point network on some high-performance-computing (HPC) applications. Furthermore, this dissertation developed a novel quality-of-service (QoS) mechanism for enabling performance isolation between multiple virtual machines (VM) running on a partitioned multichip system that employs the wavelength stealing interconnect. The proposed QoS algorithm can be deployed in the hypervisor such that it can map the VMs to the appropriate regions in the network to provide the desired isolation guarantees.

Switched photonic networks were also explored as part of this thesis. Due to the high device loss of optical switches that can be fabricated in the current technology generation, this dissertation only focused on electrical switching in photonic networks. Switched networks are commonplace in computer systems. However, as demonstrated in this thesis, silicon photonic technology imposes new constraints and provides unique opportunities that are quite different from traditional electrical networks requiring a fresh look at electri-

cal switching within the purview of this new technology. Hence, this thesis provided a thorough evaluation of popular electrically-switched networks within the constraints of silicon photonic technology. It was demonstrated that photonic topology imposes a fixed cost per network channel as opposed to traditional electrical networks where channels are assigned different costs depending on their lengths. Different categories of topologies were considered in this evaluation (direct/ indirect, low-/ high-radix) and it was demonstrated that an adaptively routed fully-connected topology in which packets can take non-minimal (2-hop) routes provides the highest performance under both favorable and adversarial traffic conditions. Since, the logic complexity of traditional input-queued, virtual-channel routers scale quadratically with the number of ports, these routers become prohibitively complex for fully-connected networks. To design efficient routers, this dissertation adopted a topology-aware design approach and developed insights that show that the expensive logic structures (crossbar switch, switch allocator and virtual-channel allocator) needed for other topologies are not required in routers designed for the fully-connected topology. Leveraging these insights, this thesis proposed novel router designs that consume significantly less area and power than a traditional input-queued router while achieving similar performance. A novel QoS mechanism to provide differentiated bandwidth guarantees in the network is also developed as part of this thesis. Using the proposed mechanism, the bandwidth in a fully-connected topology can be re-configured to realize multiple regions with different throughput guarantees. A hypervisor can be used to provide this reconfiguration functionality and map a VM to a network region whose throughput guarantee is consistent with the bandwidth demands of that VM.

6.2 Future Work

This section surveys some potential avenues for future work that extend the research conducted in this thesis.

6.2.1 Extending Wavelength Stealing Architecture

The wavelength stealing interconnect – proposed as part of this thesis – is presented in chapter 4. Multiple senders share access on a network channel in this architecture. Specifically, one of the sender nodes on the channel is called the 'owner' and the other senders are called the 'stealers' of that channel. However, as explained in chapter 4, due to link loss considerations, the sharing in this architecture is restricted to just two senders per channel, i.e. there is one owner and one stealer. To coordinate the activities of these two senders, the control mechanism presented in section 4.2 employs erasure coding and some special control wavelengths per channel. Going forward, when the optical losses of photonic components improve, then sharing beyond two senders may be more effective from a power-performance standpoint. In this case, the proposed control mechanism will need to be extended to provide the correct functionality. Extending erasure coding to higher capability is straightforward as there are methods described in literature to generate erasure codes of a desired strength. However, designing the functionality of the control wavelengths is not straightforward and needs to be worked out for higher degrees of sharing.

6.2.2 Design of Multi-macrochip Systems

The scalability of the fully-connected topology was explored in chapter 5 for the macrochip system. Scaling the fully-connected topology to higher node counts requires bandwidth to scale *quadratically* with the number of nodes. Using conservative estimates and projections,

it was shown that the scalability of the macrochip system is limited to 128 chips when targeting an inter-chip port bandwidth of 20Gbps. To scale to higher chip counts, a 'multi-macrochip' approach has to be adopted. This opens up a plethora of design challenges. For example, in a single macrochip system, the optical fibers connected along the perimeter are used to deliver optical power for two purposes: intra-macrochip communication and connecting to I/O devices. Moving to a multi-macrochip system, the optical fibers have to be provisioned for inter-macrochip communication as well. Since, the number of optical fibers that can be connected along the perimeter of a macrochip is limited, optical fibers devoted to intra- and inter-macrochip communication as well as I/O have to be carefully partitioned. This partitioning will affect the type of inter-macrochip topologies that can be realized given a desired system size. These considerations need to be investigated further to understand the trade-offs involved in mutli-macrochip design.

6.3 Reflections

This section presents my opinions and thoughts on silicon photonic technology and optical network design. It should be emphasized that these opinions may not reflect those of the my collaborators and co-authors. Needless to say, they may change at anytime in the future.

6.3.1 The Challenge of High Static Power Consumption

Channels in photonic networks are sourced with laser power regardless of whether they are being utilized or not. This is one of the reasons why the bulk of the power consumption in photonic networks is static power. To make matters worse, due to the high optical losses in the current technology generation, this static power cost is rather significant. To justify these high activity-agnostic power costs, this dissertation argued that photonic technology should be deployed in high utilization scenarios. Herein lies one of the main challenges

with this technology. It is well known that network traffic exhibits bursty behavior [10, 86, 32]. Thus there is significant down time on the network channels followed by periods of activity. However, due to the lack of energy-proportionality with network activity, this type of traffic behavior is ill-suited for photonic networks from a power-performance standpoint. Moving forward, I can see two possible ways these inefficiencies could be overcome:

- First, optical devices could improve significantly in terms of their losses, efficiencies and/ or tuning power requirements leading to a reduction in the static power consumption of photonic networks. This can either increase the energy-efficiency to such an extent that the overall power consumption becomes less of an issue or it can skew the dominant factor in the power consumption towards the dynamic component making photonic networks more energy-proportional.

- Second, efficient techniques to turn-off the laser sources during periods of down time, or divert light to only those network regions that have active communication could become viable.

6.3.2 Need Photonic Technology Roadmap

One of the biggest challenges I encountered in conducting research in nanophotonics is the lack of a technology roadmap or consensus in the device losses of optical components. Thus, the values of device losses assumed in prior papers have varied greatly leading to designs that show fundamentally different trade-offs. This creates confusion for researchers trying to learn about prior work and causes frustration for scientists trying to publish in this area. Due to this reason, I believe that all interested parties – be it industry or academia – should take part in an extensive roadmapping effort for silicon photonic technology. To learn more about the merits of undertaking this effort, I refer any interested reader to an excellent commentary paper written by Kirchain and Kimerling [45].

6.4 Closing Remarks

Optical technology has been extremely successful in long distance communication and has been instrumental in bringing about the telecommunication revolution (telephone, internet etc.). For a long time, the costs associated with optical technology prevented its mass adoption in short reach applications. However, through exceptional pace in innovation, researchers were able to fully integrate optical devices with a complementary-metal-oxide-semiconductor (CMOS) process. This meant that industry could leverage the economies and infrastructure of CMOS to fabricate photonic devices that could solve some of the problems facing computing systems today.

This dissertation has investigated one application of silicon photonic technology – providing seamless integration of multiple chips with high bandwidth-density communication. However, it is important to emphasize that the findings and contributions of this dissertation go beyond just multichip networks. That is, the solutions presented in this thesis could be applied to any silicon photonic interconnect, be it for on-chip communication in a multi- or many-core setting or between servers in a large-scale system.

I firmly believe that silicon photonic technology holds a promising future. Regardless of whether I will get a chance to work on it again in my professional career, I plan to follow any progress and development in this exciting field for years to come.

BIBLIOGRAPHY

[1] Vikas Agarwal, M. S. Hrishikesh, Stephen W. Keckler, and Doug Burger. "Clock Rate Versus IPC: The End of the Road for Conventional Microarchitectures". In: *Proceedings of the 27th Annual International Symposium on Computer Architecture*. ISCA '00. New York, NY, USA: ACM, 2000, pp. 248–259.

[2] Jung-Ho Ahn, Sungwoo Choo, and J. Kim. "Network within a network approach to create a scalable high-radix router microarchitecture". In: *High Performance Computer Architecture (HPCA), 2012 IEEE 18th International Symposium on*. 2012.

[3] Jung Ho Ahn, Nathan Binkert, Al Davis, Moray McLaren, and Robert S. Schreiber. "HyperX: topology, routing, and packaging of efficient large-scale networks". In: *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*. SC '09.

[4] Vilson R. Almeida, Roberto R. Panepucci, and Michal Lipson. "Nanotaper for compact mode conversion". In: *Opt. Lett.* 28.15 (2003), pp. 1302–1304.

[5] *Amazon Elastic Compute Cloud*. `http://aws.amazon.com/ec2/`.

[6] R Anigundi, Hongbin Sun, Jian-Qiang Lu, K. Rose, and Tong Zhang. "Architecture design exploration of three-dimensional (3D) integrated DRAM". In: *Quality of Electronic Design, 2009. ISQED 2009. Quality Electronic Design*. 2009, pp. 86–90.

[7] Krste Asanovic, Ras Bodik, Bryan Christopher Catanzaro, Joseph James Gebis, Parry Husbands, Kurt Keutzer, David A. Patterson, William Lester Plishker, John Shalf, Samuel Webb Williams, and Katherine A. Yelick. *The Landscape of Parallel Computing Research: A View from Berkeley*. Tech. rep. UCB/EECS-2006-183. EECS Department, University of California, Berkeley, 2006. URL: `http://www.eecs.berkeley.edu/Pubs/TechRpts/2006/EECS-2006-183.html`.

[8] M. Asghari and A. V. Krishnamoorthy. "Silicon photonics: Energy-efficient communication". In: *Nature Photonics* (2011).

[9] Semiconductor Industries Association. *The International Technology Roadmap for Semiconductors (ITRS)*. `http://www.itrs.net/Links/2008ITRS/home2008.htm`.

[10] Mario Badr and Natalie Enright Jerger. "SynFull: Synthetic Traffic Models Capturing Cache Coherent Behaviour". In: *Proceedings of the International Symposium on Computer Architecture*. 2014.

[11] D. H. Bailey, E. Barszcz, J. T. Barton, D. S. Browning, R. L. C̃arter, L. Dagum, R. A. Fatoohi, P. O. Frederickson, T. A. Lasinski, R. S. S̃chreiber, H. D. Simon, V. Venkatakrishnan, and S. K. Weeratunga. "The NAS parallel benchmarks – summary and preliminary results". In: *Proc. of the ACM/IEEE conference on Supercomputing*. Supercomputing '91. New York, NY, USA, 1991.

[12] G. Balamurugan, J. Kennedy, G. Banerjee, J.E. Jaussi, M. Mansuri, F. O'Mahony, B. Casper, and R. Mooney. "A Scalable 5-15Gbps, 14-75mW Low Power I/O Transceiver in 65nm CMOS". In: *VLSI Circuits, 2007 IEEE Symposium on*. 2007, pp. 270–271.

[13] A Barkai, Ansheng Liu, Daewoong Kim, R. Cohen, N. Elek, Hsu-Hao Chang, B.H. Malik, R. Gabay, R. Jones, M. Paniccia, and Nahum Izhaky. "Efficient Mode Converter for Coupling between Fiber and Micrometer Size Silicon Waveguides". In: *Group IV Photonics, 2007 4th IEEE International Conference on*. 2007.

[14] B. Ben Bakir, A. Descos, N. Olivier, D. Bordel, P. Grosse, J.L. Gentner, F. Lelarge, and J.-M. Fedeli. "Hybrid Si/III-V lasers with adiabatic coupling". In: *Group IV Photonics (GFP), 2011 8th IEEE International Conference on*. 2011.

[15] K. Bernstein, P. Andry, J. Cann, P Emma, D. Greenberg, W. Haensch, M. Ignatowski, S. Koester, J. Magerlein, R. Puri, and A Young. "Interconnects in the Third Dimension: Design Challenges for 3D ICs". In: *Design Automation Conference, 2007. DAC '07. 44th ACM/IEEE*. 2007.

[16] N. Binkert, A Davis, N.P. Jouppi, M. McLaren, N. Muralimanohar, R. Schreiber, and Jung Ho Ahn. "The role of optics in future high radix switch design". In: *Computer Architecture (ISCA), 2011 38th Annual International Symposium on*. 2011.

[17] B. Black, M. Annavaram, N. Brekelbaum, J. DeVale, Lei Jiang, G.H. Loh, D. McCauley, P. Morrow, D.W. Nelson, D. Pantuso, P. Reed, J. Rupley, Sadasivan Shankar, J. Shen, and C. Webb. "Die Stacking (3D) Microarchitecture". In: *Microarchitecture, 2006. MICRO-39. 39th Annual IEEE/ACM International Symposium on*. 2006.

[18] Erich Bloch. "The Engineering Design of the Stretch Computer". In: *Papers Presented at the December 1-3, 1959, Eastern Joint IRE-AIEE-ACM Computer Conference*. IRE-AIEE-ACM '59 (Eastern). New York, NY, USA: ACM, 1959.

[19] Shekhar Borkar and Andrew A. Chien. "The Future of Microprocessors". In: *Commun. ACM* 54.5 (May 2011).

[20] George Z. Chrysos and Joel S. Emer. "Memory Dependence Prediction Using Store Sets". In: *Proceedings of the 25th Annual International Symposium on Computer Architecture*. ISCA '98. Washington, DC, USA: IEEE Computer Society, 1998.

[21] Mark J. Cianchetti, Joseph C. Kerekes, and David H. Albonesi. "Phastlane: a rapid transit optical routing network". In: *Proc. of the International Symposium on Computer Architecture*. ISCA '09. New York, NY, USA, 2009.

[22] R. Claps, D. Dimitropoulos, V. Raghunathan, Y. Han, and B. Jalali. "Observation of stimulated Raman amplification in silicon waveguides". In: *Opt. Express* 11.15 (2003), pp. 1731–1739.

[23] William Dally and Brian Towles. *Principles and Practices of Interconnection Networks*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003.

[24] M. Dinu, F. Quochi, and H. Garcia. "Third-order nonlinearities in silicon at telecom wavelengths". In: *Applied Physics Letters* 82.18 (2003), pp. 2954–2956.

[25] Po Dong, Shirong Liao, Hong Liang, Roshanak Shafiiha, Dazeng Feng, Guoliang Li, Xuezhe Zheng, Ashok V. Krishnamoorthy, and Mehdi Asghari. “Submilliwatt, ultrafast and broadband electro-optic silicon switches”. In: *Opt. Express* 18.24 (2010).

[26] Eric Dulkeith, Fengnian Xia, Laurent Schares, William M. J Green, and Yurii A. Vlasov. “Group index and group velocity dispersion in silicon-on-insulator photonic wires”. In: *Opt. Express* 14.9 (2006), pp. 3853–3863.

[27] Eric Dulkeith, Yurii A. Vlasov, Xiaogang Chen, Nicolae C. Panoiu, and Jr. Richard M. Osgood. “Self-phase-modulation in submicron silicon-on-insulator photonic wires”. In: *Opt. Express* 14.12 (2006), pp. 5524–5534.

[28] Hadi Esmaeilzadeh, Emily Blem, Renee St. Amant, Karthikeyan Sankaralingam, and Doug Burger. “Dark silicon and the end of multicore scaling”. In: *Proceedings of the 38th annual international symposium on Computer architecture*. ISCA ’11.

[29] R. L. Espinola, M. C. Tsai, James T. Yardley, and Jr. Osgood R.M. “Fast and low-power thermooptic switch on thin silicon-on-insulator”. In: *Photonics Technology Letters, IEEE* 15.10 (2003).

[30] L. H. Frandsen, P. I. Borel, Y. X. Zhuang, A. Harpøth, M. Thorhauge, M. Kristensen, W. Bogaerts, P. Dumon, R. Baets, V. Wiaux, J. Wouters, and S. Beckx. “Ultralow-loss 3-dB photonic crystal waveguide splitter”. In: *Opt. Lett.* 29.14 (2004), pp. 1623–1625.

[31] Eiji Fujiwara. *Code Design for Dependable Systems: Theory and Practical Application*. Wiley-Interscience, 2006.

[32] Mitchell Hayenga and Mikko Lipasti. “The NoX Router”. In: *Proceedings of the 44th Annual IEEE/ACM International Symposium on Microarchitecture*. MICRO-44. New York, NY, USA: ACM, 2011.

[33] Jim Held, Jerry Bautista, and Sean Koehl. *White Paper From a Few Cores to Many: A Tera-scale Computing Research Review*. `http://www.intel.com/content/dam/www/public/us/en/documents/technology-briefs/intel-labs-tera-scale-research-paper.pdf`.

[34] R. Ho, K.W. Mai, and M.A. Horowitz. “The future of wires”. In: *Proceedings of the IEEE* (2001).

[35] R. Ho, I. Ono, F. Liu, R. Hopkins, A. Chow, J. Schauer, and R. Drost. “High-Speed and Low-Energy Capacitively-Driven On-Chip Wires”. In: *Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International*.

[36] J. Howard, S. Dighe, S.R. Vangal, G. Ruhl, N. Borkar, S. Jain, V. Ẽrraguntla, M. Konow, M. Riepen, M. Gries, G. Droege, T. Lund-Larsen, S. Šteibl, S. Borkar, V.K. De, and R. Van Der Wijngaart. “A 48-Core IA-32 Processor in 45 nm CMOS Using On-Die Message-Passing and DVFS for Performance and Power Scaling”. In: *Solid-State Circuits, IEEE Journal of* (2011).

[37] I wei Hsieh, Xiaogang Chen, Jerry Dadap, Nicolae Panoiu, Richard Osgood, Yurii Vlasov, and Sharee McNab. "Cross-Phase Modulation in Si Photonic Wire Waveguides". In: *Conference on Lasers and Electro-Optics/Quantum Electronics and Laser Science Conference and Photonic Applications Systems Technologies*. Optical Society of America, 2006, CWD6.

[38] "IEEE standards for local area networks: Token ring access method and physical layer specifications". In: *IEEE Std* 802.5 — 1989 (1989).

[39] D.R. Johnson, M.R. Johnson, J.H. Kelm, W. Tuohy, Steven S. Lumetta, and S.J. Patel. "Rigel: A 1,024-Core Single-Chip Accelerator Architecture". In: *Micro, IEEE* (2011).

[40] Ajay Joshi, Christopher Batten, Yong-JI Kwon, Scott Beamer, Imran Shamim, Krste Asanović, and Vladimir Stojanović. "Silicon-photonic Clos networks for global on-chip communication". In: NOCS '09. New York, NY, USA: ACM, 2009.

[41] Byungsub Kim and V. Stojanovic. "A 4Gb/s/ch 356fJ/b 10mm equalized on-chip interconnect with nonlinear charge-injecting transmit filter and transimpedance receiver in 90nm CMOS". In: *Solid-State Circuits Conference - Digest of Technical Papers, 2009. ISSCC 2009. IEEE International*. 2009.

[42] J. Kim, W.J. Dally, S. Scott, and D. Abts. "Technology-Driven, Highly-Scalable Dragonfly Topology". In: *Computer Architecture, 2008. ISCA '08. 35th International Symposium on*.

[43] John Kim, William J. Dally, and Dennis Abts. "Flattened butterfly: a cost-efficient topology for high-radix networks". In: *Proceedings of the 34th annual international symposium on Computer architecture*. ISCA '07.

[44] John Kim, William J. Dally, Brian Towles, and Amit K. Gupta. "Microarchitecture of a High-Radix Router". In: *Proceedings of the 32nd annual international symposium on Computer Architecture*. ISCA '05.

[45] R. Kirchain and L. Kimerling. "A roadmap for nanophotonics". In: *Nature Photonics* (2007).

[46] Nevin Kirman and Jose F. Martinez. "A power-efficient all-optical on-chip interconnect using wavelength-based oblivious routing". In: ASPLOS '10. New York, NY, USA: ACM, 2010.

[47] Nevin Kirman, Meyrem Kirman, Rajeev K. Dokania, Jose F. Martinez, Alyssa B. Apsel, Matthew A. Watkins, and David H. Albonesi. "Leveraging Optical Technology in Future Bus-based Chip Multiprocessors". In: *Proc. of the International Symposium on Microarchitecture*. MICRO 39. Washington, DC, USA: IEEE Computer Society, 2006.

[48] A.V. Kirshnamoorthy, X. Zheng, G. Li, J. Yao, T. Pinguet, A. Mekis, H. Thacker, I. Shubin, L. Ying, K. Raj, and J.E. Cunningham. "Exploiting CMOS manufacturing to reduce tuning requirements for resonant optical devices". In: *IEEE Photonics Journal* 3 (3 2011).

[49] P. Koka, M.O. McCracken, H. Schwetman, C.-H.O. Chen, Xuezhe Zheng, Ron Ho, K. Raj, and A.V. Krishnamoorthy. "A micro-architectural analysis of switched photonic multi-chip interconnects". In: ISCA '12. 2012.

[50] Pranay Koka, Michael O. McCracken, Herb Schwetman, Xuezhe Zheng, Ron Ho, and Ashok V. Krishnamoorthy. "Silicon-photonic network architectures for scalable, power-efficient multi-chip systems". In: ISCA '10. 2010.

[51] A.V. Krishnamoorthy, Ron Ho, Xuezhe Zheng, H. Schwetman, Jon Lexau, P. Koka, GuoLiang Li, I. Shubin, and J.E. Cunningham. "Computer Systems Based on Silicon Photonic Interconnects". In: *Proceedings of the IEEE* (2009).

[52] AV. Krishnamoorthy, J.E. Cunningham, Xuezhe Zheng, I Shubin, J. Simons, Dazeng Feng, Hong Liang, Cheng-Chih Kung, and M. Asghari. "Optical Proximity Communication With Passively Aligned Silicon Photonic Chips". In: *Quantum Electronics, IEEE Journal of* 45 (2009).

[53] George Kurian, Jason E. Miller, James Psota, Jonathan Eastep, Jifeng Liu, Jurgen Michel, Lionel C. Kimerling, and Anant Agarwal. "ATAC: a 1000-core cache-coherent processor with on-chip optical network". In: PACT '10. New York, NY, USA: ACM, 2010.

[54] M. Lamponi, S. Keyvaninia, C. Jany, F. Poingt, F. L̃elarge, G. de Valicourt, G. Roelkens, D. Van Thourhout, S. M̃essaoudene, J.-M. Fedeli, and G.H. Duan. "Low-Threshold Heterogeneously Integrated InP/SOI Lasers With a Double Adiabatic Taper Coupler". In: *Photonics Technology Letters, IEEE* (2012).

[55] James Larus. "Spending Moore's Dividend". In: *Commun. ACM* 52.5 (May 2009).

[56] B.G. Lee, A Biberman, Po Dong, M. Lipson, and K. Bergman. "All-Optical Comb Switch for Multiwavelength Message Routing in Silicon Photonic Networks". In: *Photonics Technology Letters, IEEE* 20.10 (2008), pp. 767–769.

[57] Charles E. Leiserson. "Fat-trees: universal networks for hardware-efficient supercomputing". In: *IEEE Trans. Comput.* 34.10 (1985).

[58] Charles E. Leiserson, Zahi S. Abuhamdeh, David C. Douglas, Carl R. Feynman, Mahesh N. Ganmukhi, Jeffrey V. Hill, Daniel Hillis, Bradley C. Kuszmaul, Margaret A. St. Pierre, David S. Wells, Monica C. Wong, Shaw-Wen Yang, and Robert Zak. "The Network Architecture of the Connection Machine CM-5 (Extended Abstract)". In: *Proceedings of the Fourth Annual ACM Symposium on Parallel Algorithms and Architectures*. SPAA '92. San Diego, California, USA: ACM, 1992.

[59] Guoliang Li, Jin Yao, Hiren Thacker, Attila Mekis, Xuezhe Zheng, Ivan Shubin, Ying Luo, Jin hyoung Lee, Kannan Raj, John E. Cunningham, and Ashok V. Krishnamoorthy. "Ultralow-loss, high-density SOI optical waveguide routing for macrochip interconnects". In: *Opt. Express* 11 (), pp. 12035–12039.

[60] Guoliang Li, Jin Yao, Hiren Thacker, Attila Mekis, Xuezhe Zheng, Ivan Shubin, Ying Luo, Jin hyoung Lee, Kannan Raj, John E. Cunningham, and Ashok V. Krishnamoorthy. "Ultralow-loss, high-density SOI optical waveguide routing for macrochip interconnects". In: *Opt. Express* 20.11 (2012), pp. 12035–12039.

[61] Shu Lin and Daniel J. Costello. *Error Control Coding, Second Edition*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2004.

[62] Mikko H. Lipasti and John Paul Shen. "Exceeding the Dataflow Limit via Value Prediction". In: *Proceedings of the 29th Annual ACM/IEEE International Symposium on Microarchitecture*. MICRO 29. Washington, DC, USA: IEEE Computer Society, 1996.

[63] Mikko H. Lipasti, Christopher B. Wilkerson, and John Paul Shen. "Value Locality and Load Value Prediction". In: *Proceedings of the Seventh International Conference on Architectural Support for Programming Languages and Operating Systems*. ASPLOS VII. New York, NY, USA: ACM, 1996.

[64] F.Y. Liu, D. Patil, J. Lexau, P. Amberg, M. Dayringer, J. Gainsley, H.F. Moghadam, Xuezhe Zheng, J.E. Cunningham, AV. Krishnamoorthy, E. Alon, and R. Ho. "10-Gbps, 5.3-mW Optical Transmitter and Receiver Circuits in 40-nm CMOS". In: *Solid-State Circuits, IEEE Journal of* 47 (2012).

[65] Gabriel H. Loh. "3D-Stacked Memory Architectures for Multi-core Processors". In: *Proceedings of the 35th Annual International Symposium on Computer Architecture*. ISCA '08. Washington, DC, USA: IEEE Computer Society, 2008.

[66] Guillaume Maire, Laurent Vivien, Guillaume Sattler, Andrzej Kazmierczak, Benito Sanchez, Kristinn B. Gylfason, Amadeu Griol, Delphine Marris-Morini, Eric Cassan, Domenico Giannone, Hans Sohlström, and Daniel Hill. "High efficiency silicon nitride surface grating couplers". In: *Opt. Express* 16.1 (2008), pp. 328–333.

[67] G.Z. Masanovic, V. M N Passaro, and G.T. Reed. "Coupling to nanophotonic waveguides using a dual grating-assisted directional coupler". In: *Optoelectronics, IEE Proceedings* - 152.1 (2005).

[68] E. Mensink, D. Schinkel, E. Klumperink, E. van Tuijl, and B. Nauta. "A 0.28pJ/b 2Gb/s/ch Transceiver in 90nm CMOS for 10mm On-Chip interconnects". In: *Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International*. 2007.

[69] R. Morris, A.K. Kodi, and A. Louri. "Dynamic Reconfiguration of 3D Photonic Networks-on-Chip for Maximizing Performance and Improving Fault Tolerance". In: *Microarchitecture (MICRO), 2012 45th Annual IEEE/ACM International Symposium on*. 2012.

[70] Andreas Moshovos, Scott E. Breach, T. N. Vijaykumar, and Gurindar S. Sohi. "Dynamic Speculation and Synchronization of Data Dependences". In: *Proceedings of the 24th Annual International Symposium on Computer Architecture*. ISCA '97. New York, NY, USA: ACM, 1997.

[71] S.S. Mukherjee, P. Bannon, S. Lang, A. Spink, and D. Webb. “The Alpha 21364 network architecture”. In: *Micro, IEEE* (2002).

[72] B.T. Murphy. “Cost-size optima of monolithic integrated circuits”. In: *Proceedings of the IEEE* (1964).

[73] Christopher J. Nitta, Matthew K. Farrens, and Venkatesh Akella. “Resilient Microring Resonator Based Photonic Networks”. In: *Proceedings of the 44th Annual IEEE/ACM International Symposium on Microarchitecture*. MICRO-44. New York, NY, USA: ACM, 2011, pp. 95–104.

[74] Kunle Olukotun and Lance Hammond. “The Future of Microprocessors”. In: *Queue* 3.7 (Sept. 2005).

[75] Jin Ouyang and Yuan Xie. “Enabling quality-of-service in nanophotonic network-on-chip”. In: *Design Automation Conference (ASP-DAC), 2011 16th Asia and South Pacific*. 2011.

[76] Yan Pan, J. Kim, and G. Memik. “FlexiShare: Channel sharing for an energy-efficient nanophotonic crossbar”. In: *High Perf. Computer Architecture (HPCA), 2010*.

[77] Yan Pan, John Kim, and Gokhan Memik. “FeatherWeight: low-cost optical arbitration with QoS support”. In: *Proceedings of the 44th Annual IEEE/ACM International Symposium on Microarchitecture*. MICRO-44 ’11.

[78] Yan Pan, Prabhat Kumar, John Kim, Gokhan Memik, Yu Zhang, and Alok Choudhary. “Firefly: Illuminating future network-on-chip with nanophotonics”. In: *Proc. of the Int’l Symposium on Comp. Architecture (ISCA)*. 2009.

[79] Ashok M. Prabhu, Zhanghua Han, Alan Tsay, and Vien Van. “Wideband 1.5µm-Radius SOI Add-Drop Microring Filter for WDM on-Chip Interconnects”. In: *Conference on Lasers and Electro-Optics/International Quantum Electronics Conference*. Optical Society of America, 2009, CFV4.

[80] Haisheng Rong, Ying-Hao Kuo, Ansheng Liu, Mario Paniccia, and Oded Cohen. “High efficiency wavelength conversion of 10 Gb/s data in silicon waveguides”. In: *Opt. Express* 14.3 (2006), pp. 1182–1188.

[81] K. Sakuma, P.S. Andry, C.K. Tsang, S.L. Wright, B. Dang, C.S. Patel, B.C. Webb, J. Maria, E.J. Sprogis, S.K. Kang, R.J. Polastre, R. R. Horton, and J.U. Knickerbocker. “3D chip-stacking technology with through-silicon vias and low-volume lead-free interconnections”. In: *IBM Journal of Research and Development* 52.6 (2008).

[82] S. Scott, D. Abts, J. Kim, and W.J. Dally. “The BlackWidow High-Radix Clos Network”. In: *Computer Architecture, 2006. ISCA ’06. 33rd International Symposium on*. 2006.

[83] Steven L. Scott and et al. *The Cray T3E Network: Adaptive Routing in a High Performance 3D Torus*. 1996.

[84] Assaf Shacham, Keren Bergman, and Luca P. Carloni. “On the Design of a Photonic Network-on-Chip”. In: NOCS ’07. Washington, DC, USA: IEEE Computer Society, 2007.

[85] John Paul Shen and Mikko H. Lipasti. *Modern processor design : fundamentals of superscalar processors*. Boston: McGraw-Hill Higher Education, 2005.

[86] F. Silla, M.P. Malumbres, J. Duato, D. Dai, and D.K. Panda. “Impact of adaptivity on the behavior of networks of workstations under bursty traffic”. In: *Parallel Processing, 1998. Proceedings. 1998 International Conference on*. 1998, pp. 88–95.

[87] Arjun Singh. “Load-Balanced Routing in Interconnection Networks”. PhD thesis. Stanford University, 2005.

[88] C.H. Stapper. “On Murphy’s yield integral [IC manufacture]”. In: *Semiconductor Manufacturing, IEEE Transactions on* (1991).

[89] Chen Sun, Chia-Hsin Owen Chen, George Kurian, Lan Wei, Jason Miller, Anant Agarwal, Li-Shiuan Peh, and Vladimir Stojanovic. “DSENT - A Tool Connecting Emerging Photonics with Electronics for Opto-Electronic Networks-on-Chip Modeling”. In: *Proceedings of the 2012 IEEE/ACM Sixth International Symposium on Networks-on-Chip*. NOCS ’12.

[90] Fei Sun, Jinzhong Yu, and Shaowu Chen. “A 2x2 optical switch based on plasma dispersion effect in silicon-on-insulator”. In: *Optics Communications* 262.2 (2006).

[91] C.-E.W. Sundberg. “Erasure and Error Decoding for Semiconductor Memories”. In: *Computers, IEEE Transactions on* (1978).

[92] James E. Thornton. “Parallel Operation in the Control Data 6600”. In: *Proceedings of the October 27-29, 1964, Fall Joint Computer Conference, Part II: Very High Speed Computer Systems*. AFIPS ’64 (Fall, part II). New York, NY, USA: ACM, 1965.

[93] J. Van Olmen, A Mercha, G. Katti, C. Huyghebaert, J. Van Aelst, E. Seppala, Zhao Chao, S. Armini, J. Vaes, R.C. Teixeira, M. Van Cauwenberghe, P. Verdonck, K. Verhemeldonck, A Jourdain, W. Ruythooren, M. De Potter de ten Broeck, A Opdebeeck, T. Chiarella, B. Parvais, I Debusschere, T. Y Hoffmann, B. De Wachter, W. Dehaene, M. Stucchi, M. Rakowski, P. Soussan, R. Cartuyvels, E. Beyne, S. Biesemans, and B. Swinnen. “3D stacked IC demonstration using a through Silicon Via First approach”. In: *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*. 2008.

[94] S. Vangal, J. Howard, G. Ruhl, S. Dighe, H. Wilson, J. Tschanz, D. F̃inan, P. Iyer, A. Singh, T. Jacob, S. J̃ain, S. Venkataraman, Y. Hoskote, and N. Borkar. “An 80-Tile 1.28TFLOPS Network-on-Chip in 65nm CMOS”. In: *Solid-State Circuits Conference, 2007. ISSCC 2007.* 2007.

[95] Dana Vantrease, Robert Schreiber, Matteo Monchiero, Moray McLaren, Norman P. Jouppi, Marco Fiorentino, Al Davis, Nathan Binkert, Raymond G. Beausoleil, and Jung Ho Ahn. “Corona: System Implications of Emerging Nanophotonic Technology”. In: *Proceedings of the 35th Annual International Symposium on Computer Architecture*. ISCA ’08.

[96] Dana Vantrease, Nathan Binkert, Robert Schreiber, and Mikko H. Lipasti. "Light speed arbitration and flow control for nanophotonic interconnects". In: *Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture*. MICRO 42. 2009.

[97] L. Vivien, G. Maire, G. Sattler, D. Marris-Morini, E. Cassan, S. Laval, A Kazmierczak, D. Giannone, B. Sanchez, A Griol, D. Hill, K. B. Gylfason, and H. Sohlstrom. "A high efficiency silicon nitride grating coupler". In: *Group IV Photonics, 2007 4th IEEE International Conference on*. 2007.

[98] Steven J. E. Wilton and Norman P. Jouppi. "CACTI: An Enhanced Cache Access and Cycle Time Model". In: *IEEE Journal of Solid-State Circuits* 31 (1996), pp. 677–688.

[99] Felix Wolf. "Scalasca". In: *Encyclopedia of Parallel Computing*. Ed. by David Padua. Springer, 2011.

[100] Yi Xu, Jun Yang, and Rami Melhem. "Channel borrowing: an energy-efficient nanophotonic crossbar architecture with light-weight arbitration". In: *Proc. of the International Conference on Supercomputing*. ICS '12. New York, NY, USA: ACM, 2012.

[101] Yi Xu, Jun Yang, and Rami Melhem. "Tolerating Process Variations in Nanophotonic On-chip Networks". In: *Proceedings of the 39th Annual International Symposium on Computer Architecture*. ISCA '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 142–152.

[102] Jin Yao, Xuezhe Zheng, Guoliang Li, I. Shubin, H. Thacker, Ying Luo, K. Raj, J.E. Cunningham, and A.V. Krishnamoorthy. "Grating-coupler based low-loss optical interlayer coupling". In: *Group IV Photonics (GFP)*. 2011.

[103] Jin Yao, Xuezhe Zheng, Guoliang Li, Ivan Shubin, Ying Luo, Hiren Thacker, Attila Mekis, Thierry Pinguet, Subal Sahni, Kannan Raj, John E. Cunningham, and Ashok V. Krishnamoorthy. *Grating-coupler-based optical proximity coupling for scalable computing systems*. 2011.

[104] Reza Zamani and Ahmad Afsahi. "Communication Characteristics of Message-Passing Scientific and Engineering Applications". In: *IASTED PDCS'05*.

[105] Xuezhe Zheng, Ivan Shubin, Guoliang Li, Thierry Pinguet, Ãttila Mekis, Jin Yao, Hiren Thacker, Ying Luo, Joey Costa, Ǩannan Raj, John E. Cunningham, and Ashok V. Krishnamoorthy. "A tunable 1x4 silicon CMOS photonic wavelength multiplexer/demultiplexer for dense optical interconnects". In: *Opt. Express* (2010).

[106] Xuezhe Zheng, P. Koka, H. Schwetman, J. Lexau, R. Ho, J.E. Cunningham, and AV. Krishnamoorthy. "Silicon photonic WDM point-to-point network for multi-chip processor interconnects". In: *Group IV Photonics, 2008 5th IEEE International Conference on*. 2008.

[107] Xuezhe Zheng, Jon Lexau, Ying Luo, Hiren Thacker, Ţhierry Pinguet, Attila Mekis, Guoliang Li, Jing Shi, Ƥhilip Amberg, Nathaniel Pinckney, Kannan Raj, Ron Ho, John E. Cunningham, and Ashok V. Krishnamoorthy. "Ultra-low-energy all-CMOS modulator integrated with driver". In: *Opt. Express* (2010).

[108] A. J. Zilkie, P. Seddighian, B. J. Bijlani, W. Qian, D. C. Lee, S̃. Fathololoumi, J. Fong, R. Shafiiha, D. Feng, B. J. Luff, X. Zheng, J̃. E. Cunningham, A. V. Krishnamoorthy, and M. Asghari. "Power-efficient III-V/Silicon external cavity DBR lasers". In: *Opt. Express* (2012).

[109] Arslan Zulfiqar, Pranay Koka, Herb Schwetman, Mikko Lipasti, Xuezhe Zheng, and Ashok V. Krishnamoorthy. "Wavelength Stealing: An Opportunistic Approach to Channel Sharing in Multi-chip Photonic Interconnects". In: *Proceedings of the 46th Annual IEEE/ACM International Symposium on Microarchitecture*. MICRO-46 '13.