Design and Analysis of Large Scale Nanophotonic On-Chip Networks

By

CHRISTOPHER NITTA
B.S. (University of California at Davis) 2000
M.S. (University of California at Davis) 2004

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Computer Science

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Professor Matthew Farrens, Chair

Professor Venkatesh Akella

Professor Rajeevan Amirtharajah

Committee in Charge
2011

I dedicate this dissertation to my loving wife and best friend Amy.

Acknowledgments

I would like to express my sincere gratitude to my committee chair Professor Matthew Farrens for advising me, encouraging me, thoroughly editing my work, and providing me with my graduate and undergraduate foundations in computer architecture. My sincere thanks also goes to my co-adviser Professor Venkatesh Akella for his high expectations of my work, without him spurring me on I would not be the quality of researcher that I am today. I would also like to thank Professor Rajeevan Amirtharajah for his insights into thermal dynamics, and for providing a fresh perspective on this work.

It is a great pleasure to thank Kevin Macdonald for his multiple month long simulation runs that were essential to this work.

I owe my gratitude to Dr. Mark Duvall for encouraging me to apply to graduate school and eventually to pursue my PhD. I am grateful for Professor Andrew Frank who built the UC Davis HEV Center, its unparalleled creative environment provided me with the broad knowledge base that I continually relied upon during my doctoral work. I would like to thank all of the alumni of Team Fate who I had the pleasure to worked with over my six years at the HEV Center.

I must give my deepest thanks to my friends and company partners, William Allan for his boundless creativity and endless optimism, and Charnjiv "Chief" Bangar for meeting any of our hardware needs and for taking care of even the most minute details. Their efforts to continue our company despite both having full time jobs, allowed me to tackle interesting problems that prevented me from burning out on my doctoral research.

I owe my deepest gratitude to my family for all their love and support. I would especially like to thank my mother Gwen for providing a loving, stable environment that fostered my early intellectual curiosity, and to my late father John for always expecting I do my very best in all my endeavors; I know he would have been proud to have another doctor in the family.

Last but not least, I cannot thank my wife Amy Nitta enough for her unwavering support, without her this may never have been a completed work.

Contents

List of Figures

List of Tables

List of Acronyms

HP Hewlett-Packard . . . 51
IBM International Business Machines . . . 3
ITRS International Technology Roadmap for Semiconductor . . . 14
K Kelvin
KB Kilobyte(s)
LFSR Linear Feedback Shift Register . . . 42
LPCVD Low-Pressure Chemical Vapor Deposition . . . 124
LU Lower and Upper matrix decomposition . . . 65
LVDS Low Voltage Differential Signaling . . . 42
MASTAR Model for Assessment of cmoS Technologies And Roadmaps . . . 14
MBDS Multi-Bit Differential Signaling . . . 42
MIT Massachusetts Institute of Technology . . . 2
mm Milimeter(s)
MTBF Mean Time Between Failure . . . 5
mW Miliwatt(s)
MWSR Multiple Writer Single Reader . . . 73
NAK Negative AcKnowledge . . . 47
NcK N choose K . . . 43
NED Negative Exponential Distribution . . . 64
nm Nanometer(s)
Ω Ohm(s)
ORION Open Research Infrastructure for Optimizing Networks . . . 14
PDG Packet Dependency Graph . . . 64
PECVD Plasma Enhanced Chemical Vapor Deposition . . . 123
PMMA Polymethyl Methacrylate . . . 26
pJ Picojoule(s)
pm Picometer(s)
RAID Redundant Array of Independent Disks
RAW raw Architecture Workstation . . . 2
RIE Reactive Ion Etching . . . 123
s Second(s)
SAW Stop-And-Wait . . . 46
SECDED Single Error Correction and Double Error Detection . . . 42
SERDES SERializer/DESerializer . . . 97
SiO_2 Silicon Dioxide . . . 123
SOI Silicon-On-Insulator . . . 16
SPLASH-2 Stanford ParalleL Applications for SHared-memory 2 . . . 64
SRW Sliding Ring Window . . . 24
SWMR Single Writer Multiple Reader . . . 37
TB Terabyte(s)
TCW Temperature Control Window . . . 22
TED Triple Error Detection . . . 42
TFLOPS Tera FLoating-point OPerations per Second . . . 2
THz Terahertz
TIA Transimpedance Amplifiers . . . 95
μA Microamp(s)
μm Micrometer(s)
μW Microwatt(s)

UV Ultraviolet 12
V Volt(s)
W Watt(s)
WDM Wavelength Division Multiplexing 3
XOR eXclusive OR

Abstract

In recent years both the core clock frequencies and the amount of power that can cost-effectively be removed from a chip have begun to level off, while chip manufacturers have continued to increase the number of cores per processor. These facts have driven computer architects to examine high bandwidth, energy efficient on-chip optical networks, which feature microring resonators as a basic building block. In this thesis I investigate ways to compensate for the thermal sensitivity of these resonators, and I present an architect's view of the nature of microring malfunctions and propose techniques to overcome errors. I demonstrate that on-chip networks with microring counts in the hundreds of thousands will be feasible, and that using photonics, it is possible to create directly connected topologies that eliminate the need for arbitration. The Directly-Connected Arbitration-Free (DCAF) topologies that I propose in this work are actually a family of networks that allow the computer architect to configure the degree of simultaneous communication in order to meet the available power budget. Finally, I show that because photonics do not follow Moore's Law, it will be increasingly difficult to get data to and from the microrings, and that designers must take this into account when choosing a topology. Based on the work presented in this thesis, it is clear that architects must take a holistic view of the entire system when designing photonic networks.

Chapter 1

Introduction

In 1965, Gordon Moore [62] predicted that the number of transistors that can cost effectively be placed on an integrated circuit would double every two years[1]. This became known as Moore's Law, and for the first few decades after it was articulated microprocessors experienced huge improvements in performance as the exponentially growing number of transistors available on-chip were used to implement techniques originally developed for use in supercomputers. Recently clock frequencies of individual cores have leveled off because issues such as overall power consumption and heat dissipation have become of greater concern. However, the number of transistors continues to double, leaving chip designers in an interesting position. What should be done with all these transistors? Continuing to place more and more of the memory hierarchy on-chip in order to address the processor/memory speed mismatch has reached a point of diminishing returns, so currently companies are using the extra chip real estate to increase the number of cores per processor – for example most commercially available desktop processors are now dual or quad core and processors already exist with dozens of cores (Intel 80-core Polaris [90] and Tileras 64-core chip). Chip manufacturers are planning on hundreds [9] and even thousands [17] of cores in the future.

Unlike the clock frequency increases and Central Processing Unit (CPU) improvements of the past, it is unclear if the increased number of cores will result in the same level of performance improvements that have been observed over the past decades. Amdahl's Law [2] states that the speedup due to the improvement of a feature is limited by the frac-

[1] Moore's Law is commonly quoted as every 18 months

tion of the time the feature is actually used, so increasing the number of cores will be of limited use if the cores sit idle. And if the cores are not idle, then providing sufficient data bandwidth may be a significant challenge. Bell, Szalay, and Gray [7] convincingly show the need for balanced computing systems, and according to the authors of [99], it is expected that chips with a thousand cores will be able to perform in excess of 10 Tera FLoating-point OPerations per Second (TFLOPS). A balanced computing system would need a 10TB/s memory bandwidth in order to support such a chip.

Before one can suggest techniques to provide the desired memory bandwidth, the memory system structure must first be considered. The address space in multicore processors is either shared or distributed. In a shared memory system, a cache coherency protocol is required to maintain consistency. Bus snooping is one of the most common mechanisms used to support cache coherency protocols [33]; unfortunately, bus based systems do not scale well. Therefore, other mechanisms to connect processors (such as on-chip networks) must be used in the design of large high-performance cache coherence systems. Distributed memory systems also require an interconnection network, and they scale better - however, they present a much more complex programming model. Either way, the memory system of future multicore processors is going to require an on-chip interconnection network.

Future multicore systems will also require high bandwidth communication networks, and electrical networks are not likely to scale up well, primarily for latency and power consumption reasons [90]. According to Miller [60] the rate at which heat can cost-effectively be removed from a chip will plateau around 200W; furthermore, the amount of power consumed in the interconnection network is becoming a major portion of the total power budget. For example, the 16-tile Massachusetts Institute of Technology (MIT) raw Architecture Workstation (RAW) and Intel Polaris interconnection networks consume 36% and 28% of total chip power, respectively [35]. This problem will only become worse as the number of cores on a chip continues to climb, since the size of the on-chip network will need to grow in order to accommodate additional cores.

Recent advances in silicon nanophotonics have led researchers to explore the feasibility of using optics in future multicore designs. In particular, the ability to fabricate microring resonators and optical waveguides on silicon [85] has enabled the use of optics for

both on and off-chip communication [52, 53, 61, 60]. Using an external laser and a comb generator, researchers are now able to create dozens of wavelengths and steer them around on-chip using these resonators, creating optical interconnection networks [92, 6, 97].

Optical interconnects promise to provide more energy efficient communication than electrical interconnects. For example, the power required to send a 1Gbps electrical signal from one corner to an opposing corner of a 22mm x 22mm die is 1.25mW (assuming 16nm technology with 231fF/mm wire capacitance [34], at 0.7V), while the equivalent optical connection would require 117μW (assuming a 13.5dB attenuation – 3dB photo conversion, 4.5dB interconnect, 1dB coupling, and 5dB laser and a 10fF photodetector capacitance). In addition, the optical interconnect power is the power required at the wall plug and includes all components of link power, while the electrical interconnect calculation presented is only that to switch the wire (repeaters, drivers, buffers, etc. are not included). A minimum of 22 repeaters would be required for the electrical interconnect since a global wire in 16nm technology can carry a 1GHz signal a maximum $\sim$2mm (according to the methodology used in [64]), meaning that such an electrical link would either have significant additional latency or would consume considerably more power.

Energy efficient communication is not the only advantage of optical interconnects – the unique properties of optics can also be exploited in a way that allows topologies which are impractical (or impossible) to build using only electronics. Fully-connected topologies, for example, are extremely desirable because they ease the burden of parallel programming and improve performance. In fact, the network being considered for the International Business Machines (IBM) Blue Waters (the most ambitious machine to date with 300000+ nodes) is direct and fully connected to the extent possible [3]. Fully-connected electrical networks are rarely feasible because of their extraordinary wiring complexity - however, the ability of optics to provide multiple wavelengths within the same waveguide (accomplished using Wavelength Division Multiplexing (WDM)) means a single waveguide per node connection is all that is required. The use of WDM, coupled with the fact that waveguides can intersect without complete signal interference, makes it possible to create a Fully-Connected Optical Network (FCON).

1.1 Thesis Contributions

The goal of this work is to investigate the design and implementation of large scale microring based on-chip nanophotonic networks – the practical challenges that have not been addressed by other researchers (such as trimming, resilience, and electro-optical interfaces), as well as the potential topologies that can be built by exploiting the unique capabilities of photonics.

This dissertation grew out of an initial investigation into the thermal stability of large scale microring based on-chip networks, and resulted in the following contributions (presented in order of appearance in this work, not in order of importance):

- **Investigation of System-Level Trimming in On-Chip Networks** Microring resonators [53] are designed to respond to particular (unique) wavelengths, and are also known to be extremely sensitive to thermal variations. Temperature fluctuations cause the rings to shift their resonance wavelengths [59, 12, 31] – a change in temperature of as little as one degree can cause a microring to respond to a completely different wavelength than intended. Active *trimming* techniques have been proposed in [1] that dramatically increase the network power requirements, and it was unclear if these techniques could truly be implemented systemwide, across hundreds of thousands (or even millions) of rings.

  In order to perform the investigation I developed a power and floor-plan simulation library, which was integrated with an existing thermal simulator to provide a full power/floor-plan/thermal simulator. The thermal simulations showed that microring based on-chip networks are thermally stable (when trimming is not included). Once the thermal stability of the microring network had been established, active trimming was included in the simulations, and I discovered that the heating power required for trimming has a non-linear relationship with microring count, and that using current injection for trimming can quickly lead to thermal runaway. These observations led me to the proposal and analysis of a technique called the Sliding Ring Window, which aims to increase the thermal window in which the network must be maintained. The results of this work are published in [66].

- **Development of Microring Based Fault Model** The trimming work dealt with thermal drift errors, but did not address problems due to fabrications errors. Including fabrication variations in the simulation models was a natural extension of the trimming investigation. During this extended investigation we decided that instead of just expanding the models, the focus should be shifted toward determining modes of microring failure. It was this investigation that led to the development of the first microring based fault model. The fault model was used when analyzing a range of techniques to increase the reliability of optical links. The results from the link reliability study were then used to determine the fault rate microring resonators must attain in order to meet a target Mean Time Between Failure (MTBF), or the encoding scheme that should be used to meet the MTBF target if the microring fault rate is known. The results show that unless microring fabrication techniques dramatically improve, an error detecting or correcting code will almost certainly be necessary in order for large scale microring based on-chip networks to be realized. These results will appear in [67].

- **Proposal and Evaluation of a Family of Directly Connected Arbitration Free Networks** The proposal and evaluation of the family of directly connected arbitration free networks is perhaps the most important contribution of this thesis.

  The trimming and reliability studies demonstrated that the issues of fabrication tolerances and thermal drift could be overcome, and that large scale microring based on-chip networks are feasible. Therefore, I decided to examine the potential benefits and costs of a FCON. The investigation showed that the performance gains that a FCON can realize are significant when compared to even the highest performing photonic crossbar that has been previously proposed. It was determined that while the FCON outperformed the crossbar, the large amount of photonic power required to support all possible links simultaneously is not feasible.

  The observation that much of the photonic power in a FCON was not being used prompted the proposal of a directly connected network that is limited in the amount of simultaneous communication it can perform. Additional microring resonators are

used to accomplish this, and provide a family of networks that are almost identical to a FCON, but use a more reasonable amount of photonic power. In addition, these networks do not require arbitration, only flow control – this is a significant advantage, because arbitration is an overhead that must be paid for all communication, while flow control only occurs when the network is becoming overwhelmed. This work also showed that the energy efficiency of networks with an external laser is highly dependent upon workload, and that solutions to make the energy efficiency independent of workload should be further investigated.

- **Investigation of WDM Trade-offs** During the detailed power analysis of the network topologies, it became clear that the portion of power attributed to the wiring of the microring resonators is significant, because the length of a group of microring resonators used for 64-bit WDM is approximately a fifth the width of a network node. Therefore, I examined the trade-offs between switching speed and data path width, in order to meet a target link bandwidth. This investigation showed that a 64-bit link is not always the most energy efficient, even though many researchers have assumed the use of 64 wavelengths to create 64-bit data paths modulated at 10GHz in their proposed photonic networks [92, 40, 15, 71, 70]. My investigation showed that different data path widths should be considered for on-chip nanophotonic networks, and that other photonic technologies merit further study since microring resonators do not scale according to Moore's Law.

1.2 Dissertation Structure

The rest of this dissertation is structured as follows: Chapter 2 presents a fairly detailed introduction into the basic building blocks of nanophotonic interconnects, while Chapter 3 discusses the feasibility of on-chip optical networks with large microring counts. The results of the performance and power investigation of a FCON is presented in Chapter 4. Chapter 5 discusses a family of novel directly connected networks, and presents the results of performance and power simulations. An investigation into the trade-offs of WDM and the ramifications on the electrical requirements is presented in Chapter 6. This dissertation

concludes in Chapter 7 with observations and suggestions of future work.

Chapter 2

Background

Before the feasibility of optical communication can be addressed, the construction of an optical link must first be understood. Figure 2.1 presents a typical on-chip optical link that uses an external laser as a light source. The necessary set of wavelengths used for communication can be created by either an external laser with a comb filter, or by the use of a mode locked laser [44]. These wavelengths are delivered to the transmitter section of the source node via an optical waveguide.

The transmitter, consisting of electrical drivers and optical modulators, uses the modulators to remove certain wavelengths (in this case λ_2 and λ_n), creating the desired pattern. This pattern then travels down the waveguide from the source to the destination node. When the transmitted value arrives at the destination, the optical detectors convert the photonic power back to an electrical signal and the transmission is complete. The rest of this chapter provides more details of how the individual components of the optical link

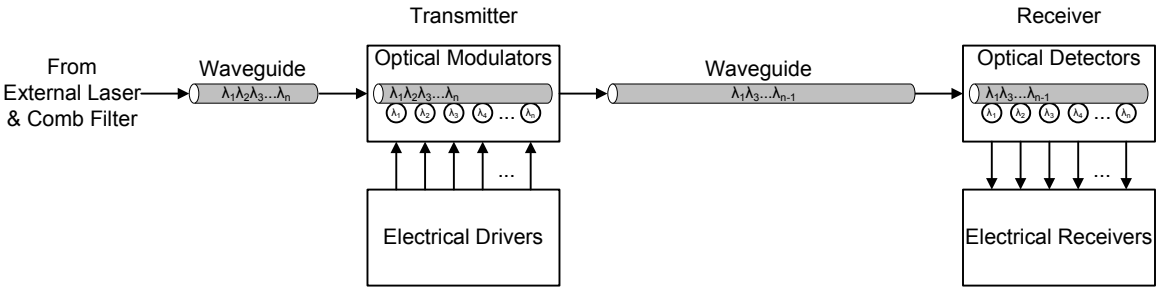
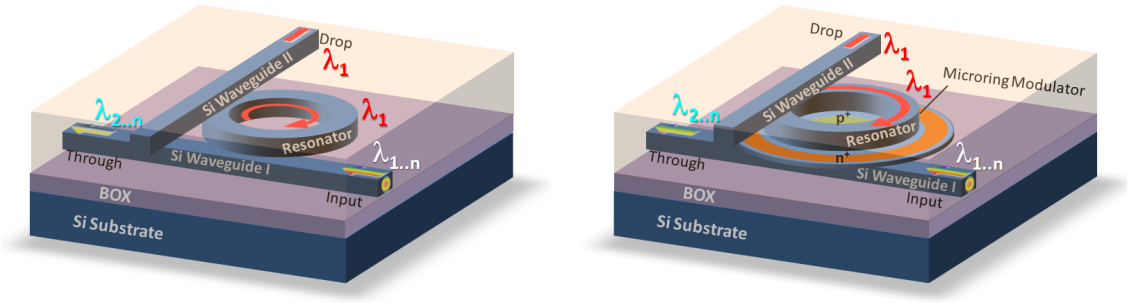


Figure 2.1: Example of an Optical Link



(a) Filter (PASSIVE) Microring (b) Modulator (ACTIVE) Microring

Figure 2.2: Microring Resonators

function.

2.1 Ring Resonators

Microring resonators are designed to resonate when presented with specific individual wavelengths and remain quiescent at all other times. The resonance wavelengths to which microring resonators respond repeat at an interval known as the Free Spectral Range (FSR) – for example, a microring designed to resonate at λ_1 will also resonate at $\lambda_1 \pm$ FSR, $\lambda_1 \pm$ 2FSR, etc. The ability to respond to specific wavelengths enables the removal (filtering) of specific wavelengths from a waveguide, and these resonators are the primary technology used to bundle the high quantity of wavelengths per waveguide needed for Dense Wavelength Division Multiplexing (DWDM). This filtering can be achieved using either passive or active microrings. Figure 2.2(a) shows a passive microring that is biased during fabrication to extract only λ_1 from the incoming waveguide and steer it down a perpendicular waveguide.

Since the passive microrings are biased during fabrication to always respond to a single wavelength, they cannot be used for modulation. Modulating a given wavelength requires an active microring resonator, which is designed to change its resonance frequency based on the amount of current present in the n^+ base. Figure 2.2(b) illustrates an active microring resonator modulating wavelength λ_1. If the electrical current is present, λ_1 is extracted and sent down waveguide II – if there is no current applied, λ_1 will continue down waveguide I unaffected.

Generally, it is assumed that the presence of a wavelength represents a logic 1 and the absence represents a logic 0, and the method by which an active microring modulates depends upon the configuration of the incoming and outgoing waveguides. For example, if the incoming waveguide is also the outgoing waveguide, then a zero can be created by using the microring to remove the wavelength by bending it onto a dead end drop waveguide, and a one is created by allowing the wavelength to pass unaffected (this is shown in Figure 2.2(b)). If the incoming and outgoing waveguides are not the same, then ones are created by bending the wavelength onto the outgoing waveguide, and zeros by allowing the wavelength to continue unperturbed along the incoming waveguide. (This is shown in Figure 2.2(b) if waveguide II is the outgoing waveguide, and not a dead-end drop.)

2.2 Photonic Vias and Vertical Coupling

Waveguides can intersect without complete signal interference, unlike wires carrying electronic signals. Intersections of waveguides at 90 degrees allow signals traveling down each waveguide to continue on intact, although each signal will suffer a small attenuation (often modeled as ∼0.1dB), possible signal crosstalk[1], and partial signal reflection. This characteristic of photonics has allowed on-chip optical networks to be laid out on a single layer – however, the cumulative effect of a large number of intersections may make a single layer waveguide layout infeasible. Therefore, waveguides may need to be routed on different layers to avoid excessive intersections.

In the electronic domain signals can easily move from layer to layer using vias, and transitioning photonic signals to different layers is done in a similar manner. Grating couplers are used to couple optical fibers and waveguides [86, 57], and vertical grating couplers can be used to connect waveguides on different layers. In this work it is assumed that the signal attenuation of such a coupling is 1dB(a conservative estimate considering optical fiber and waveguide couplings of less than 1dB loss have already been demonstrated). Figure 2.3 illustrates a grating coupler being used as a photonic via.

Grating couplers are not the only possible structure for use as a photonic via.

[1]In electronics, crosstalk is any phenomenon by which a signal transmitted on one circuit or channel of a transmission system creates an undesired effect in another circuit or channel.

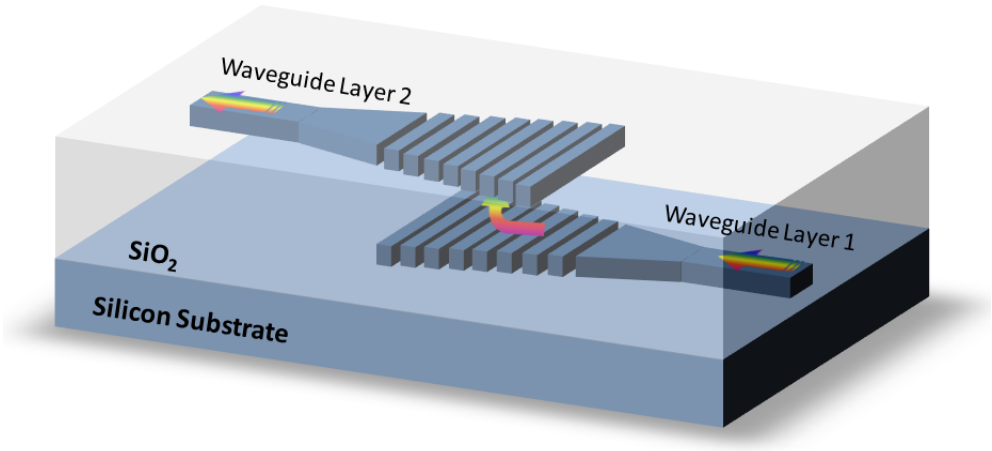


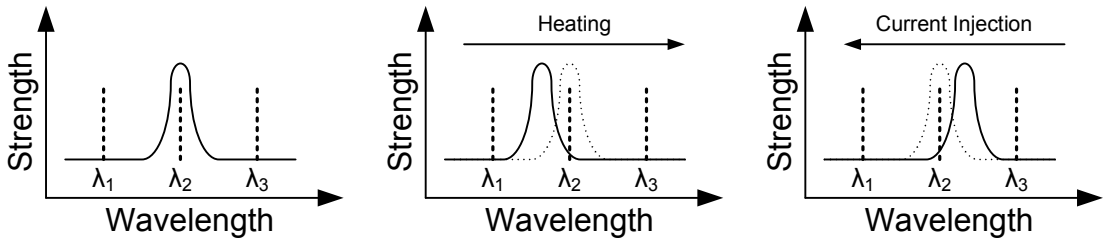
Figure 2.3: Grating Coupler (Photonic Via)

Plasmonics have the capability to drastically change the direction of light, which could be useful when changing layers; however, plasmonics suffer from high path attenuation (typically ~0.2dB/μm [19]). Over the relatively short distances required for an inter-layer via (assumed less than 10μm), the loss experienced by a plasmonic based photonic via may be acceptable; the possibility of using plasmonics as a photonic via is not investigated in this work, but only mentioned as an example of a possible alternative to grating couplers.

The microring resonators described previously are also capable of vertical coupling. The microrings in Figure 2 are vertically coupled, allowing perpendicular waveguides to be coupled to a single microring without the waveguides intersecting. See Appendix A for more details on the fabrication processes involved in creating these structures.

2.3 Trimming

The wavelengths individual microrings respond to are set during fabrication - however, variations in fabrication tolerances may require that certain microrings be "trimmed" to move the resonance frequency up or down slightly. In addition, the refractive index (n) of silicon changes due to changes in ambient temperature (ΔT), which can be modeled as -$\Delta n \approx 1.84 \times 10^{-6} \times \Delta T$. As a result microring resonators are very sensitive to temperature and drift spectrally approximately 0.09nm/°C. Trimming can be used to dynamically modify the resonance frequency of a microring to overcome both thermal drift



(a) On Resonance λ_2

(b) Red Shift Needed, Resonance Wavelength Moved Using Heating

(c) Blue Shift Needed, Resonance Wavelength Moved Using Current Injection

Figure 2.4: Microring Resonance vs. Wavelength – (a) Shows On Resonance λ_2, (b) Shows that Red Shift is Needed (Solid Line) and Heating is Applied (Dotted Line), and (c) Shows Blue Shift is Needed (Solid Line) and Current Injection is Used (Dotted Line)

and fabrication inaccuracies. This trimming can be accomplished dynamically by increasing the current in the n^+ region (to shift the resonance towards the blue, referred to as current injection) or by heating the ring [2](to shift towards the red) [1]. Figure 2.4 illustrates how resonance frequency changes when heating (Figure 2.4(b)) and when injecting current (Figure 2.4(c)). Unfortunately, these active trimming techniques can result in a dramatic increase in the overall power requirements. Passive or post fabrication techniques such as using Ultraviolet (UV) light to correct for fabrication errors have been proposed [80, 48], and they have distinct advantages - for example, once all the rings are trimmed power is only needed to address thermal drift. However, this approach requires each ring to be analyzed/trimmed individually, so it is not clear how practical this will be at the system level.

[2]This effect can be implemented in devices, for example, with thin-film platinum surface heaters near waveguide sections [23]

Chapter 3

Feasibility of Large Nanophotonic On-Chip Networks

This chapter explores two of the principal challenges to building large scale nanophotonic networks – thermal sensitivity and reliability. These issues have not been addressed thus far by the research community because the initial focus has been on showing the potential benefits of optical interconnects compared to electrical interconnects. Now that the benefits have been established [60], the practical issues underlying the implementation of large scale photonic networks need to be addressed. Given the extreme thermal sensitivity of microring resonators, for example, it is not clear if microring based on-chip networks can be made thermally stable[1]. Performing such an analysis requires knowing both the floor plan and the amount of power dissipated in each floor plan unit; furthermore, since some of the components of power dissipation (such as static leakage and trimming) are a function of temperature, the power and thermal models must be integrated.

In addition to thermal sensitivity, fabrication inaccuracies further complicate and amplify the importance of addressing reliability issues from the outset. Creating an integrated solution to deal with these issues is of great importance if large scale microring based on-chip networks are to be realized. In this chapter I will present my work on these two problems; Section 3.1 presents a description of the integrated power/floor-plan/thermal

[1] In this work I consider that the network is thermally stable if the microrings do not experience thermal runaway and if they drift less than 1nm.

simulation library used in the feasibility study, while Section 3.2 contains an investigation of the system level trimming issues present in large scale microring based networks. A Mean Time Between Failure (MTBF) analysis based on a derived nanophotonic fault model is presented in Section 3.3, Section 3.4 describes the related work to this feasibility study, and this chapter concludes in Section 3.5 with a brief discussion.

3.1 Mintaka Simulation Library

At the onset of this work no publicly available on-chip optical network power simulator existed; therefore, I developed the Mintaka[2] Simulation Library in order to evaluate the power and thermal characteristics of on-chip optical networks. The motivation to create a new power simulator instead of modifying an existing simulator such as ORION 1.0 [93] was driven by the fact that ORION 1.0 was grossly inaccurate for deep sub-micron technologies [41].

The photonic power estimates in Mintaka were developed using a link loss approach similar to that done in [1], with relevant data gathered from the literature and extrapolated from laboratory test results (see Table 3.1). Hot-Spot 5.0 [37, 36] was chosen to perform the thermal analysis since it can be compiled as a library, allowing it to be integrated into Mintaka. In a similar fashion to Hot-Spot, Mintaka has been designed to be compiled and linked as a library to facilitate future integration into other network performance simulators.

Many of the electrical components used in Mintaka were modeled in a manner similar to that discussed in [18] and used in ORION 1.0, although electrical technology data such as transistor capacitances were taken from Cache Access and Cycle Time Information (CACTI) 6.5 [89], and Model for Assessment of cmoS Technologies And Roadmaps (MASTAR) using the International Technology Roadmap for Semiconductor (ITRS) 2009 [81] parameters. CACTI was used for for technology parameters from 90nm to 32nm, and MASTAR was used for technology parameters below 32nm. Unlike ORION and CACTI, Mintaka does not size transistors by directly scaling from the 0.8μm technol-

[2]Mintaka -faintest of the three belt stars in ORION, which hints at the faint influence of the Open Research Infrastructure for Optimizing Networks (ORION) simulator.

Table 3.1: Simulation Optical Parameters

Description	Value	Description	Value
Waveguide		**Microring**	
Width	0.3μm	Diameter	3μm
Spacing	1μm	Spacing	5μm
Minimum Bend Radius	1.5μm	Resistance	10Ω
Attenuation*	0.3dB/cm	Capacitance	10fF
Intersection Attenuation	0.1dB	Quiescent Current*	10μA
Grating Attenuation	1dB	On Resonance Attenuation	0.5dB
Bend Attenuation	2.25e-3dB	Off Resonance Attenuation*	1.5e-3dB
Photodetector			
Width	3μm	Attenuation*	3dB
Height	0.3μm	Capacitance*	10fF

* The numbers are taken from the Corona published works [1].

ogy point[3] – rather, it sizes transistors based on the required switching period and load to be driven. Once the overall transistor width is determined from the load and switching period, the amount of transistor folding that is necessary is calculated. The wire technology data is based on work done by Ho in [34], and the methodology described in [64] is used in Mintaka to determine the correct wire sizing (local, semi-global, or global) given the desired bandwidth and wire length. The energy required per transition for each sub-component is calculated, and the number of transitions per sub-component is maintained as an integer in order to mitigate potential floating point round off errors that can occur during long simulation runs if a floating point only implementation is used. Static power loss is accounted for in Mintaka, since CACTI 6.5 calculates static power values as a function of temperature; the temperature of each floor-plan unit is passed into the simulated network in order to ascertain the static power loss.

The link loss calculation starts at a photodetector and works backwards towards the source, adding the attenuation losses along the way; this is done for the worst case path for all links in the modeled network. Attenuation sources include the photodetector, waveguide, waveguide intersection, waveguide bends, grating coupling, on-resonance rings,

[3]ORION and CACTI size transistors based upon constant factors of the feature size. These constant factors originate from the actual sizes used in 0.8μm technology. This approach has become less accurate as technology has moved into the deep sub-micron.

and off-resonance rings. Once the total attenuation has been calculated for the worst case path the power output necessary to switch the photodetector at the desired rate is determined, and given these two values the required laser power per wavelength is a simple calculation of $P_{PD}10^{\frac{A}{10}}$ (where P_{PD} is the required power at the photodetector and A is the attenuation of the path.)

Once the minimum laser power has been calculated for the worst-case path, multiplying by the total number of wavelengths provides the total amount of laser power required by the network. Once this value has been determined, the energy lost (absorbed) by each optical component can then be calculated, starting at the source and working downstream towards the photodetector.

In Mintaka, the floor-plan layout for each network is integrated into the electrical/optical power/sizing calculations. The floor-plan layout is necessary since both the optical and electrical power requirements depend upon the distance which the signals must travel, and the minimum size of some sub-components is dependent upon the power requirements.

The power consumed in the network, both electrical and optical, is maintained for each floor-plan component. This floor-plan, floor-plan power, and floor-plan temperature data is passed to the Hot-Spot library (using appropriate thermal constants for Silicon-On-Insulator (SOI)) to calculate the new floor-plan temperatures. The Hot-Spot steady state solver is used to determine the updated temperatures for the floor-plan components, and these values are then used by Mintaka to calculate more accurate power consumption numbers for the next iteration, since some components (such as static power loss and trimming power) are a function of temperature. The iterative process continues until either Mintaka/Hot-Spot converges on a steady state solution, or a thermal runaway is detected.

Mintaka was validated by comparing its link loss calculations to those published for Corona [1], given the same parameters. Mintaka calculated an attenuation loss of 13dB for Corona, which matched the published values – a 3dB photodetector loss, plus the 11dB transmission loss, minus the 1dB coupling loss (the coupling loss is not included since Mintaka calculates the required on-chip laser power only). The validation of electrical components was done by comparing their values to the intermediate values generated inside

CACTI and ORION when using the same technology data. The differences observed were primarily due to the dynamic transistor sizing done by Mintaka vs. the static scaling from 0.8μm assumed in ORION and CACTI. Mintaka was also validated by hand-calculating the link losses of some other optical network configurations and comparing those results to the ones generated by Mintaka. As an additional sanity check, the total power of the floor plan units was compared to the laser power plus the estimated electrical switching and static power loss for all the simulations. None of the simulations deviated from the expected power consumption, given the required laser power and the input traffic pattern.

3.2 Trimming Analysis

Considering the thermal sensitivity of microring resonators and the fact that published research indicated trimming required a significant amount of power, I decided to use Mintaka to investigate the impact of active trimming on the power consumption and thermal behavior of an optical network. The base architecture used in the study was a 64 node crossbar network with a 64-bit data path (or phit[4] width) between nodes, built using 16nm technology. The nodes were assumed to operate at 5GHz and were capable of generating and consuming one 128-bit flit[5] per cycle. A crossbar was chosen as the base model because it uses a similar number of microring resonators as other proposed on-chip optical networks, and in addition a very detailed floor-plan was available. In this case, the on-chip network consists of ∼524K microring resonators and occupies an entire level of a 3D stacked processor design, with an area of 484mm^2. The workload used was a synthetic random traffic pattern, since the goal of the work was to determine the power/thermal sensitivity of large on-chip optical networks and not to analyze the performance.

3.2.1 Stability Issues with Trimming

The first step in the investigation was to determine if the network itself was thermally stable (i.e. does the network generate too much internal heat even in the absence

[4]A phit is basic unit of data transfer at the physical layer; in other words, the number of bits used in each physical data transfer.

[5]A flit is a flow control digit, or in other words the smallest unit of flow control.

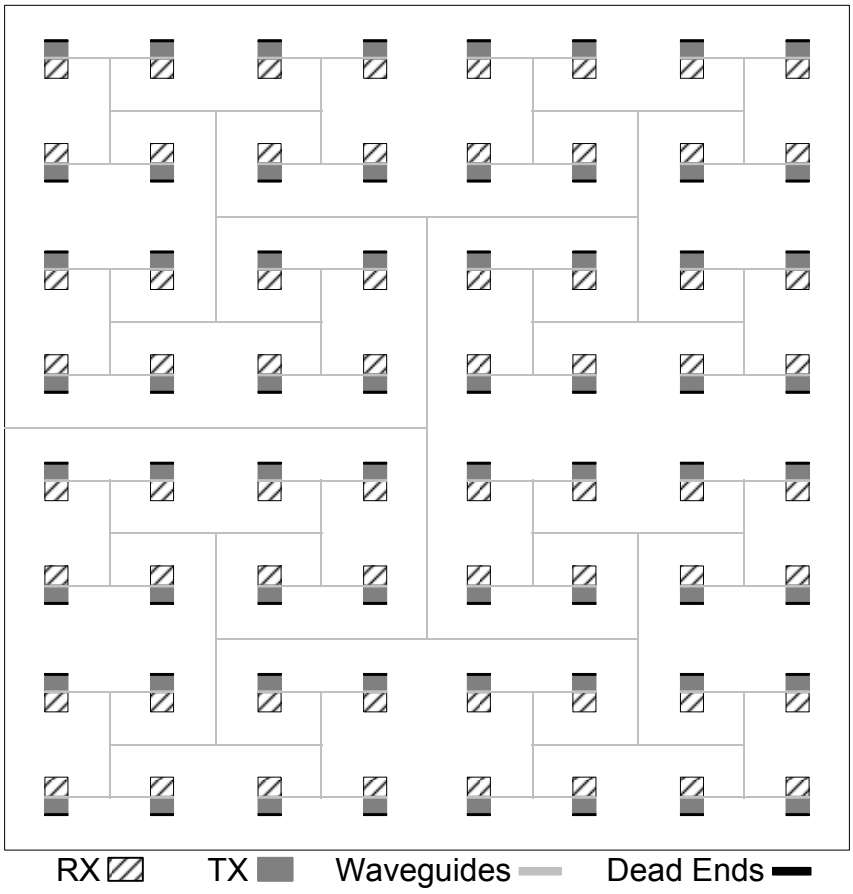


Figure 3.1: Network Floor Plan

of trimming?). Figure 3.1 shows the floor plan of the 64 node crossbar network that was used in this study. Figure 3.2 shows that the network settled at less than half a °C above the ambient. The initial simulation also yielded the "ideal" operating temperature for the network, 45.38°C. This value was used to estimate the required trimming power for the subsequent experiments – as the temperature of the microrings drops below the "ideal" temperature trimming using heating is required, and as the temperature climbs above the "ideal" current must be injected.

The initial thermal simulation established that the network itself was thermally stable, so the next step was to include trimming in the model. The first trimming experiment consisted of varying the ambient temperature from 310K to 325K to determine the required trimming power as a function of ambient temperature. The microring thermal sensitivity was assumed to be 0.09nm/°C, and the channel width was assumed to be 0.16nm.

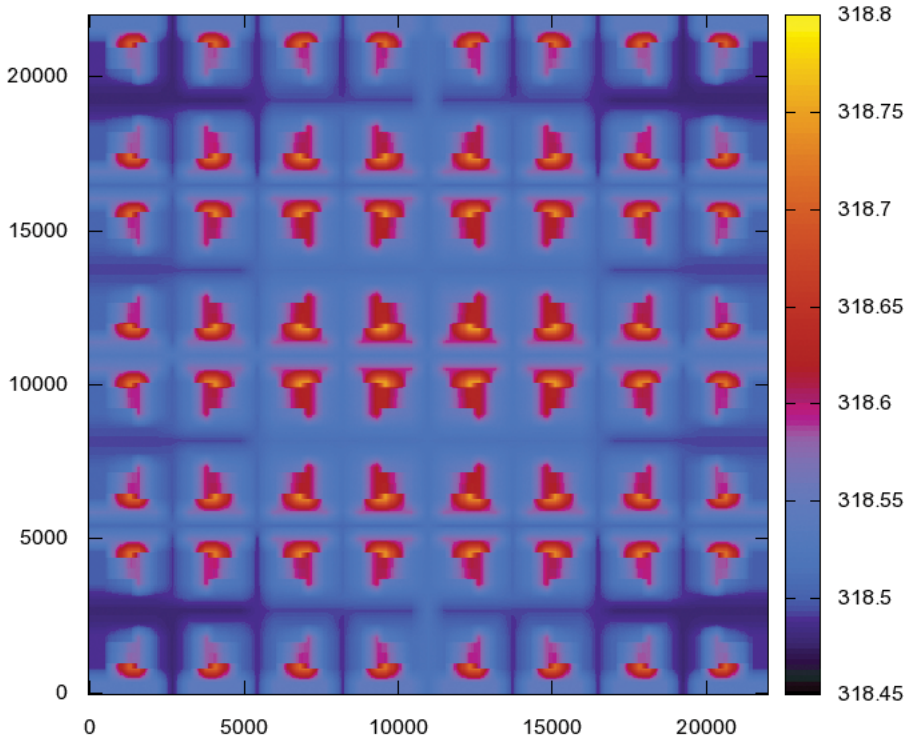


Figure 3.2: Temperature (K) vs. X,Y (μm) for Network with 318K (45°C) Ambient

The microring resonance tolerance was assumed be ±2% – if the resonance drifted more than 3.2pm away from the desired resonance point, the microring was considered out of specification.

The required trimming power was initially assumed to be 130μW/nm [1] for current injection (blue shift) and 240μW/nm [22] for heating (red shift). The fixed value approach for red shift was quickly abandoned in favor of a closed loop solver that determined the actual power required to maintain a minimum required temperature. Since resonance deteriorates under high current, blue shift using current injection will be capable of less than 1nm of shift, so it was assumed that a maximum of 1nm of blue shift could be achieved using current injection.

The results of the trimming power sensitivity simulations can be seen in Figure 3.3. The network required approximately 5.1W of heating for every degree the ambient temperature dropped below the design target of 318.15K (45°C). In addition, the network became thermally unstable within a one degree increase above the optimal ambient temperature – this is because the current injection becomes a positive feedback system with current injec-

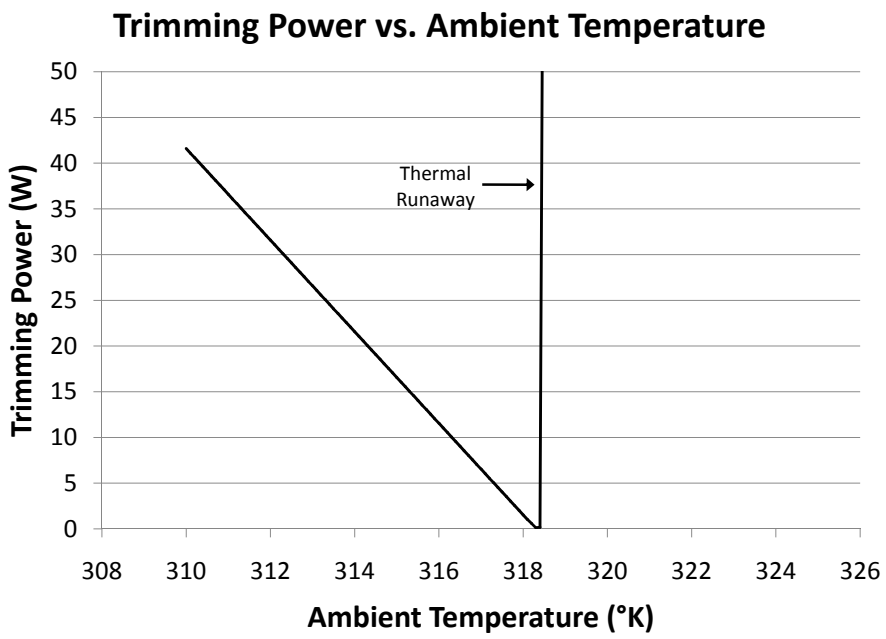


Figure 3.3: Trimming Power (W) vs. Ambient Temperature (K) for Network

tion heating the rings, heat in the rings causing red shift, requiring more current injection, etc. (The network would in fact settle at a steady state when the ambient temperature rose above one degree, but the required blue shift was beyond 1nm, which as previously stated cannot be accomplished using current injection.)

The thermal runaway observed in the baseline model led me to look at a 32-bit version of the network operating at twice the frequency. The 32-bit network uses approximately half as many microrings (∼270K), and it has been assumed by some researchers [1, 40, 70] that reducing the number of microrings will reduce the required trimming power. The results of the trimming power sensitivity simulations using the 32-bit network are shown in Figure 3.4. As one might expect the power required for current injection is lower for the 32-bit network, and the network is slightly more stable than the 64-bit version (becoming thermally unstable within four degrees above the optimal ambient temperature, instead of within a single degree). What seems surprising is that the required power for heating is almost identical for both networks. In situations where heating is required for trimming, the amount of trimming power required appears to have a non-linear relationship with microring count. It was this observation that motivated my investigation into varying the die

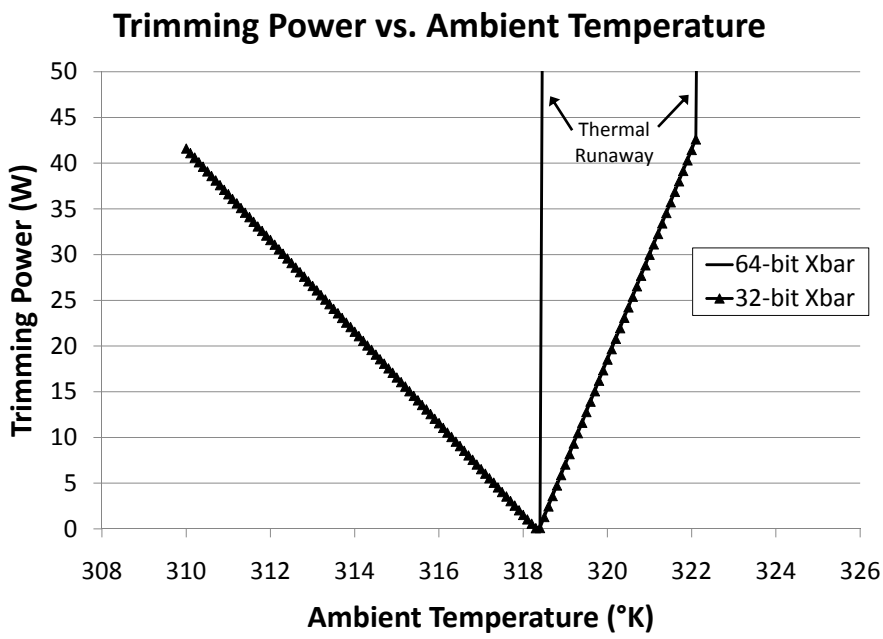


Figure 3.4: Trimming Power (W) vs. Ambient Temperature K for 64-bit and 32-bit Network

area.

Figure 3.5 shows the required trimming power for the 64-bit network with varying die areas. The die areas shown are 484mm^2, 400mm^2, and 324mm^2 (22mm, 20mm, and 18mm squares). The trimming power required for heating is clearly related to (although not directly proportional to) the die area. This is because the power required to maintain a given temperature is dominated by the area that has to be heated and not the number of microrings. Using a simple thermodynamic analysis, the required rate for heating should equal the rate at which heat can be removed from the die – and as the die area is reduced, so is the rate at which heat is removed from the die (the rate of heat transfer is directly proportional to the conduction surface area).

The amount of trimming power necessary for current injection is not visible in Figure 3.5 because of the thermal runaway. Further analysis of the simulation results shows that the trimming power for current injection does appear to have a direct correlation to average microring density. As the die area shrinks the microring density increases, leading to an increase in the required current injection trimming power, because the heating that occurs during injection is spread among the rings. This can be seen in Figure 3.4, since

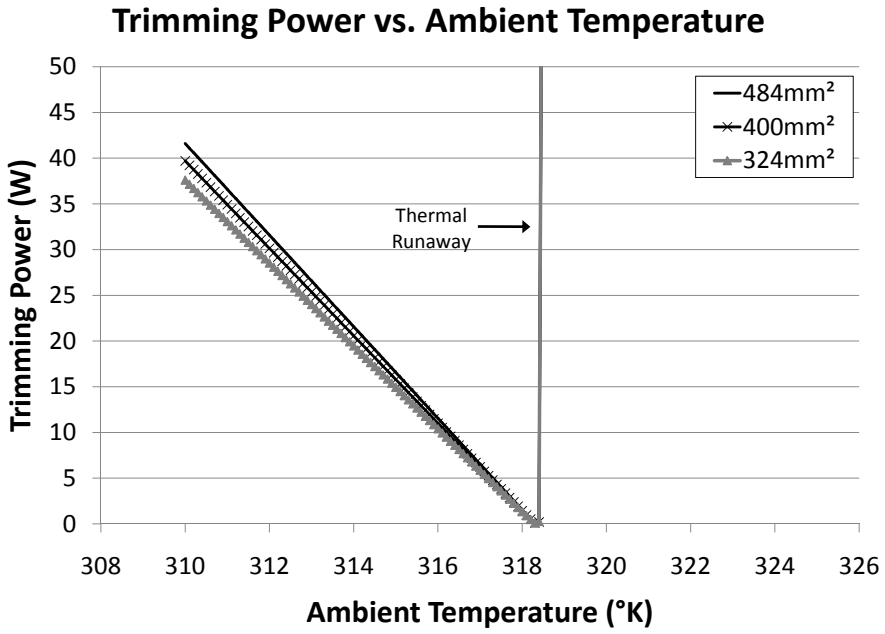


Figure 3.5: Trimming Power (W) vs. Ambient Temperature (K) for 64-bit Network with 484mm^2, 400mm^2, and 324mm^2 Die Area

the 32-bit network has a dramatically lower microring density than the 64-bit equivalent. These results show that reducing the number of microrings or the microring density can reduce the trimming power necessary for current injection, but the only effective method of reducing the required trimming power when heating is to reduce the die area.

We define the Temperature Control Window (TCW) as the range of temperatures within which the network must be kept in order to remain within a given trimming budget and prevent thermal runaway. The TCW for the 64-bit network is less than 1.1°C and 4.1°C for a trimming budget of 5W and 20W, respectively. If heating is the only technique used to keep the temperature within the TCW and avoid thermal drift, then power will constantly be used to heat the microrings unless the system is in an environment on the high edge of the operational range. The simulation results indicate that a 20K TCW as suggested in [70] and [40] would require a maximum trimming power of 103W (51.6W average) for the 484mm^2 die area of the simulated network. Even for the 400mm^2 die area assumed in [40] the maximum power needed for heating across the 20K TCW is 98.9W (49.4W average). According to the American Society of Heating, Refrigerating and Air-conditioning Engineers (ASHRAE) [4] the recommended temperature range of data centers

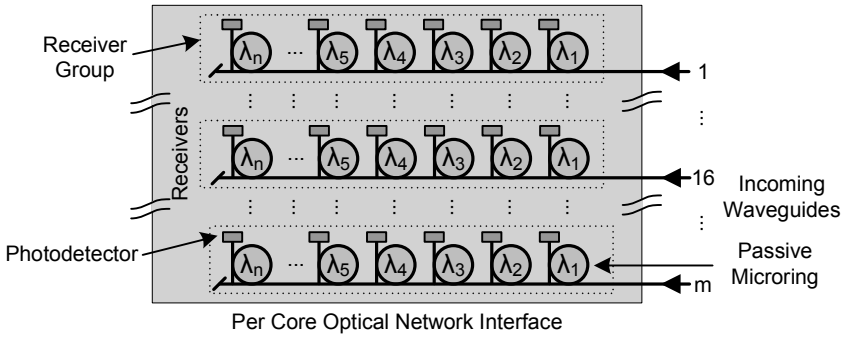


Figure 3.6: Receive Section of a Node Illustrating Microring Grouping

is 18-27°C, which is a 9°C temperature range for the room ambient temperature - the temperature range the chip will experience will most likely be larger. Thus, a 20K TCW is appropriate given the air conditioning recommendations of ASHRAE.

3.2.2 Increasing the Temperature Control Window

The granularity of the floor-plan units used in the previous experiments was set to an entire group of co-located rings used to implement a transmitter or receiver (i.e. 64-bits), and it was assumed a constant temperature was maintained across the entire group. Figure 3.6 illustrates the receive section of a node and shows how microrings were grouped together into single floor-plan units.

This relatively coarse granularity was chosen because simulating approximately half a million floor plan units, which would be necessary if each floor plan unit contained a single microring, is not currently practical. In order to ascertain the impact of the chosen granularity, the simulator was modified to keep track of each individual microring in a specific co-located group of transmitter or receiver rings. Simulations were run over sweeps of ambient temperatures ranges with trimming disabled (having trimming on would defeat the purpose of the test), and transmitter and receiver groups to be studied were chosen strategically (i.e. corners, centers, and edges).

The largest intra-group temperature delta observed was 5.13e-4°C, which corresponds to less than 0.03% channel separation when assuming a wavelength separation of 0.16nm and 0.09nm/°C thermal drift. This is good news for a number of reasons – it shows that the original set of simulations are valid because the results are not sensitive to the

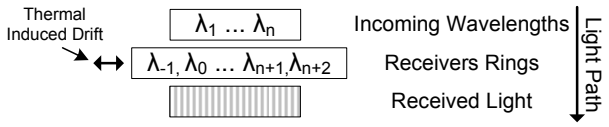


Figure 3.7: Node to Node Drift Resilience

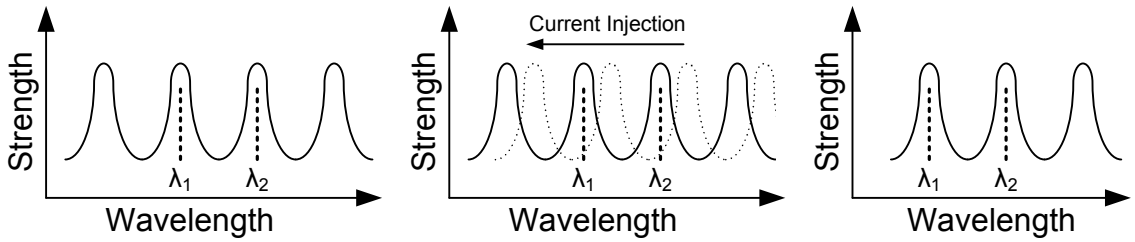
simulated granularity, it means that the amount of circuitry necessary to support trimming can be greatly reduced since microrings can be trimmed as a group instead of needing to be trimmed individually, and perhaps most importantly it means that we can extend the TCW by adding resonators at each end of the incoming wavelength spectra.

3.2.3 Sliding Ring Window

The results presented so far indicate that a reasonable TCW will require a potentially unreasonably large trimming budgets. In order to overcome this problem, I investigated incorporating additional rings on either end of the spectral range in order to maintain the same usable data path width. The additional microrings will create a Sliding Ring Window (SRW), exploiting the fact that the entire group will slide the same amount spectrally in either direction.

This concept is shown in Figure 3.7, and works as follows: current injection is used to maintain the spectral position of the entire group of rings (see Figure3.8(b)) until the rings become so hot that with trimming removed they will resonate at the next (red shifted) frequency (see Figure3.8(c)). At this point the current injection is turned off and the entire group begins resonating one wavelength over. As the rings cool current injection can be reapplied to correct the spectral position of the microrings to the previous (blue shifted) channel. The existence of the additional rings prevents the current injection positive feedback system from thermally running away, by creating a lower power trimming state at a higher temperature. In the following discussion of Figures 3.9 through 3.11 I will designate the number of additional rings in the name, so that SRW-1 indicates 1 extra resonator, SRW-2 indicates there are two extra rings, etc.

Guiding the correct electrical signals to/from the correct microrings of the SRW will require integration with the trimming circuitry. However, the SRW control is expected



(a) On Resonance (b) Blue Shift Needed (c) Remove Current Slide to Next Ring

Figure 3.8: Sliding Ring Window Microring Resonance vs. Wavelength for On Resonance (a), Current Injection (b), and Current Removal (c)

to be of minimal additional circuit complexity since any trimming system implemented must already maintain microrings in the correct spectral position. In addition, the control circuitry for the SRW only requires local information (a global feedback channel is unnecessary) since the since the feedback is the temperature of the microring group.

In order to evaluate the effectiveness of the SRW, I began by running simulations using SRW-2 – increasing the network total microring count to ∼540K. Whenever possible the steady state thermal solver was used, although it was not always possible since the microrings (by design) cycled back and forth between lower temperature/higher power and higher temperature/lower power states. The transitional temperatures were determined using the Hot-Spot time step solver until the total trimming power converged.

Figure 3.9 shows the impact of SRW-2 on the amount of trimming power required. The sawtooth (or wave-like) pattern is due to the fact that the microrings require maximum current injection trimming before becoming hot enough to be on resonance at the next channel. The peak current injection trimming power is a function of the channel separation of 0.16nm – the peak trimming power could be reduced if channels were able to be more densely packed, although this would also shrink the TCW. The TCW for the 64-bit network using SRW-2 is ∼5.6°C for a trimming budget of 10W (which is more than double the ∼2.1°C TCW for the 64-bit network without SRW), and the TCW is greater than 7.5°C for the 20W trimming budget mentioned in the previous section.

3.2.3.1 Increasing TCW with SRW

The SRW mechanism can be expanded to incorporate more than just two additional microrings. Each additional microring per group will result in an additional peak and will increase the TCW. The separation of the peaks seen in Figure 3.9 is ∼1.8°C, which corresponds to the channel separation divided by thermal sensitivity (0.16nm / 0.09nm/°C). Achieving the 20K TCW discussed previously would require roughly nine more microrings (creating SRW-9), raising the total microring count to ∼595K for the 64-bit network.

Expanding the TCW by using additional microrings requires an increase in area, trimming and laser power. Compared to the existing microring count and die area of the proposed 64-bit network the additional area for SRW microrings is not a concern. The required laser power increases with each additional microring since the number of off-resonance microrings which light must travel through is increased, but this increase is also not a great concern since the additional attenuation of off-resonance microrings is relatively small (assumed to be 1.5e-3dB). The increase in trimming power is likely to be the greatest concern, since each microring added will also need to be trimmed. These additional microrings will cause an increase in the peak current injection power seen in the SRW sawteeth, though the increase in peak trimming power is relative to the number of additional rings. The proposed SRW-9 would have an approximately 10% greater peak power than that of Figure 3.9, yielding a roughly 9.7W peak.

3.2.4 Impact of Reducing Microring Thermal Sensitivity

Another promising approach to increasing the TCW is to use rings which have been clad with polymethyl methacrylate [101]. These microrings are substantially less sensitive to temperature variations - unclad rings change approximately 0.09nm/°C, while for Polymethyl Methacrylate (PMMA) clad rings the change is closer to 0.027nm/°C. In order to analyze their impact on the TCW, the original set of simulations were rerun using PMMA clad rings instead of unclad ones. Figure 3.10 shows the results for both the baseline and PMMA-clad 64-bit network. As expected, both networks require the same amount of additional heating power for every degree below the ambient temperature. While

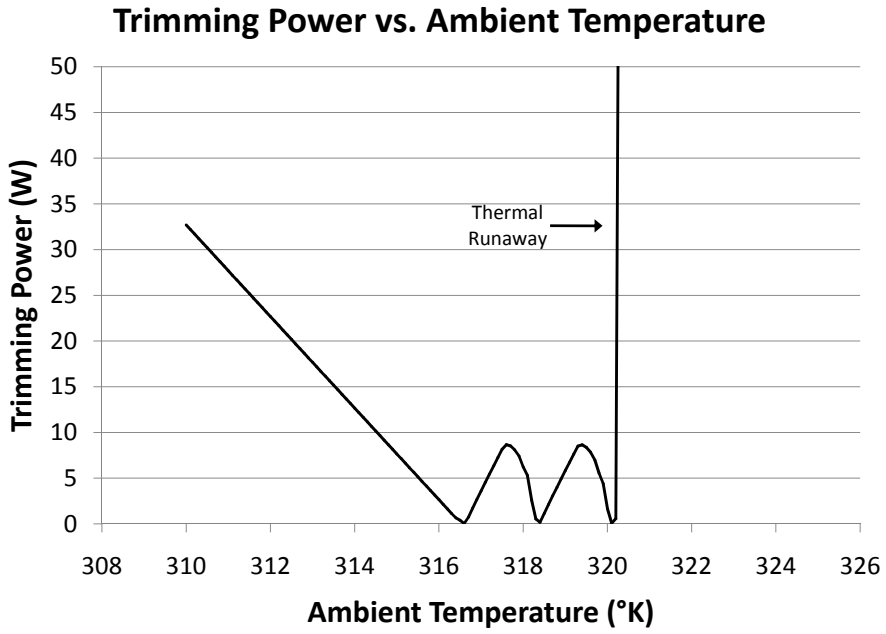


Figure 3.9: Trimming Power (W) vs. Ambient Temperature (K) for 64-bit Network Using SRW-2

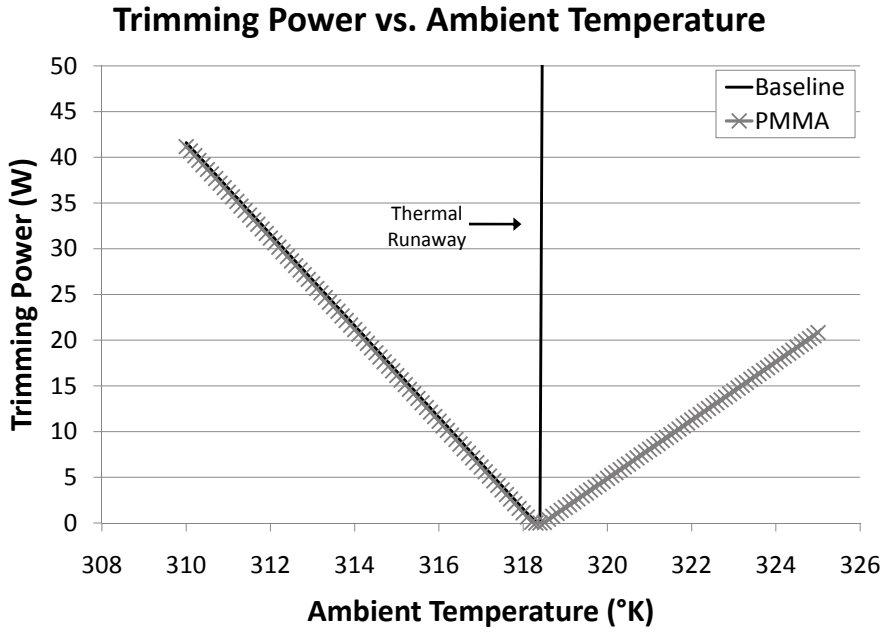


Figure 3.10: Trimming Power (W) vs. Ambient Temperature (K) for Baseline and PMMA 64-bit Network

the PMMA cladding reduces the thermal sensitivity of the microring resonators, it does not change the power required to maintain a minimal temperature (although the minimal

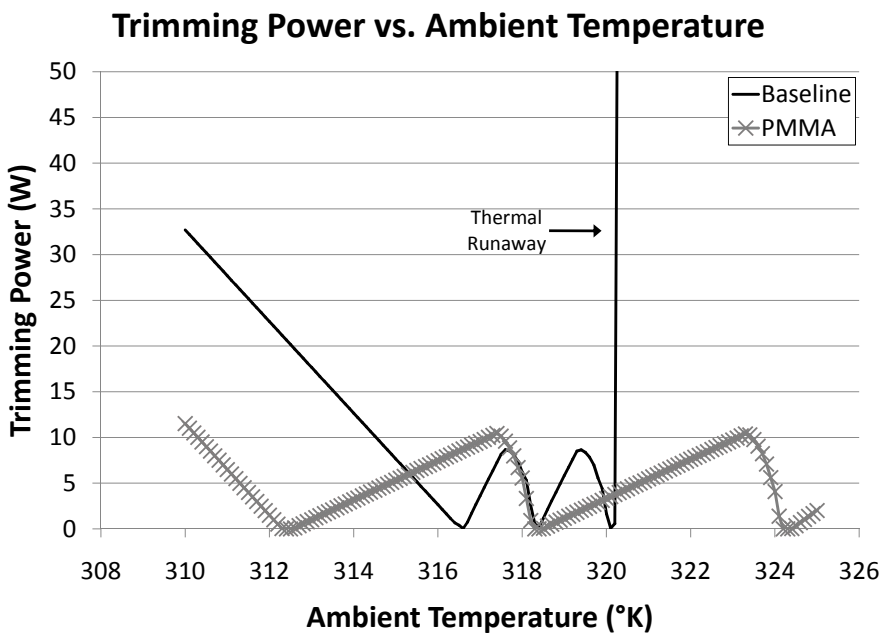


Figure 3.11: Trimming Power (W) vs. Ambient Temperature (K) for Baseline and PMMA 64-bit Network with SRW-2

temperature that must be maintained is slightly lower for the PMMA-clad microrings than the baseline, which can be seen in the offset of the PMMA-clad line in the figure). On the other hand, when the rings are too hot and current injection is required, the PMMA-clad microrings substantially outperform their unclad counterparts. As the figure shows, only ∼3W of current injection is require for every degree the ambient temperature climbs above the optimal. Thus the TCW for a PMMA-clad network is approximately 3°C and 10.6°C for a trimming budget of 5W and 20W, respectively. While this is substantially better than unclad rings, it still implies that a trimming budget of nearly 39W would be required to meet the 20K TCW.

Since cladding and SRW are orthogonal techniques, I decided to see how well they would work in conjunction. The simulations using SRW-2 were rerun assuming PMMA-clad rings, and the results are presented in Figure 3.11. In this figure one can see that the PMMA-clad network using SRW-2 provides a TCW of 19°C with a trimming budget of less than 10.5W, and the 20K target TCW can be obtained with a trimming budget below 12.2W.

Another interesting point to note in the figure is that the peaks of the baseline

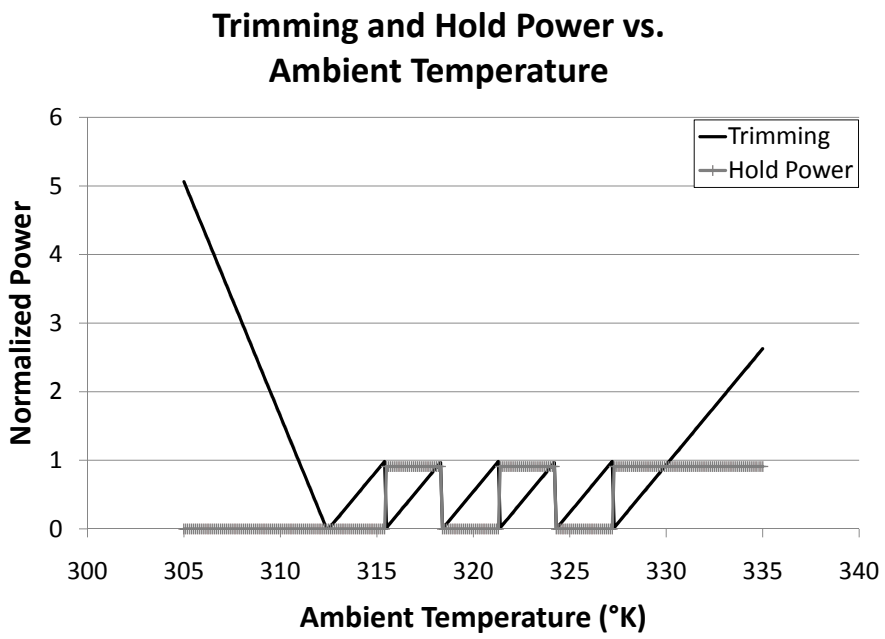


Figure 3.12: Trimming Power and Hold Power vs. Ambient Temperature (K) for PMMA 64-bit Network with SRW-2 and Dual Mode Modulation

network are lower than those of the PMMA-clad network. One would expect both the baseline and the PMMA-clad networks to require the same peak potential trimming power - however, the baseline network has a much higher thermal sensitivity, and therefore it oscillates between lower temperature/higher power and higher temperature/lower power more often than does the PMMA-clad network. Thus, the baseline's peak is actually smoothed out by its thermal sensitivity.

In an attempt to reduce the peak trimming power while maintaining the same operational range I investigated reversing the method of modulation while the microrings were in certain temperature bands. Since the quiescent state of the modulator microrings is off-resonance and current is injected to bring the microring on-resonance, I observed that for certain temperature ranges the modulator microrings would be on-resonance in their quiescent state and current would need to be injected to keep the microrings off-resonance. The simulations showed that little power could be saved by using the proposed dual mode of modulation. While the trimming power was greatly reduced in the bands of reverse modulation, the additional hold power to keep the microrings off-resonance negated the gain. Figure 3.12 shows a normalized example of the trimming power and the hold power

for the microring modulators. Notice that the trimming appears to have three additional sawteeth - this is due to the reverse modulation bands. Also notice that the hold power almost completely negates the reduction in trimming power.

3.2.5 Future Work

The optimal amount of channel separation when using SRW merits further investigation. In my simulations I assumed the channels had minimum possible separation, but it is possible to widen the TCW by separating the channels further, which will increase the peak trimming power. The SRW scheme could be employed without the need of additional microrings if the channel separation times the data path width is equal to the microrings FSR. The use of SRW causes the trimming power to be a non-continuous function – therefore, trimming will be most efficient in periodic temperature bands. An investigation of system level techniques to maintain the network within those temperature bands should be completed.

3.2.6 Trimming Discussion

Based on what has been presented, there are some things that architects must keep in mind when designing large nanophotonic systems:

1. The power required to maintain the "ideal" temperature using heating has a non-linear relationship with microring count and thermal sensitivity, and is more affected by the die area, ambient temperature, and rate at which heat can be removed from the die.

2. A nanophotonic network that only uses heating must keep the microrings at the "ideal" temperature, and if the microring temperature is ever above the ideal the network will not function (since there is no way to move the resonance back toward the blue). Since there are many heat sources in a processor (cores, cache, etc.) the ideal must be set very high, which means that the microring heaters will essentially always be on, pumping heat into the system and potentially raising the operating temperature of the other components (cores, cache, etc.) as well.

3. Trimming using current injection also faces challenges – it is highly sensitive to microring count, density, and the thermal sensitivity of the rings, and thus thermal runaway can happen very easily - networks with ∼524K microrings and a 484mm^2 die area experience runaway within a change of a single degree. This is significant, because optical network topologies have been proposed that employ higher microring counts and densities than those analyzed in this paper. The use of current injection in these networks will only be feasible if they employ techniques like PMMA cladding or the Sliding Ring Window.

4. Any microring based nanophotonic network cooling system will need to be carefully designed, since it impacts the efficiency of the trimming system. The design of a nanophotonic network that trims using only heating will obviously benefit from a simpler cooling system. A cooling system that only removes the amount of heat generated by the laser and modulation power would be ideal, since the microrings could be brought up to temperature and then maintained by the heat generated by the laser and modulation. Conversely, any system that utilizes current injection as a form of trimming will benefit from an efficient cooling system - otherwise, the network will become thermally unstable.

Reducing the thermal sensitivity of microrings has great potential for improving the energy efficiency of trimming. Recently researchers [88, 77, 28] have shown a significant reduction in thermal sensitivity from that of PMMA upper cladding [101] – these researchers have demonstrated the ability to reduce the thermal sensitivity to as little as 1pm/°C. Therefore, in the rest of this dissertation I assume only current injection-based active trimming of microrings with a thermal sensitivity of 1pm/°C and a TCW of 20°C. Current injection-based trimming was chosen over heating since heating (as previously discussed) has the potential for increasing the static leakage of other components in the processor.

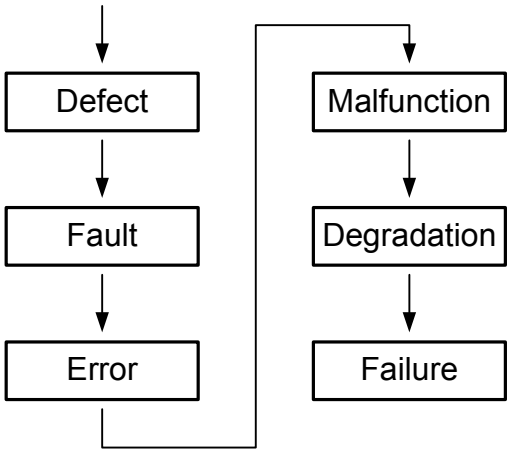
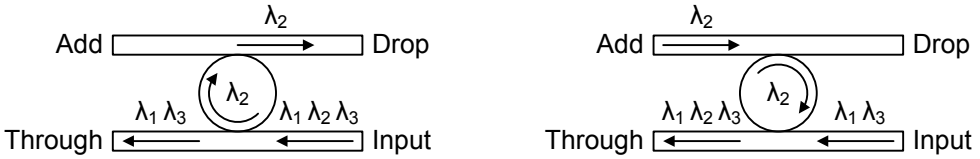


Figure 3.13: Six Level View of Impairments

3.3 Photonic Link Resilience Analysis

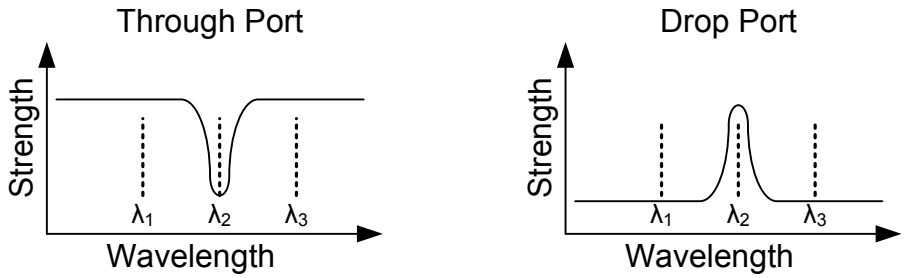
In Section 3.2 it was shown that thermal drift may be addressed using trimming. If fabrication defects are prevalent and post fabrication techniques are not practical then any trimming scheme will be further complicated – trimming rings as a co-located group requires that all rings in the group are uniformly spaced spectrally. Figure 3.13 shows the flow of how a defect may ultimately manifest as a system failure – this follows the terminology presented in [72]. Looking at the flow, it is clear that microring fabrication defects will likely reduce the reliability of individual links (as well as the entire network); therefore, I explored schemes to address the link reliability problems.

In order to evaluate reliability and resilience schemes, information is needed on the types of faults that are likely to occur in the optical domain. Unfortunately, since the fabrication of nanophotonic components is still in the nascent stage, there is very little in the literature on either the nature of defects or how to model them, so it was necessary to derive a microring based optical fault model. Understanding the sources of optical faults and their resultant bit errors enables one to propose techniques which can improve resilience. The performance/power vs. resilience trade-off is well understood in the electrical domain; unfortunately, given the nature of photonics, this trade-off is not as clear in the optical realm.



(a) Filtering λ_2 from Input to Drop (b) Injecting λ_2 from Add to Through

Figure 3.14: Microrings being used to filter (a) and inject (b) λ_2



(a) Power at Through port, Microring Tuned to λ_2 (b) Power at Drop port, Microring Tuned to λ_2

Figure 3.15: Power at Through (a) and Drop (b) ports for Microring Tuned to λ_2

3.3.1 Photonic Link Fault Modeling

This section provides an in-depth explanation of the derived abstractions for the faults in a photonic link and attempts to classify them in a way that is useful to a computer architect, to help them make decisions about the design trade-offs such as performance/power vs. resilience.

3.3.1.1 Microring Resonators

Microring resonators are designed to resonate only when presented with specific individual wavelengths, behaving in essence as band pass/reject filters. They are typically configured to have two input (*input* and *add*) and two output (*through* and *drop*) ports. Figure 3.14(a) shows a microring resonator filtering (removing) λ_2 from the *input* port, which carries multiple wavelengths. Figure 3.14(b) shows a microring resonator injecting λ_2 from the *add* port onto the *through* port, where it joins other wavelengths.

Figure 3.15 shows the theoretical *through* and *drop* power as a function of wavelength for a microring tuned to λ_2. The Y-axis of the graphs in Figure 3.15 represents the

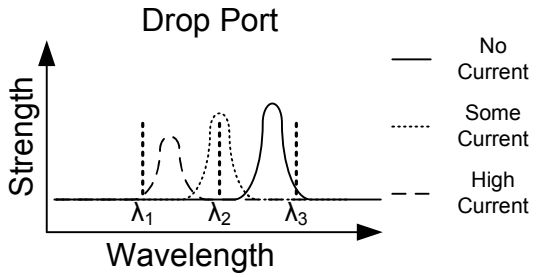


Figure 3.16: Degradation in signal quality

signal strength, which is the percent of power provided at the *input* port that makes it to the *through* port (Figure 3.15(a)) or the *drop* port (Figure 3.15(b))[6]. The same is true of Figure 3.15(b) which represents the percent of power for *input* → *drop* and *add* → *through*.

As discussed previously the resonance frequency of a microring can be changed by heating the microring; in addition, injecting current into the microring can also change the resonance frequency, but in the opposite direction. Microrings respond quickly enough to current injection that it is also be used for modulation; unfortunately, current injection causes a significant degradation in the quality of the modulated signal. Figure 3.16 shows that as current is injected, the resonance wavelength moves to the left and the strength of the signal decreases. Signal degradation due to current injection is a further complication to the implementation of reliable nanophotonic networks based on microring resonators. The proposed methods of modulation and blue shift trimming increase the likelihood that either insufficient light will transfer from the *input* port to the *drop* port (if filtering wavelengths) or from the *add* port to the *through* port (if injecting wavelengths).

3.3.1.2 Photonic Waveguides

Unlike electrical wiring, photonic waveguides are designed to carry multiple bits of information along a single waveguide. Photonic waveguides have relatively low signal crosstalk [59] and are capable of carrying signals over a longer distance at higher signaling rates with lower losses than their electrical counterpart. Signals do suffer some losses

[6]The units were purposely left off since the actual values are not important, but ideally the top value would be 100% (0dB attenuation) and the bottom would be 0% ($-\infty$dB attenuation). The function shown in Figure 3.15(a) represents both the percent of power that reaches the *through* port from the *input* port, as well as the percent power that reaches the *drop* port from the *add* port.

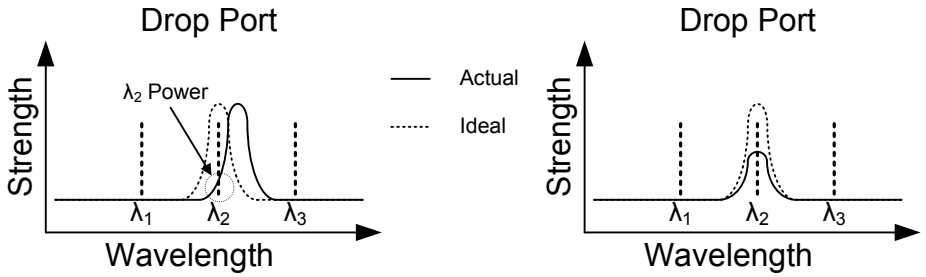
in waveguides, however, due to effects such as scattering and radiation mode coupling. According to Lipson in [52], waveguide scattering losses are highly dependent upon the fabrication process. Increased waveguide losses (or higher path attenuation) reduces the photonic power that will reach the photodetectors, and thus must be accounted for in the fault models.

3.3.1.3 Faults

Faults can be classified as either permanent or temporary. Permanent faults are primarily due to fabrication errors, while temporary faults may be due to environmental factors such as fluctuations in temperature or ElectroMagnetic interference (EMI). Permanent faults that cannot be overcome through architectural resilience techniques will lower the fabrication yield of on-chip photonic networks. Temporary microring faults due to temperature fluctuations can cause higher path attenuation, since the shifting of the resonant wavelength due to temperature changes can increase the ring attenuation (if not perfectly corrected via trimming). However, the ultimate impact on the photonic network is whether or not faults manifest themselves as bit errors, not if the fault is temporary or permanent.

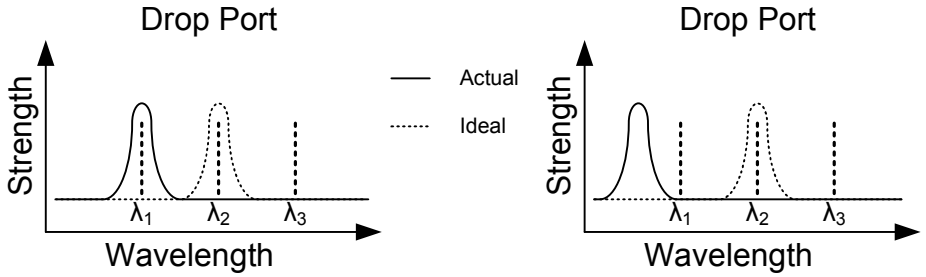
Microrings that do not resonate at their designed spectral position and waveguides with increased attenuation will be abstracted further. Microrings that do not resonate as designed will be considered faulty, which will happen if they are resonating to the wrong wavelength, the signal attenuation is too great, or both. The two exclusive cases are illustrated in the graphs in Figure 3.17. In this figure the dashed lines show the desired function, while the solid lines show the actual behavior. Figure 3.17(a) shows that when the microring is not accurately tuned to the desired wavelength λ_2, the amount of λ_2 that appears on the *drop* port is very small – the amount of the λ_2 line that lies below the solid black line. This misalignment could be the result of thermal drift, improper fabrication, insufficient trimming, etc. Figure 3.17(b) shows the power to the *drop* port of a microring that is excessively attenuating the signal, which results from improper fabrication or too much current injection (since resonance deteriorates during current injection).

Some of the behavior of faulty optical components (such as higher than designed attenuation or resonance drift) can potentially be overcome by increasing the amount of



(a) Drop for Off Resonance Microring (b) Drop for Attenuated Microring

Figure 3.17: Drop for Off Resonance (a) and Attenuated (b) Microrings Designed to Resonate on λ_2

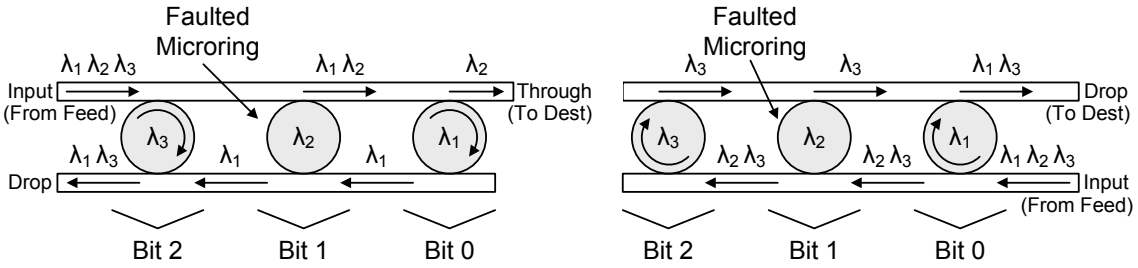


(a) Drop for Interfering Microring (b) Drop for Non-Interfering Microring

Figure 3.18: Drop for Interfering (a) and Non-Interfering (b) Microrings Designed to Resonate on λ_2

system power used. Off resonance microrings that are being trimmed, for example, simply require more signal power to operate correctly. Waveguides that have higher path attenuation than normal may also be compensated for by increasing the photonic power, enough so that sufficient power reaches the photodetectors. The model presented in this section focuses on the cases for which addressing the fault will not be as simple as increasing the power.

Microrings that do not resonate at their designed spectral position (as in Figure 3.17(a)) can be put into one of two categories, *interfering* or *non-interfering*. Interfering microrings are those whose resonance frequency has drifted so far that they are actually resonating at another wavelength channel. Figure 3.18(a) shows the power to the *drop* port of an interfering microring that is designed to resonate at λ_2, but is interfering with λ_1. Non-interfering microring are those that do not resonate at the desired wavelength, but do not interfere with any other wavelength either. Figure 3.18(b) shows the power to the *drop*



(a) Modulating Zero (Removing Wavelengths) (b) Modulating Ones (Adding Wavelengths)

Figure 3.19: Faulty Microring for Modulation of Zeros (a), and Modulation of Ones (b)

port of a non-interfering microring that is designed to resonate at λ_2, but is resonating below λ_1. Microrings that have increased attenuation are considered to be non-interfering microrings, since the end result is the same as a slightly off-resonance non-interfering microring (in both cases, a diminished amount of the desired wavelength appears on the output port.) A microring that is partially interfering (less than 100% interference) is modeled as a non-interfering microring if the interference is insufficient to cause a bit error in the other wavelength channel, otherwise it is modeled as a interfering microring.

3.3.1.4 Link Component Structure Dependent Errors

The types of errors that will result from faults depend upon the structure of the link components. For the time being we will focus on the transmitter and receiver sections of the on-chip optical networks, since the proposed networks all have similar transmitter/receiver structures and differ primarily in the interconnection topology. Transmitting data is done in one of two ways: by actively modulating ones (transitioning wavelengths from the *input* waveguide to *drop* waveguide) or by actively modulating zeros (removing wavelengths from the *through* waveguide). The receiver section for a link will consist of a set of microring resonators that are either always on-resonance (as in [92]), or enabled whenever a message is sent (as in Single Writer Multiple Reader (SWMR), proposed in [71]).

3.3.1.5 Errors Resulting from Non-Interfering Microring Faults

Non-Interfering faults do not move the desired wavelength from the *input* port to the *drop* port (or *add* port to *through*), but do not transition any other wavelength either.

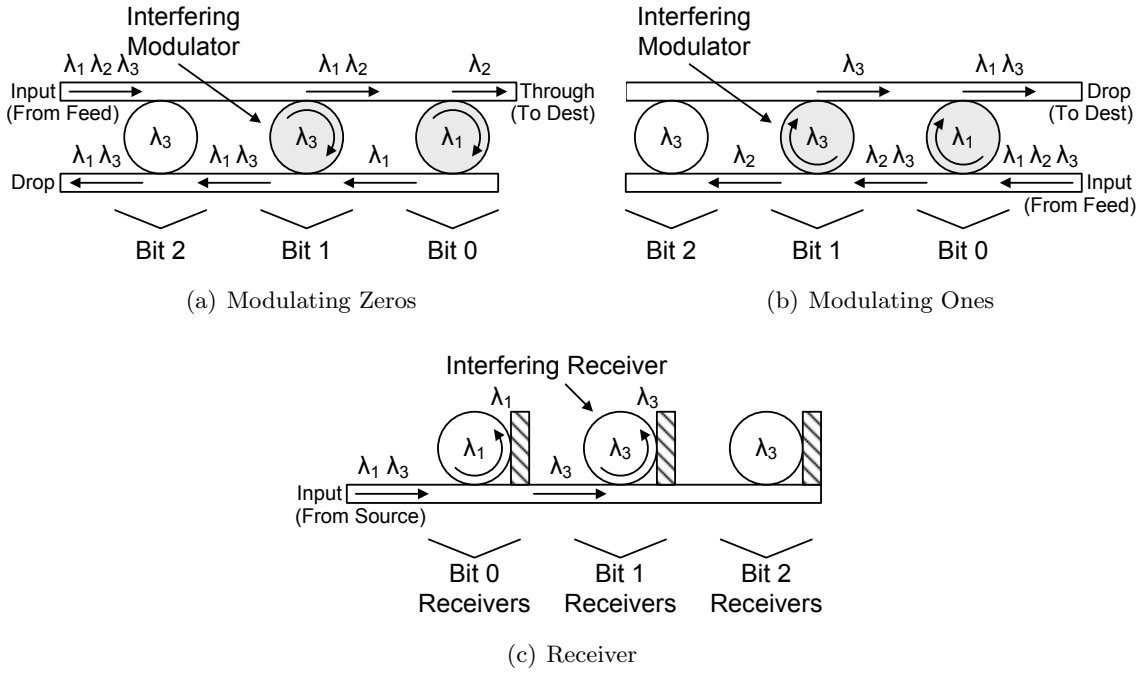
Thus, a non-interfering faulty microring that is in the receiver section will result in zeros always being received for that bit, since the proper wavelength will never transition from the *input* port to the *drop* port. This is essentially a "stuck-at-zero" fault, and only results in a bit error when a one is being sent on that bit.

The types of errors generated by a non-interfering faulted microring in the transmitter section will depend upon the method of modulation. In the case where zeros are actively modulated (a wavelength is removed from the *through* waveguide), a faulty microring will result in the wavelength always being present at the destination (a one will always be detected, which corresponds to a "stuck-at-one" fault). This is shown in Figure 3.19(a), where a three bit transmit section is attempting to send all zeros. The bit 1 modulator is faulty, so λ_2 is not being removed from the waveguide.

In the case where ones are actively modulated (a wavelength is transitioned from the *input* to *drop* ports), a faulted microring will result in its resonant wavelength never being present at the destination (a zero will always be detected, which corresponds to a "stuck-at-zero" fault). Figure 3.19(b) illustrates a three bit transmit section that is attempting to send all ones - again, bit 1 has a faulted modulator and therefore is not transitioning λ_2 from the *input* feed to the *drop*. In summary, non-interfering faulty microrings in the transmitter and receiver sections will result in "stuck-at" faults.

3.3.1.6 Errors Resulting from Interfering Microring Faults

Interfering faults are much more problematic than non-interfering faults. It is possible for double bit errors to occur when an interfering faulty microring is involved, for example. Figure 3.20 illustrates double errors for both forms of modulation and for reception. Figure 3.20(a) shows a three bit transmit section attempting to transmit the value 100, but since bit 1 is interfering with bit 2 (λ_3 is removed instead of λ_2), a double error occurs. In Figure 3.20(b), the transmit section is attempting to send the value 011, but again bit 1 is interfering with Bit 2 (λ_3 is transitioned instead of λ_2), resulting in a double error. Finally, Figure 3.20(c) shows a three bit receive section that has been sent the value 101, but since bit 1 is interfering with bit 2 (λ_3 is removed by Bit 1 instead of by Bit 2) a double error occurs.



(a) Modulating Zeros

(b) Modulating Ones

(c) Receiver

Figure 3.20: Double Bit Error from Interfering Microring Fault for Modulation of Zeros (a), Modulation of Ones (b), and Reception (c)

Interfering modulators will result in the interfer*ing* bit being "stuck-at" (similar to a non-interfering fault), and the interfer*ed* bit being a logical function of the interfering and interfered bits (similar to a "bridged" fault). In the case where zeros are actively modulated, the interfered bit will be a logical AND of the interfering and interfered bits, since either modulator will remove the wavelength in the case of a zero, and only both bits being a one will result in the wavelength passing unperturbed. In the case where ones are actively modulated, the interfered bit will be a logical OR of the interfering and interfered bits. The case where ones are actively modulated is symmetric to that of the case where zeros are actively modulated, as one might expect.

In the receive section, microrings that are resonating at another wavelength may or may not actually be interfering. Looking at Figure 3.20(c) it should be clear that Bit 2 cannot interfere with Bit 0, even if it is resonating at λ_1. A microring resonating at another wavelength but not interfering behaves like a non-interfering microring ("stuck-at-zero"). However, in the case where one microring *is* interfering with another, the interfered bit will manifest as a "stuck-at-zero", and the interfering bit will receive the interfered bit's

information.

3.3.1.7 Unidirectional Bit Errors

The choice of modulation and reception structures can lead to an asymmetry of errors when certain faults occur. It is clear that interfering faults can lead to double bit errors, but non-interfering faults lead to errors in a single direction. Modulators that actively modulate zeros will have non-interfering faults that yield $0 \rightarrow 1$ bit errors. Non-interfering faults for receivers and modulators that actively modulate ones will cause $1 \rightarrow 0$ bit errors, but not $0 \rightarrow 1$ bit errors since they are "stuck-at-zero" faults. In addition, increased waveguide path attenuation will only lead to $1 \rightarrow 0$ bit errors (since insufficient photonic power to switch from $0 \rightarrow 1$ reaches the photodetector).

Non-interfering faults in components result in the following unidirectional bit errors:

- **Modulator (Active Zeros)** – Light will not be successfully removed from the *through* waveguide. When attempting to send a zero, a one will be received. Results in a $0 \rightarrow 1$ (stuck-at-one) bit error.
- **Modulator (Active Ones)** – Light will not be successfully transitioned to the *drop* waveguide. When attempting to send a one, a zero will be received. Results in a $1 \rightarrow 0$ (stuck-at-zero) bit error.
- **Receiver** – Light will not be successfully transitioned from the *input* to the photodetector. When attempting to send a one to it, a zero will be received. Results in a $1 \rightarrow 0$ (stuck-at-zero) bit error.
- **Waveguide** – Increased waveguide attenuation results in insufficient light being received at the end of the waveguide. When attempting to send a one, a zero will be received. Results in a $1 \rightarrow 0$ (stuck-at-zero) bit error.

The type of single bit errors that will occur in a photonic link can be designed to be unidirectional if the correct link component structure is chosen. This is significant, because unidirectional errors can be dealt with more efficiently. If one is willing to give up some bandwidth and separate the channels more, one may be able to minimize/eliminate interfering faults.

3.3.2 Link Reliability/Throughput Trade-off

Improving communication link reliability can be accomplished by increasing the probability that each transmission will be received correctly, by retransmitting until the transmission is received correctly, or both. Increasing the probability of a correct reception can be done using fairly straight-forward techniques, such as reducing the error rate (reducing the device fault rate) and/or adding bits in order to correct for errors. Retransmitting messages until they are properly received is a little more complicated, since it requires a feedback communication link and a communication protocol.

A common method of providing reliable data transmission over an unreliable communication channel is to use an Automatic Repeat reQuest (ARQ) protocol. In order to implement an ARQ protocol, errors must be detectable, and since the communication rate of on-chip networks is very high, the codes used for error detection must enable fast encoding and decoding. The implementation of an ARQ protocol and the additional error detecting bits will reduce the potential bandwidth of a given network, but will increase its resilience.

An ARQ protocol alone will not guarantee reliable communication for all fault sources, however. If the faults are permanent, for example, ARQ protocols will unsuccessfully repeat transmissions until the maximum retransmission count is reached. To circumvent this problem, a Hybrid Automatic Repeat reQuest (HARQ) protocol can be employed, which utilizes Forward Error Correction (FEC) in order to correct a small number of errors and only requests retransmission for cases where transient errors cannot be corrected. Thus, a HARQ protocol can make on-chip networks reliable even in the presence of some extended (or even permanent) faults, as long as they are correctable by the FEC. There are two main types of HARQ protocols; Type I, in which the FEC bits are sent with each transmission, and Type II, which sends error detecting bits with the initial transmission and only sends FEC bits if needed (FEC bits are not sent at all if the transmission is received correctly) [51]. Type II HARQ protocols do not map well for a parallel data path, because of mismatches in the sizes of data path and messages - therefore, this work focuses only on Type I HARQ protocols.

3.3.2.1 Error Detecting Codes

Traditional error detection and correction techniques are well understood and have been implemented in electrical systems for decades [84]. In this subsection I discuss the most appropriate candidate codes to use for error detection, and provide a justification for choosing to include or omit them in my further analysis.

Cyclic Redundancy Check (CRC) is one of the most widely used error detection codes in digital networks and storage devices. An n bit CRC is capable of detecting any single error burst of up to n bits in length. CRC may not be well suited for this environment, though, since communication is not a serial stream of bits (making burst errors less likely), and the block length is relatively short. Furthermore a CRC code is typically calculated in hardware using a Linear Feedback Shift Register (LFSR), which would have difficulty keeping up with the communication rates of on-chip networks (although parallel implementations do exist [11, 55]).

Berger codes can detect any number of unidirectional bit errors with the addition of $k = \lceil log_2(n+1) \rceil$ check bits, where n is the number of data bits [8]. The efficiency of the coding makes Berger codes good candidates for use in this setting - unfortunately, Berger codes require the computation of the *weight* of the codeword. This can be done using $log_2(n)$ layers of adders, arranged in a tree structure – the use of fine grain pipelining may possibly allow for the data throughput requirements to be met, but considering the required complexity, they will not be explored further in this work.

Extended Hamming [29] codes for Single Error Correction and Double Error Detection (SECDED) have been used in a number of memory systems, including the CRAY-1. The same SECDED codes can be utilized as a Triple Error Detection (TED) code, if no correction is performed. A SECDED or TED code can be implemented for 64 and 32 data bits with the addition of 8 and 7 check bits, respectively.

Another approach to error detection is to use multiple signals to transmit a single bit of information. This approach is commonly used in high speed communications, such as Low Voltage Differential Signaling (LVDS) [5]. Multi-Bit Differential Signaling (MBDS) [49]

Table 3.2: N Choose K Code Counts

N	K	Codes	Bits/Block	Efficiency
2	1	2	1	50%
4	2	6	2	50%
6	3	20	4	66.7%
8	4	70	6	75%
10	5	252	7	70%
12	6	924	9	75%

has been proposed to overcome the low code rate efficiency[7] of LVDS, and has been suggested for use in short range (board to board) optical communication [14]. These approaches are essentially an N choose K (NcK) encoding - for example, LVDS is a 2C1 encoding, since only one of the two signals will be a one at any given time. NcK encodings can detect all odd number of bit errors, and may be able to detect some even number of bit errors as well. Significantly, NcK encoding can detect *any* number of unidirectional errors.

The number of valid codes in an NcK encoding is $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. A single NcK block need not be used to cover the entire data width – multiple NcK blocks can be used to create a larger data path while maintaining the same error detection capability of a single NcK block. Table 3.2 shows the number of codes and bits per block that can be encoded using various values for N and K. Looking at Table 3.2, it should be clear that a single 4C2 block does not improve the coding efficiency over two 2C1 blocks (although two 4C2 blocks does yield 36 codes, which is sufficient to encode 5 bits). Encoding and decoding of the NcK blocks will need to be efficient in order to work at the speeds necessary in this environment.

3.3.2.2 Forward Error Correction (FEC)

As discussed previously, HARQ requires FEC. One possible FEC code that could be used is the extended Hamming SECDED code described earlier. Another approach is to combine the NcK encoding with either a parity block or a Reed Solomon code. Since any odd number of bit errors can be detected with an NcK encoding, the detected errors could be treated as block erasures, and an additional parity block could be used to recover from

[7]The code rate efficiency is measured as $\frac{K}{N}$ where K is the number of bits of raw information and N is the number of bits used in the encoded word.

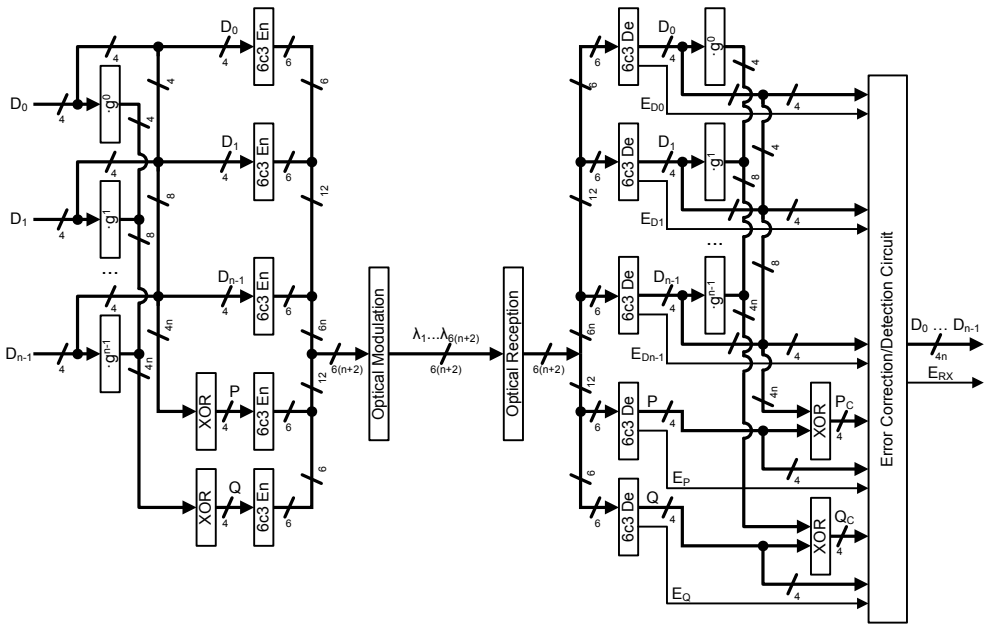


Figure 3.21: Reed Solomon Integrated with 6C3 Encoding Circuit

a single erasure (as is commonly done in RAID-5 [26]).

Redundancy could be extended to protect against double erasures as in RAID-6, as long as the size of the Galois Field (GF) being used for the Reed Solomon code blocks is large enough. A $\mathbf{GF}(2^n)$ can cover $2^n - 1$ data blocks; therefore, a 2C1 code could only cover a single data block, while a 6C3 could cover 15 data blocks $(2^4 - 1)$ or up to 60 bits of data. Equations 3.1 and 3.2 show how parity and the Reed-Solomon code is calculated, respectively:

$$\mathbf{P} = D_0 + D_1 + ... + D_{n-1} \tag{3.1}$$

$$\mathbf{Q} = g^0 \cdot D_0 + g^1 \cdot D_1 + ... + g^{n-1} \cdot D_{n-1} \tag{3.2}$$

In these equations, "addition" is handled by an XOR, and "multiplication" is done in the GF. At first glance it may seem that calculating the GF multiplication may be too complex, but since I am proposing only a 4-bit code word and the multiplication is being done with a constant value, it can be realized with a simple look-up-table. Figure 3.21 illustrates a potential Reed-Solomon circuit that utilizes 6C3 block encoding. Details of the error correction/detection circuit have been omitted as they are not pertinent to the discussion, but would primarily be composed of a network of multiplexers.

3.3.3 Throughput Experiment

Since this area is so new, there are no measured fault and error rate numbers to work with. Therefore, I developed an optical link simulator, which uses statistical sampling to determine the average rate of error for various proposed detection/correction schemes.

The simulator categorizes each transmission as:

- **correct** – A correct transmission occurs when there are no bit errors.
- **incorrect** – An incorrect transmission occurs when an error goes undetected.
- **detected error** – A detected error is returned when the encoding detects an uncorrectable error.
- **corrected** – Corrected results from an error being detected and the data being corrected properly.
- **corrected wrong** – When an error is detected but corrected improperly.

The simulator takes as input the number of faults F (which can be interfering or non-interfering), the encoding scheme, and the number of samples to perform. The simulator selects the F microrings which will have faults at random, and then a (random) bit pattern is created and sent to the receiver. The simulator determines the pattern that is detected at the receive side, and if it is correct, incorrect, has a detected error, is corrected, or is corrected wrong. This process is repeated for each sample until the count reaches the desired number (10M samples per configuration in this case). By doing this statistical sampling one can determine the average rate of corrects, incorrects, etc. for a given number of faults. Using a set probability of a microring faulting, one can also determine the probability of 1 fault, 2 faults, etc. All this information can be combined to enable one to determine the probability of an undetected error given a set probability of a single ring faulting.

The error detection techniques simulated were 32-bit TED (TED32), 64-bit TED (TED64), 32-bit 2C1 (2C1-32), and 32-bit 6C3 (6C3-32). The error correction techniques I examined were 32-bit SECDED (SECDED32), 64-bit SECDED (SECDED64), 32-bit 2C1 with parity (2C1P-32), 32-bit 6C3 with parity (6C3P-32), and 32-bit 6C3 with Reed-Solomon (6C3RS-32). The 32-bit or 64-bit versions of the encoding schemes refers to the number of bits of information being encoded, not the number of bits in the encoded word. Table 3.3 shows the number of non-interfering (NI) and interfering (I) microrings that each

Table 3.3: Error Detection/Correction

Encoding	Max Detect				Max Correct			
	NI 0's	**NI 1's**	**I 0's**	**I 1's**	**NI 0's**	**NI 1's**	**I 0's**	**I 1's**
TED32	3	3	1	1	N/A	N/A	N/A	N/A
TED64	3	3	1	1	N/A	N/A	N/A	N/A
2C1-32	1	ANY	0	0	N/A	N/A	N/A	N/A
6C3-32	1	ANY	0	0	N/A	N/A	N/A	N/A
SECDED32	2	2	1	1	1	1	0	0
SECDED64	2	2	1	1	1	1	0	0
2C1P-32	2	ANY	1	1	1	1	0	0
6C3P-32	2	ANY	1	1	1	1	0	0
6C3RS-32	2	ANY	1	1	1	2	0	0

technique is guaranteed to detect or correct, respectively. The nomenclature of 0's and 1's refers to whether zeros or ones were being actively modulated. Notice that for NcK protocols, error detection and correction capabilities when ones are actively modulated are greater than or equal to that when zeros are actively modulated.[8]

In order to evaluate the negative impact on network throughput when using the ARQ and HARQ protocols, a separate link simulator was developed. The simulator determines the average throughput per cycle given link latency, error rates (for both data and feedback channels), data path width, and packet width. The results presented here assumed an error-free feedback link – simulations were performed in which the feedback channel was faulty, but the results indicated that the throughput is more dependent upon the error rate of the data channel than it is upon the error rate of the feedback link.

Only two of the three main types of ARQ protocols, Go-Back-N (GBN) and Stop-And-Wait (SAW), were evaluated when calculating the maximum throughput. The third, Selective Repeat, was not analyzed since it requires the packet segments or flits each contain a unique segment identification number for the selective retransmission, and this additional information would greatly impact the payload capacity of each segment.

The GBN protocol is relatively simple to implement and has the potential to

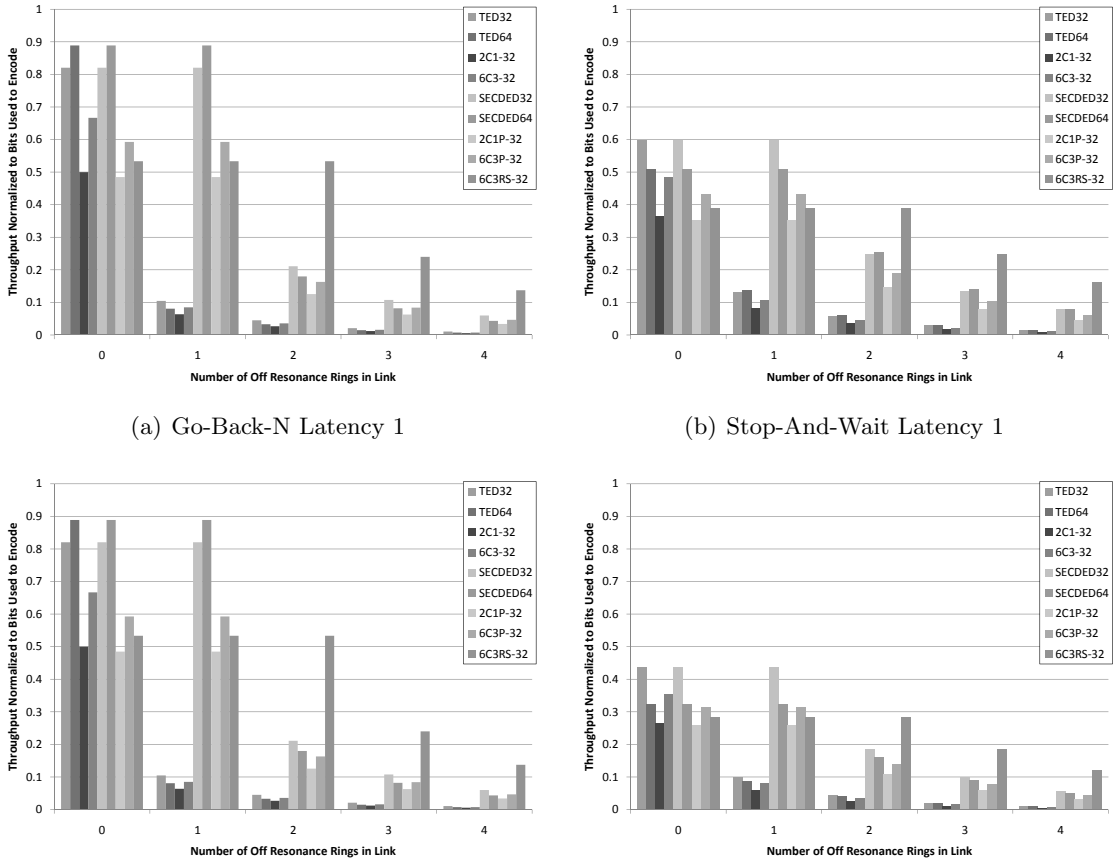
[8]Actively modulating a one means that the microring resonator directs a wavelength onto a waveguide, while modulating zeros means that the resonator removes a wavelength from a waveguide. This is explained in greater detail in Section 3.3.1

maximally utilize the data channel, although it may not be suitable for all network topologies due to its reliance on continual transmission of unacknowledged packets. As stated earlier, however, in photonic systems the external laser power is a static overhead, which means the cost of transmitting is essentially pre-paid. The more you transmit, the lower the average cost/bit becomes, making this approach quite attractive.

The use of a SAW protocol in the optical domain is not a completely new concept – the authors in [15] suggest the use of a similar protocol for signaling a dropped packet in the Phastlane architecture. In Phastlane, a Negative AcKnowledge (NAK) is sent back if the packet is dropped due to insufficient buffer space, and if the NAK is not received within a specific time window it is assumed that the packet has been successfully received and buffered. A NAK only SAW will not protect against errors in the feedback link, since it is impossible to distinguish between a lost NAK and a NAK never being sent, but this protocol could be extended to provide reliable communication by changing the NAK to an ACKnowledgement (ACK) and retransmitting if the ACK is not received.

Figure 3.22 shows the maximum throughput results for the two ARQ protocols, with the normalized throughput on the Y axis and the number of non-interfering microrings that were faulted on the X axis. The throughput results are normalized to the number of bits required for each encoding. Packets of 256-bits were assumed, and ones were actively modulated. Notice that the GBN protocol appears immune to link latency - this is because the window size is sufficient for the latencies presented. Eight cycles are required to transmit a single packet with a 32-bit data channel, so a window size of two is enough to receive an acknowledgement before "going back n" even when the link latency is three cycles. It should also be noted that the SECDED codes are only capable of reliably correcting or detecting up to two misbehaving rings - the results for three and four off resonance rings are provided purely for the sake of comparison.

In the absence of faults, TED and SECDED are the most efficient coding schemes analyzed in terms of effective data throughput given the amount of bandwidth used to encode. However, as faults start to occur (and errors are manifested), the differences between the schemes become noticeable. SECDED in particular performs better than TED when there are single and double non-interfering faults, since a double fault does not guarantee



(a) Go-Back-N Latency 1

(b) Stop-And-Wait Latency 1

(c) Go-Back-N Latency 3

(d) Stop-And-Wait Latency 3

Figure 3.22: Normalized Throughput

that there is a double error.

What may seem somewhat counter-intuitive is that the 64-bit versions of TED and SECDED are less efficient than their 32-bit counterparts when the SAW protocol is used. This result is a byproduct of the fact that 64-bit versions have more "unused" bandwidth while waiting for the acknowledgement. Since the number of cycles that the link is stopped is the same for all encodings, the 64-bit versions wind up having a lower utilization of the data link.

The results show that the use of FEC can dramatically improve throughput if microring faults are common. The efficiency of the SECDED encoding for single ring faults is evident – it is likely a desirable choice when single ring faults are typical, and three or more microring faults rarely occur (since in that case errors could be corrected incorrectly

or go completely undetected). The 6C3RS-32 encoding, on the other hand, clearly has the highest throughput efficiency (actual throughput / raw potential throughput) in the case of two or more non-interfering faulted microrings. Ultimately, the choice of which encoding and protocol to use will be driven by the reliability of the underlying nanophotonic devices, and the actual required throughput of the link.

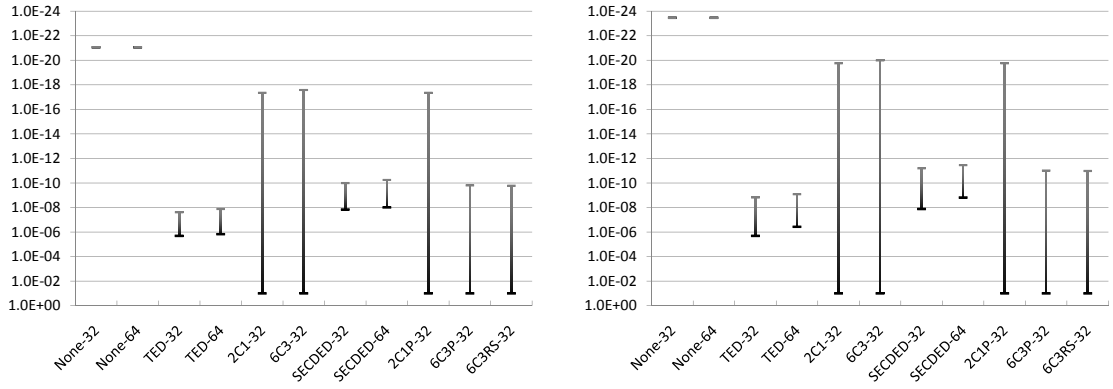
3.3.4 Mean Time Between Failure (MTBF) Analysis

In order to justify choosing one encoding scheme over another one must know both the microring fault rate and the rate of interfering vs. non-interfering faults. Since this information is not available, I have taken a different approach; I have determined the fault rate that microrings must attain in order to meet a particular MTBF for a single link, and also for an entire network (such as a photonic torus). These calculations can not only guide architects when choosing encoding schemes once microring resonators mature, but equally as important these results provide goals and targets for device researchers and manufacturers.

The MTBF for a link can be calculated given the fault rate, the rate that a fault is interfering/non-interfering, and simulation results from Section 3.3.3 which categorize received transmissions as either correct, incorrect, etc. Since the MTBF target is known, the fault rate can simply be varied until the target MTBF is achieved. I had to use great care when developing this MTBF solver, since the numbers involved differ by orders of magnitude – a naive approach could have easily lead to "catastrophic cancellations"[9] during the floating point calculations. The use of polynomial expansion and summation of terms using an algorithm that was proved in [79] to calculate the exact sum is just one example of the care that was taken when calculating the MTBF.

Figure 3.23 shows the required microring fault rate given a particular encoding and a desired MTBF of 1M hours. Figures 3.23(a) and 3.23(b) show the required fault rates for a single link and an 8-ary 2-cube Torus, respectively. The torus was assumed to

[9] According to Goldberg [27], "Catastrophic cancellation occurs when the operands are subject to rounding errors. For example in the quadratic formula, the expression b^2 - $4ac$ occurs. The quantities b^2 and $4ac$ are subject to rounding errors since they are the results of floating-point multiplications. ... When they are subtracted, cancellation can cause many of the accurate digits to disappear, leaving behind mainly digits contaminated by rounding error."



(a) 1M hr Link MTBF

(b) 1M hr Torus MTBF

Figure 3.23: Required Microring Fault Rate to Attain 1M hr MTBF for a Link (a) and a 8-ary 2-cube Torus (b) by Encoding Scheme

have direct links between the nodes (no microrings were assumed for routing). The Y-axis is the required fault rate that must be obtained; note that the values on the axis are such that a lower fault rate (meaning higher quality microrings) is placed higher on the axis.

The spread of the values is due to different rates of interfering faults. The values at the lower fault rates assume the probability of a fault being interfering is uniformly random – in other words, the probability of an interfering fault is equal to the number wavelengths minus 1 times the channel width in nm divided by the FSR in nm. The values requiring a higher fault rate, on the other hand, assume the resonance point of a microring will drift from the desired based on a normal distribution, centered at the desired resonance frequency (which yields a dramatically lower rate of interfering faults.)

The results show that the NcK encodings such as the 2C1-32 or the 6C3RS-32 are the best choice when fault rates are very high but the rate of interfering faults is very low. Hamming codes are best when the fault rates are moderately high, with TED winning out over SECDED in the absence of extended or permanent faults. In order for nanophotonic links/topologies to meet a 1M hour MTBF without using error detection or correction schemes, microrings will need to be fabricated such that fault rates are in the range of 10^{-21} to 10^{-24}/link cycle (the links were assumed to be modulated at 10GHz).

The conservative assumption I used in my simulations (that any undetected bit error will result in a failure) means these numbers are probably a little high, but it is un-

likely that the actual bit error rate that results in failure will change these results by very much (certainly not orders of magnitude.) Assuming the low fault rate cannot be attained, it is clear that some type of error detection scheme will be needed if large scale microring resonator-based networks are to become a reality, and that microring-based photonic networks that do not implement error detection or correction schemes will be inherently unreliable due to their low MTBF.

3.4 Related Work

This section describes the related work for the study presented in this chapter. This related work is separated into two sub-sections for the sake of clarity. Section 3.4.1 provides the related work for Section 3.2, while Section 3.4.2 is the related work for Section 3.3.

3.4.1 Trimming

The idea of using microring resonators for modulation in on-chip optical networks has been around for some time, although as stated in [52], "... the disadvantage of using resonators for modulation is the high temperature sensitivity of the device." However, as discussed before, this can be compensated for using trimming. In current literature, researchers typically estimate the required microring trimming power by multiplying the estimated average trimming power per microring by the number of microrings [1, 70, 40, 42]. A global estimate for microring heating was provided in [71], and I assume that a similar approach was used to derive the estimate.

In [92] Hewlett-Packard (HP) researchers describe a 64x64 WDM based crossbar (called Corona) for a 256-core Chip MultiProcessor (CMP) and the authors in [1] estimates that a total of ∼26W is necessary for trimming of the Corona network (which is ∼54% of the estimated∼48W total network power.) Cornell researchers described a bus-based scheme to connect clusters of processors in [43], and more recently propose a hybrid opto-electronic on-chip network called Phastlane [15] that uses a low complexity nanophotonic crossbar supported by an electrical network for buffering and arbitration. Neither [43] or [15] explicitly discuss the required power for trimming of the nanophotonic networks.

MIT and Berkeley researchers describe a multistage Clos network in [40], which uses a mixture of electronic routers that are connected by WDM based photonic links. A fixed thermal power was assumed to tune the microrings over a 20K range. The researchers also use a different set of constraints - they assume the microrings are 10μm in diameter, and place the rings on the same die as the cores. These two constraints lead to their conclusion that optical crossbar designs are impractical. However, microrings can be as small as 3μm and still function correctly, and if performance is important it is possible to implement the communication network on another level of a 3D design as illustrated in [92]. The Clos design reported in [40] uses only thermal trimming, which as discussed in Section 3.2.6 could be problematic.

The authors in [83, 82] propose a photonic 2D torus network that employs an electrical network for arbitration and flow control, but no estimate for trimming power of the nanophotonic network is explicitly discussed. Firefly [71] is another hybrid opto-electronic network proposal that uses an electrical network for intra-cluster communication and a nanophotonic crossbar for inter-cluster communication. A global estimate of 3.6W is assumed for microring heating in the Firefly network. The FlexiShare network crossbar [70] uses a token stream for arbitration and credit sharing, and a 1μW per ring per K with a 20K tuning range was assumed for trimming power.

These approaches to estimating trimming power are reasonable given the absence of a full integrated power/thermal simulator. And it could be argued that theoretically any microring based network will not require any trimming power if operated under ideal conditions. Temperature fluctuations in the environment external to the chip will occur in the real world, however, and it is vital to understand how temperature fluctuations affect the amount of power necessary to support trimming.

3.4.2 Photonic Link Resilience

In on-chip nanophotonic networks, the use of the ARQ protocol was proposed in [15]. The authors suggest the use of a SAW protocol designed to request retransmission in the case of a dropped packet. The Phastlane architecture presented in [15] uses a NAK to signal that the packet was dropped due to insufficient buffer space. If the NAK is

not received within a specific time window it is assumed that the packet was successfully buffered. As described earlier, it is clear that this scheme could be extended to provide reliable communication with the addition of error detection bits and the additional hardware to make the feedback channel reliable as is done in [25].

Our work proposes using error detection/correction in order to overcome the faults that will occur in microring resonators due to fabrication defects and thermal fluctuations. This is somewhat analogous to what is done in Razor [21], where a flip-flop is added to "ensure correct operation in the face of a number of environmental and process related variabilities". Razor is an example of on-line error detection and correction.

Recently in [100] the use of CRC was proposed for the Macrochip system. Many of the links in the Macrochip design are serial or pseudo-parallel[10] and have maximum packet payloads of 4KB, making CRC suitable for the application. The focus in [100] is on Bit Error Rate (BER), and is not concerned with the source of the bit errors.

Nanophotonic interconnects do have the potential for being unreliable, but electrical on-chip interconnects are not expected to be error free either - especially when implemented using very deep submicron technology [10, 16]. Fu and Ampadu in [25] investigate the use of a Type-II HARQ protocol for electrical interconnects. In this study Hamming product codes [75, 13] with Type-II HARQ are compared against Hamming, ARQ with CRC-5, Extended Hamming (SECDED), and Bose-Chaudhuri-Hocquenghem (BCH) for delay, area and power efficiency. (Reed-Solomon codes can be considered a non-binary BCH code). It is interesting to note that the authors chose not to implement the CRC-5 with a LFSR, opting for a more complex implementation. A GBN protocol was implemented for the ARQ portion of the work, and triple modular redundancy was implemented to protect the ACK/NAK feedback signal.

Fu and Ampadu also investigated the use of a dual-mode HARQ scheme in [24]. The proposed scheme uses a SECDED code for 64 bits, or four 16 bit SECDED codes in the case of the environment being very noisy. The use of Extended Hamming codes for both encoding techniques allows hardware sharing, which increases the area only slightly. The

[10]Pseudo-parallel links are very narrow (usually two or four bits wide) parallel links, that require many cycles to send a flit. This distinguishes them from serial links which are a single bit wide, and traditional parallel links which are typically are assumed to only require one or two cycles to send a flit.

proposed scheme yields up to a 35% energy improvement compared to previous solutions. However, it is unclear if the signal interleaving that is beneficial in the dual-mode work would be as beneficial in a WDM environment.

The authors in [20] investigate the energy efficiency and performance of ARQ, FEC, and HARQ in on-chip electrical networks. The ARQ scheme utilizes a CRC-8 that was implemented with a parallel bit code generator, while the FEC scheme analyzed uses overlapping parity bits. The results showed the trade-offs of performance and energy, and depending upon the environment (voltage swing, noise power, wire length, etc.) one scheme may be better than another.

3.5 Summary

The previous sections showed that large nanophotonic on-chip networks with microring counts in the hundreds of thousands will not be feasible without the aid of additional resilience techniques. Fortunately, the problems of thermal stability and fabrication errors can be overcome using trimming techniques such as Sliding Ring Window, through the addition of error detecting/correcting schemes, or a combination of both. Since the actual fault rate is not currently known, the rest of this dissertation will assume networks without error detecting/correcting bits – investigating the performance/power impact of various error detecting/correcting schemes on entire networks is beyond the scope of this work.

Chapter 4

Optical Network Topologies

The previous chapter showed that it will be possible to create large photonic on-chip networks with microring counts in the hundreds of thousands, if the correct trimming and resilience techniques are employed. Having established the feasibility of large photonic networks, in this chapter I explore the potential advantages of a Fully-Connected Optical Network (FCON). This topology is of interest because it could enable terascale multicore systems to fully realize their performance potential. In addition to evaluating the performance, I also look at the area, physical layout, and power requirements.

The rest of this chapter is organized as follows: Section 4.1 provides a detailed description of the baseline network, including why it was chosen. In Section 4.2 an overview of the FCON topology is presented, which is followed by a description of the experimental infrastructure that was used in the performance analysis in Section 4.3. The performance and power results are in Section 4.4 and 4.5 respectively, and a discussion of the related optical on-chip topology work appears in Section 4.6.

4.1 Crossbar Optical Network

In order to analyze and evaluate the various metrics (power, area, etc.) of an FCON, a suitable baseline must be chosen to compare it to. This was accomplished by creating the Crossbar Optical Network (CrON), a flat topology which has an identical link bandwidth to FCON. CrON is modeled closely after the Corona [92] design, primarily

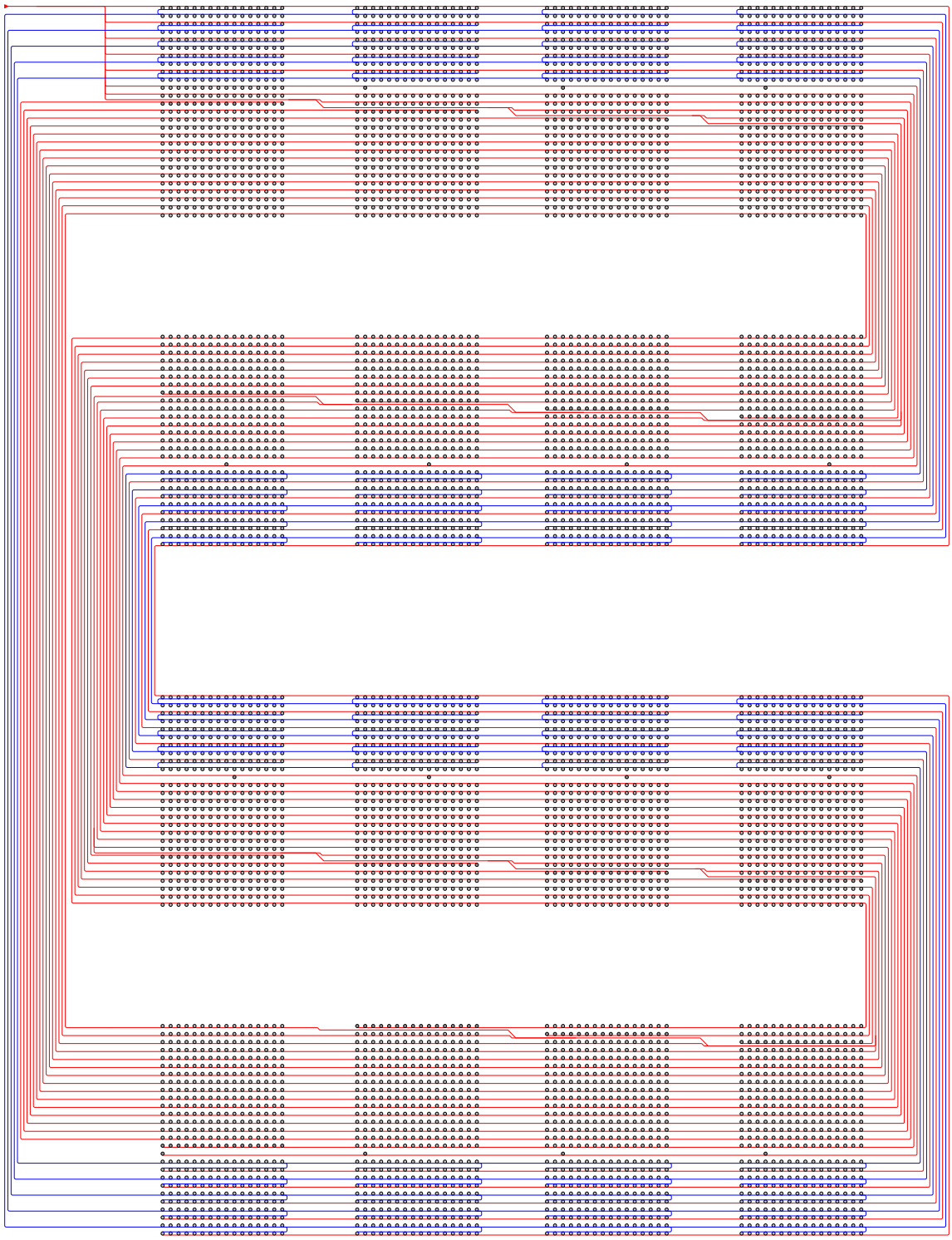


Figure 4.1: CrON Layout 16 Node 16-bit

because Corona has been carefully vetted and there are enough details publicly available to allow it to be modeled relatively accurately. This section describes CrON in detail and points out how it differs from Corona.

The Corona design is a 64 x 64 256 bit crossbar operating at 10GHz (double clocked at 5GHz) with four cores electrically connected to each network node. The waveguides in Corona are laid out in a serpentine fashion with 64 bits of data carried on each waveguide. The CrON configuration analyzed in this dissertation also assumes 64 nodes, but assumes each network node is connected to a single core and the bus width is 64 bits instead of 256. The decision to model a 64 instead of 256 bit data path was driven by the fact that CrON was being modeled as a 64 core system. Table 4.1 highlights the structural differences between Corona and CrON.

Figure 4.1 shows the entire layout of a 16 node 16-bit CrON – the 16 node version of CrON is shown for the sake of clarity (the 64 node 64-bit CrON would have four times as many waveguides in the serpentine, four times as many node groupings, and each node grouping would have four times as many microrings, which is too much detail to render legibly.) Figure 4.2 illustrates a subset of the transmitter and receiver section for a single CrON node.

Arbitration in CrON is handled in a manner similar to the Token Channel with Fast Forward described in [91]. Due to the nature of the protocol, a processor can wait up to 8 clock cycles (at 5GHz) to receive an uncontested token. Increases in die area and node count will increase the serpentine waveguide length and therefore increase propagation delay, meaning that the delay for uncontested tokens will grow with increased clocking speeds, die area, and node count. (The CrON design, however, does have the capability of a simultaneous one-to-many transmission if a single node were by chance to acquire arbitration tokens for multiple receivers.) The Token Channel with Fast Forward protocol was chosen over the Fair Slot protocol since a broadcast waveguide is required in order to support Fair Slot [91].

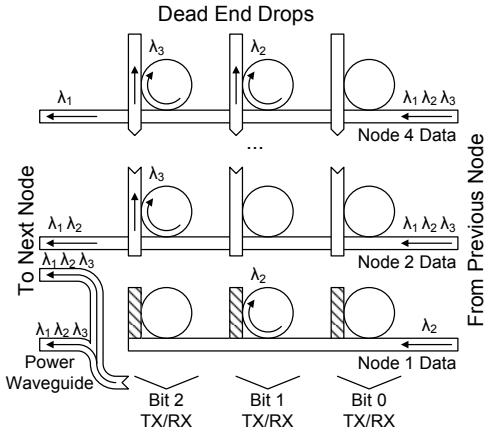


Figure 4.2: CrON Node 1 (Receives data 010, Transmits 011 to node 2 and 001 to node 4)

Table 4.1: Corona/CrON Network Parameters

			Microrings		Bandwidth		
Network	**Tech**	**WGs**	**Active**	**Passive**	**Total**	**Bisection**	**Link**
Corona	17nm	257	∼1M	∼16K	20TB/s	20TB/s	320GB/s
CrON	16nm	75	∼292K	∼4K	5TB/s	5TB/s	80GB/s

4.2 Fully Connected Optical Network (FCON)

Fully connected topologies are extremely desirable because they ease the burden of parallel programming and provide low-latency communication paths, which can be used to improve performance. Fully-connected *electrical* networks are infeasible because of their extraordinary wiring complexity – however, the ability of optics to use WDM means an FCON can be built using a single waveguide per node connection.

Considering the number of node connections (and hence the number of required waveguide crossings) and an assumed 0.1dB loss per intersection, a single layer implementation of FCON would suffer from too much loss to be practical. However, using multiple photonic layers and photonic vias (described in Section 2.2) the losses can be lowered to the point where an FCON is possible. It is important to do a more detailed evaluation of how FCON might actually be laid out, of course, since the number of waveguides needed in FCON grows quadratically with node count[1] – simply estimating the area necessary is not

[1]The number of waveguides increases as the square of the number of nodes, and is independent of the data width (as long as the number of wavelengths that can be bundled on a single waveguide is greater than or equal to the phit width).

sufficient. The layout is shown in Figure 4.3, which presents the entire layout for a 16 node FCON using a 16-bit bus. Assuming an 8μm ring pitch (3μm ring and 5μm ring spacing) and a 1.5μm waveguide pitch (0.5μm waveguide and 1μm waveguide spacing), the network as illustrated occupies an area of $\sim$1.12mm^2.

In Figure 4.3 there are four layers; green waveguides reside on one layer and connect node groups in the vertical direction, red waveguides reside on another layer and connect node groups in the horizontal, and the blue waveguides are on a third layer and connect nodes within a cluster of four. The purple waveguides are the photonic feeds – notice that the main feed enters on the center left of the network splitting out in an H-tree pattern until it reaches a node where it then fans out in a tree structure. It should be clear that a 64 node FCON could be constructed by clustering four groups of 16 nodes and interconnecting them in the same way four node clusters are interconnected in the 16 node case. Laying out an FCON network in this fashion requires that the number of layers grows as $log_2(N)$, though fewer layers could be used at a cost of more complicated waveguide routing. Given my assumed layout technique (which avoids routing waveguides through the middle of a node) a 64 node FCON will require $\sim$54.9mm^2; this is large, but not unreasonably so.

Having described the layout of FCON, it is time to explain in detail how it works. The receiver section for node 4 of a 4 node 3-bit FCON is shown in Figure 4.4(a). This figure shows data arriving at node 4 from each of the other 3 nodes – node 1 is sending a 010, node 2 a 101, and node 3 a 111. FCON transmits all data to a target node on a single waveguide, like CrON. An FCON transmitter section can be seen in Figure 4.4(b), which shows node 4 transmitting a binary 011 to node 1, a binary 100 to node 2, and a 000 to node 3. (A separate clock wavelength, not shown in this figure, is used to indicate to the receiver when the incoming waveguides contain meaningful information).

While it may seem at first that an FCON would require dramatically more microrings than a CrON, both are bound by an O($D \cdot N^2$) growth rate, where D is the data path width (or phit-size) and N is the number of nodes. FCON requires the same number of microrings to transmit data as CrON, although it does require additional microrings for reception. The additional microrings required by an FCON over a CrON are all passive, however, and will not require additional electrical power to function. CrON, on the other

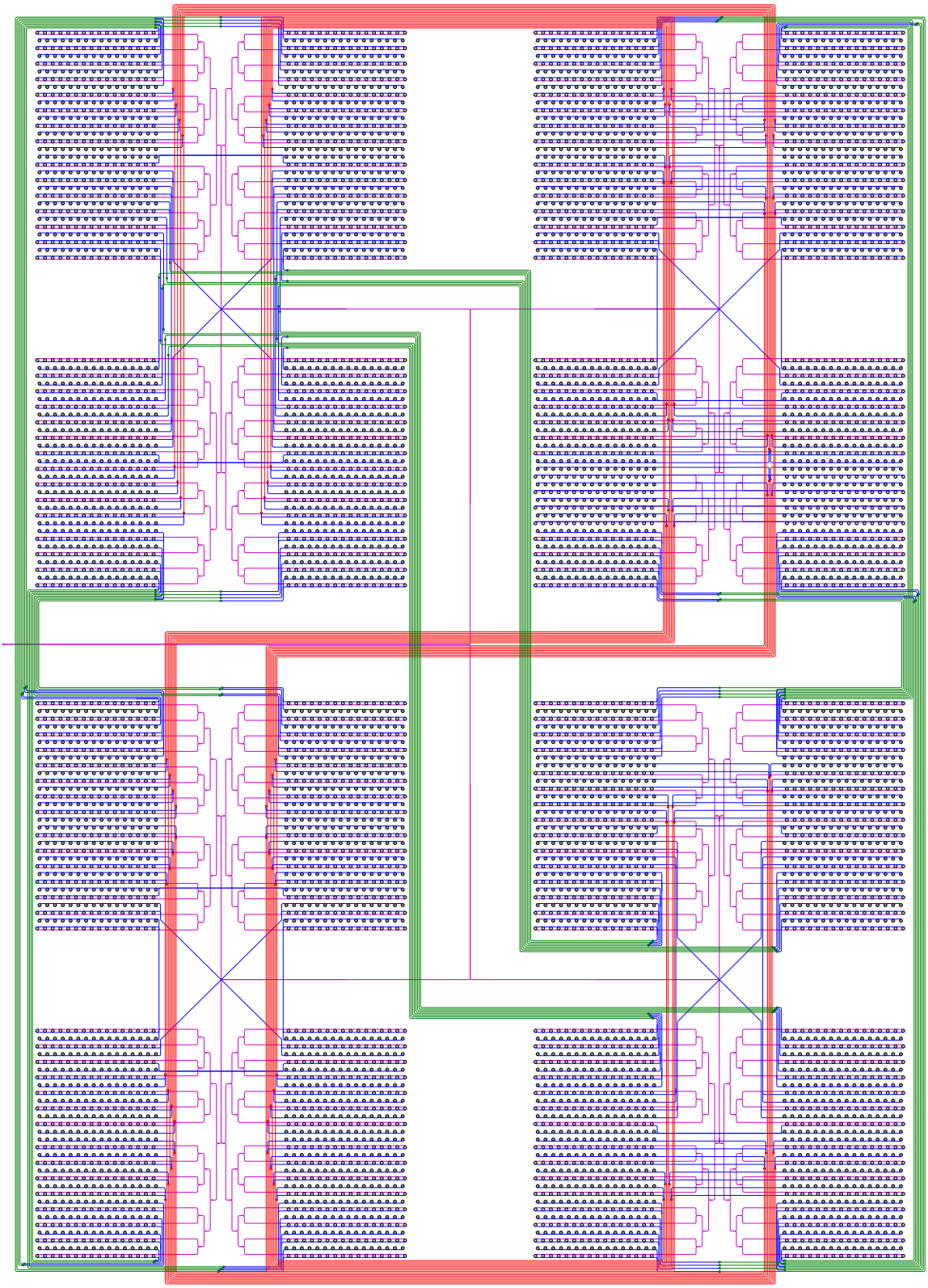
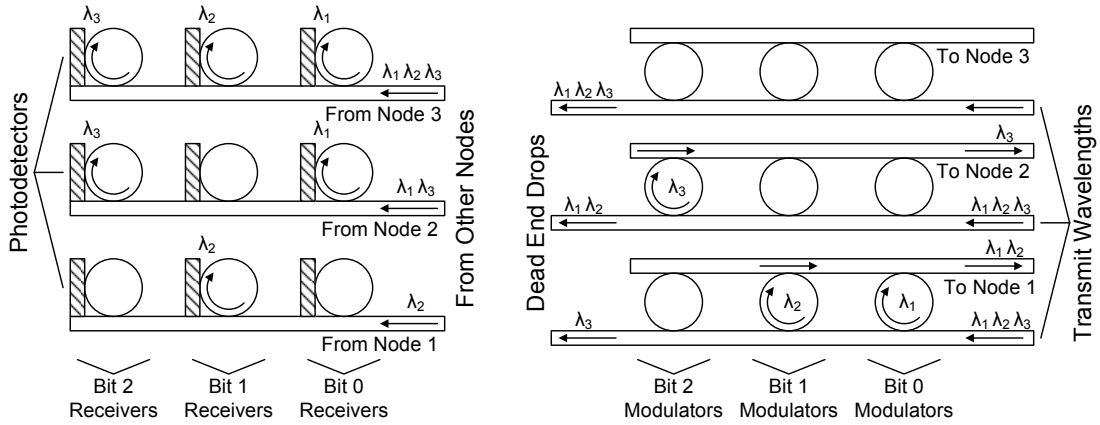


Figure 4.3: FCON Layout 16 Node 16-bit



(a) FCON RX Section for Node 4 (3-bit datapath)

(b) FCON TX Section for Node 4. Note the waveguide to Node 3 has all wavelengths blocked, signifying all zeros

Figure 4.4: Detailed FCON Node 4

hand, requires more active rings than FCON once all required rings (clocking, flow control and/or arbitration) have been included.

The major advantage of FCON over CrON, though, is that since each destination has dedicated receive hardware for every potential source, there is no shared resource that needs to be arbitrated for; furthermore, transmissions are asynchronous and (as mentioned above) are indicated by adding a clock wavelength per waveguide, removing the need for a global clock. Flow control will be necessary to prevent a node being overwhelmed with multiple messages every cycle, of course, and can be provided in a number of ways – the most straightforward is to include a flow bit by adding one more wavelength per link, which will exchange the control information of whether flow is enabled or disabled. This in essence increases the data width to $D + 1$, and requires a corresponding increase in microrings and laser power. Disabling flow is accomplished by turning off the flow bit to a given node.

There are a variety of ways the re-enabling of flow can be accomplished: all flows can be simultaneous re-enabled, flows can be re-enabled in a round robin fashion, or based on proximity, or most recently received, etc. In the experiments presented in Section 4.4 a round robin re-enabling scheme was employed because it is fair, and will not result in large bursts of traffic from all sources during the re-enabling stage.

Table 4.2: Example Trace

Packet #	Time Sent	Source	Destination
1	20	A	C
2	22	B	C
3	24	C	D
4	26	D	A

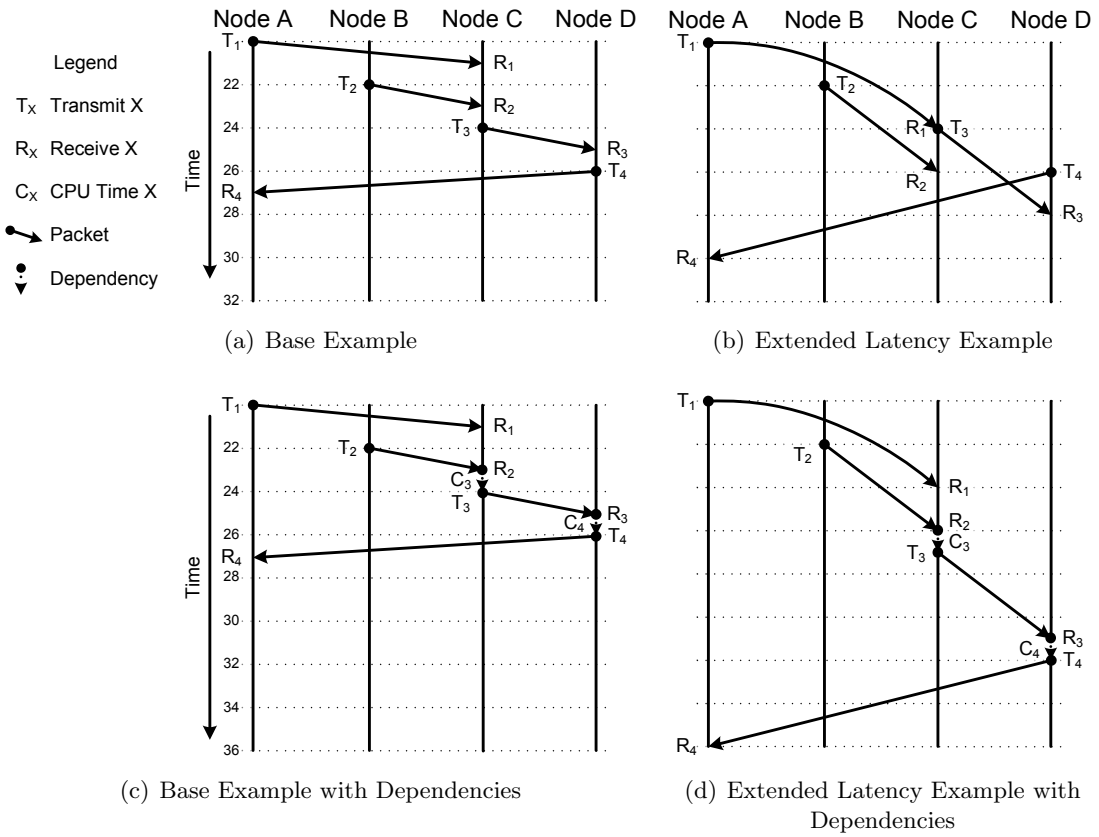
4.3 Experimental Infrastructure

The use of full system simulation is the most accurate way to perform a network/system analysis, but it is very slow and thus limits design space exploration. In order to overcome this problem researchers frequently use trace based simulation to study different network topologies and properties, which can be done much faster. Unfortunately, trace based simulation that does not include dependencies between packets can provide results that are misleading. The following subsection demonstrates the importance of including dependency information in traces.

4.3.1 Simple Example

In order to better understand the potential pitfalls of simulating networks using traces that do not include dependencies, consider the example trace shown in Table 4.2, which was obtained from a full-system simulation which used a network with a single cycle latency. If this trace is run on a network simulator which also has a latency of one cycle (see Space-Time Diagram in Figure 4.5(a)), the simulation will indicate that the program completes at time 27 (one cycle after packet 4 is sent). If this trace were to be simulated on a network which has a four cycle latency (see Figure 4.5(b)), then the program will complete at time 30 (four cycles after packet 4 is sent). As expected, a difference in completion time and also a change in average latency is seen.

But what if node C was actually gathering information from nodes A and B, calculating a sum, and then sending the result off to node D? And, what if the resulting sum was sent back to node A from node D? In this case there would be dependencies within the trace – packet 3 cannot be sent until both packets 1 and 2 arrive, for example, and



(a) Base Example

(b) Extended Latency Example

(c) Base Example with Dependencies

(d) Extended Latency Example with Dependencies

Figure 4.5: Example Space-Time Diagram (Without Dependencies (a), (b), and with Dependencies (c) and (d))

packet 4 cannot be sent until it receives packet 3. There will also be a minimum amount of "processing" time (in this case, to perform the addition) that must elapse between the reception of the last dependent packet and the transmission of the next one. Thus, in this example, if the processing time is one cycle then packet 3 can be sent one cycle after both packet 1 and 2 have been received, and packet 4 can be sent one cycle after it gets packet 3.

These dependencies do not explicitly appear in the full system trace generated on the single cycle latency network – packet 1 arrives at node C at time 21, packet 2 arrives at time 23, a single cycle is spent performing the addition, at time 24 node C transmits its value, and at time 27 the simulation completes (see Figure 4.5(c)). However, if this same program were run on a full-system simulator using a network with a four cycle latency one would see that packet 3 will not be sent until time 27 (since packet 2 is received at time 26),

which delays the reception of packet 3 until time 31. This in turn delays the transmission of packet 4 until time 32, and the completion time climbs to 36 (see Figure 4.5(d)). So the fact that there are dependencies in the trace that are not explicitly identified means the trace-based simulation will report a completion time that is artificially low.

4.3.2 Dependency Inference

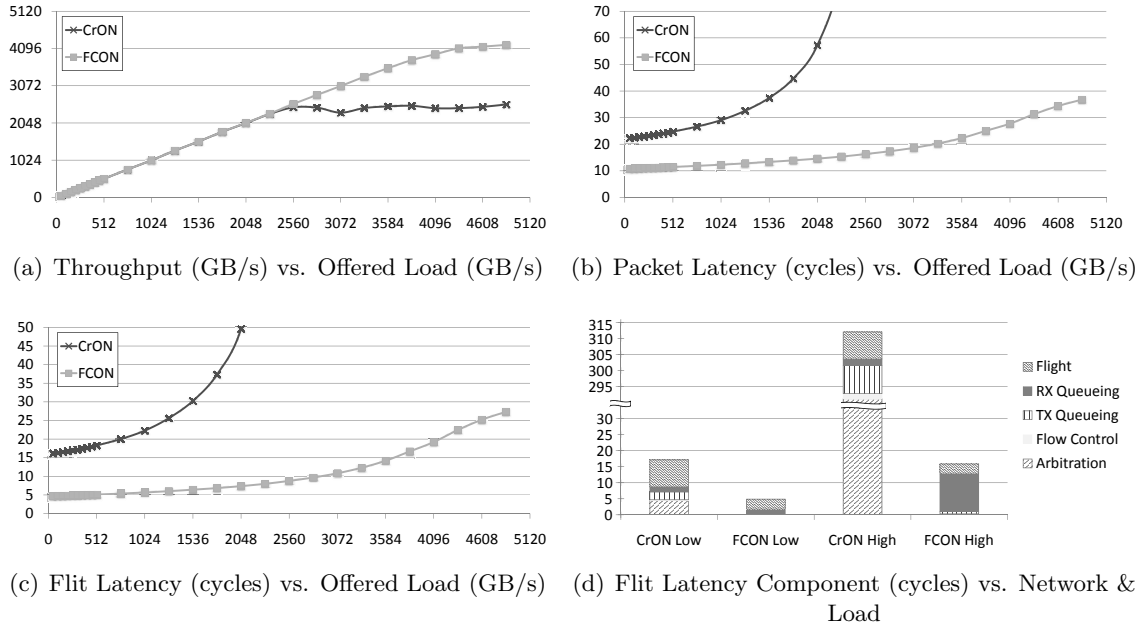
The simple example highlights the importance of including reception dependencies in traced based network simulation. In [68] we presented an inference-based technique for identifying and including packet dependencies, and showed that using our technique resulted in much better simulation accuracy without excessively extending simulation time.

4.4 FCON Performance

In order to evaluate the performance of FCON and CrON I created a trace-driven network performance simulator[2] capable of determining the latency, average and maximum queue depths, average and peak bandwidth, and total execution time. The FCON and CrON architectures I modeled were 64 node networks with a 64-bit data path between nodes, built using 16nm technology. The cores were assumed to operate at 5GHz and capable of generating and consuming one 128-bit flit per cycle. The on-chip network occupies an entire level of a 3D stacked processor design, with an area of 484mm^2. The limitation of one 128-bit flit per cycle was imposed to allow for a somewhat "fair" comparison between CrON and FCON. Although FCON does have the capability to transmit to every destination simultaneously, the ability was not fully evaluated (even in the form of multicasting) since the performance advantages of FCON can be demonstrated with unicast traffic alone.

The Packet Dependency Graphs (PDGs) used in the performance simulations were a combination of synthetic traffic patterns and a number of the Stanford ParalleL Applications for SHared-memory 2 (SPLASH-2) benchmarks. The synthetic traffic patterns chosen were Uniform Random, Negative Exponential Distribution (NED) [78], Hotspot, Tornado, Bit Inverse, Nearest Neighbor, and Transpose. All synthetic traces were run with a range of

[2]This dependency tracking simulator was the same one used in [68].



(a) Throughput (GB/s) vs. Offered Load (GB/s)

(b) Packet Latency (cycles) vs. Offered Load (GB/s)

(c) Flit Latency (cycles) vs. Offered Load (GB/s)

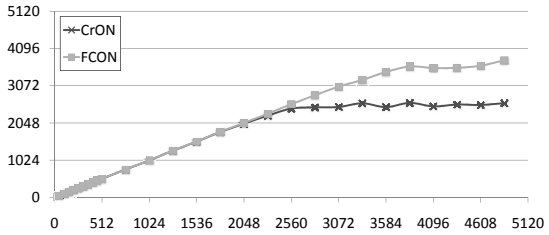
(d) Flit Latency Component (cycles) vs. Network & Load

Figure 4.6: FCON & CrON Throughput (a), Packet Latency (b) & Flit Latency (b) vs. Offered Load (GB/s) Flit Latency Components (d) vs. Network for Uniform Random Traffic Pattern

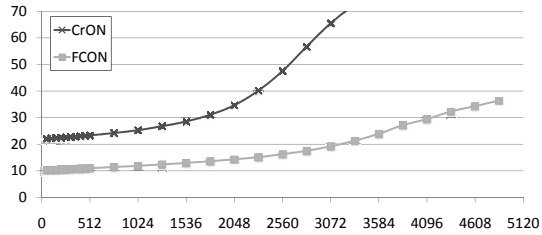
offered load[3] in order to determine maximum network throughput and average packet/flit latency – the synthetic traces do not include packet dependencies because the goal is to determine the points at which each network saturates, not to determine an execution time for a given trace (as is done for the SPLASH-2 benchmarks). The SPLASH-2 benchmark PDGs used were a 16 million point Fast Fourier Transform (FFT), Water with Spatial Data, Lower and Upper matrix decomposition (LU), Integer Radix Sort, and Raytrace Teapot. The PDGs were obtained from multiple 64 node full system simulations on Simics 3.0 [56] using the General Execution-driven Multiprocessor Simulator (GEMS) 2.1.1 framework [58] that includes the Garnet [73] network simulator; packet dependencies were then inferred using the algorithm outlined in [68].

The synthetic traffic PDGs provided an average offered load with an average packet size of 4 flits per packet, using a burst/lull distribution. The burst/lull injection distribution was chosen over a Bernoulli distribution since real traffic tends to be more "bursty" in nature. The throughput in GB/s is shown as a function of offered load in GB/s for FCON and CrON in Figures 4.6(a), 4.7(a), 4.8(a), 4.9(a), 4.10(a), 4.11(a), and 4.12(a) for Uniform

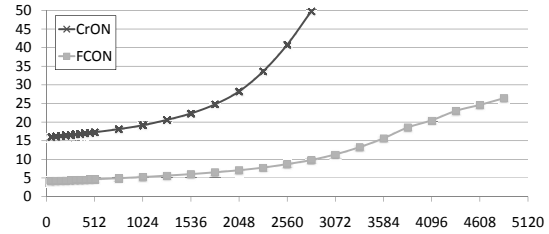
[3]The offered load is the total traffic throughput that is attempted to be injected into the network.



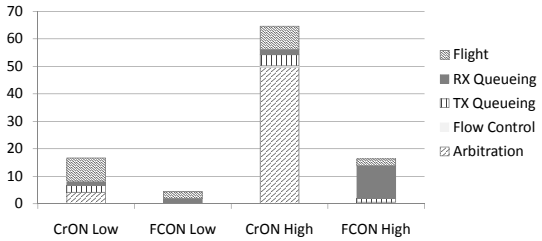
(a) Throughput (GB/s) vs. Offered Load (GB/s)



(b) Packet Latency (cycles) vs. Offered Load (GB/s)

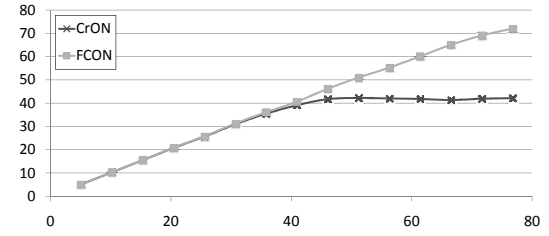


(c) Flit Latency (cycles) vs. Offered Load (GB/s)

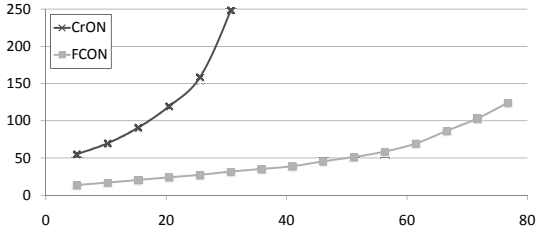


(d) Flit Latency Component (cycles) vs. Network & Load

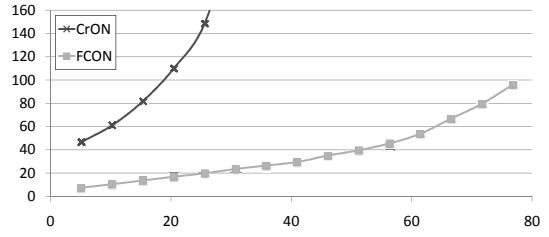
Figure 4.7: FCON & CrON Throughput (a), Packet Latency (b) & Flit Latency (b) vs. Offered Load (GB/s) & Flit Latency Components (d) vs. Network for NED Traffic Pattern



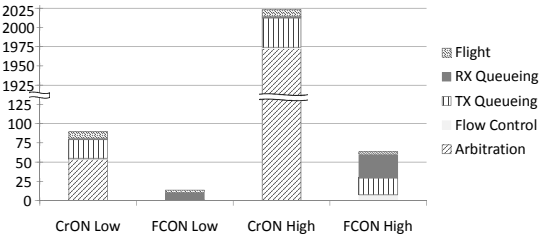
(a) Throughput (GB/s) vs. Offered Load (GB/s)



(b) Packet Latency (cycles) vs. Offered Load (GB/s)



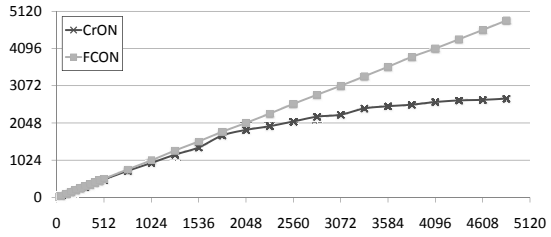
(c) Flit Latency (cycles) vs. Offered Load (GB/s)



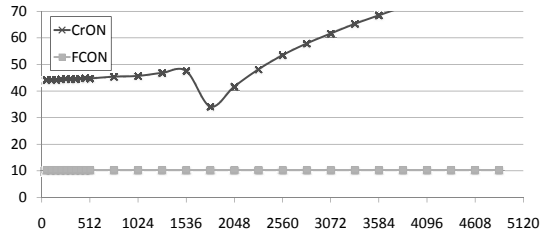
(d) Flit Latency Component (cycles) vs. Network & Load (Note break in y-axis)

Figure 4.8: FCON & CrON Throughput (a), Packet Latency (b) & Flit Latency (b) vs. Offered Load (GB/s) & Flit Latency Components (d) vs. Network for Hotspot Traffic Pattern

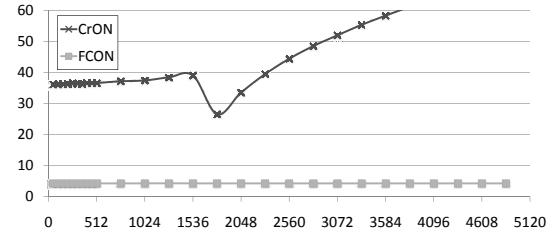
Random, NED, Hotspot, Tornado, Bit Inverse, Nearest Neighbor, and Transpose traffic patterns, respectively. These figures show that FCON provides higher throughput than CrON in terms of throughput on every one of the synthetic traffic patterns. Note that for the



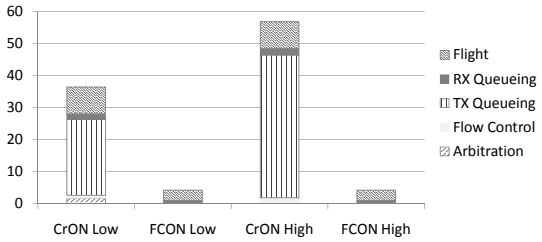
(a) Throughput (GB/s) vs. Offered Load (GB/s)



(b) Packet Latency (cycles) vs. Offered Load (GB/s)

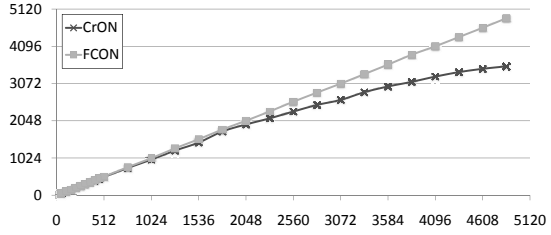


(c) Flit Latency (cycles) vs. Offered Load (GB/s)

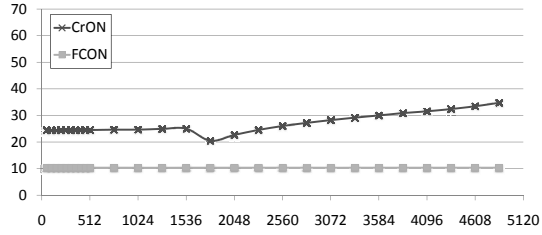


(d) Flit Latency Component (cycles) vs. Network & Load

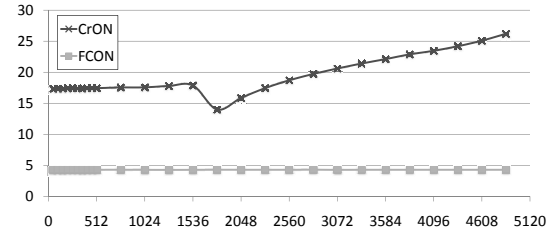
Figure 4.9: FCON & CrON Throughput (a), Packet Latency (b) & Flit Latency (b) vs. Offered Load (GB/s) & Flit Latency Components (d) vs. Network for Tornado Traffic Pattern



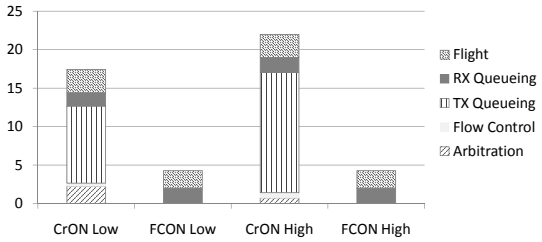
(a) Throughput (GB/s) vs. Offered Load (GB/s)



(b) Packet Latency (cycles) vs. Offered Load (GB/s)



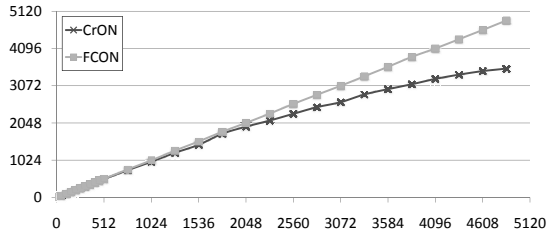
(c) Flit Latency (cycles) vs. Offered Load (GB/s)



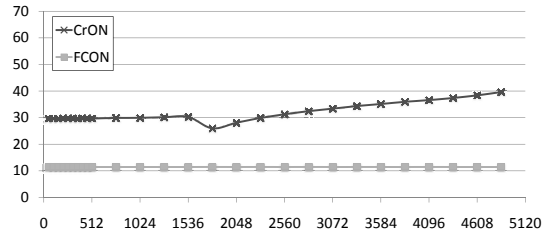
(d) Flit Latency Component (cycles) vs. Network & Load

Figure 4.10: FCON & CrON Throughput (a), Packet Latency (b) & Flit Latency (b) vs. Offered Load (GB/s) & Flit Latency Components (d) vs. Network for Nearest Neighbor Traffic Pattern

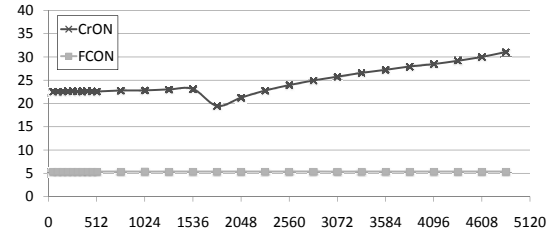
Hotspot traffic pattern the offered load is limited to 80GB/s – this is because the maximum throughput of a single node is 80GB/s and any offered load above that is guaranteed to overwhelm any network, regardless of topology. The packet and flit latencies in cycles are



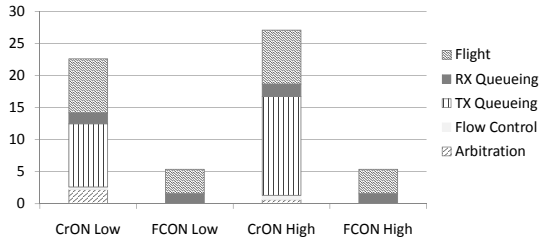
(a) Throughput (GB/s) vs. Offered Load (GB/s)



(b) Packet Latency (cycles) vs. Offered Load (GB/s)

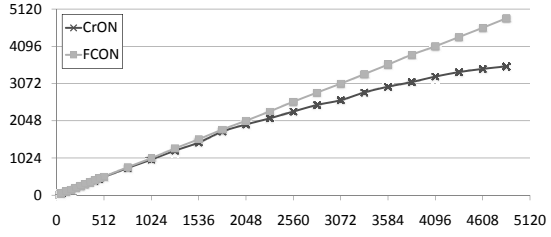


(c) Flit Latency (cycles) vs. Offered Load (GB/s)

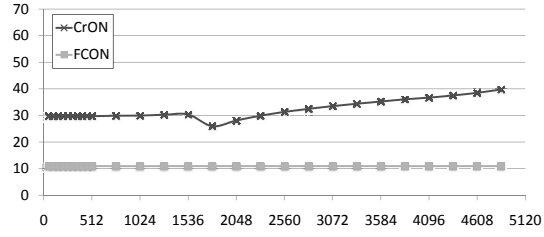


(d) Flit Latency Component (cycles) vs. Network & Load

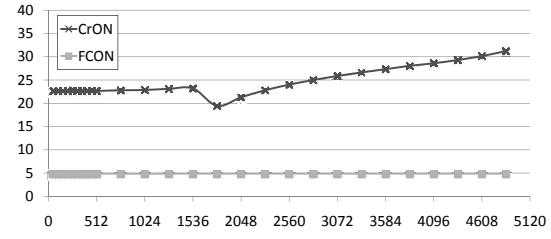
Figure 4.11: FCON & CrON Throughput (a), Packet Latency (b) & Flit Latency (b) vs. Offered Load (GB/s) & Flit Latency Components (d) vs. Network for Bit Inverse Traffic Pattern



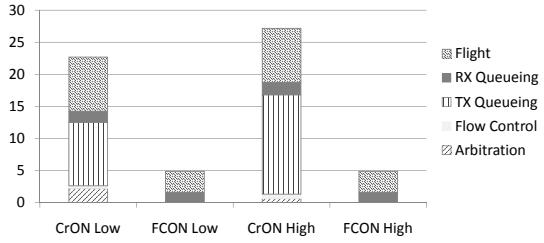
(a) Throughput (GB/s) vs. Offered Load (GB/s)



(b) Packet Latency (cycles) vs. Offered Load (GB/s)



(c) Flit Latency (cycles) vs. Offered Load (GB/s)



(d) Flit Latency Component (cycles) vs. Network & Load

Figure 4.12: FCON & CrON Throughput (a), Packet Latency (b) & Flit Latency (b) vs. Offered Load (GB/s) & Flit Latency Components (d) vs. Network for Transpose Traffic Pattern

shown as a function of offered load in GB/s in subfigures (b) and (c) of Figures 4.6 through 4.12 for the synthetic traffic patterns. Again, these figures show that FCON outperforms CrON by providing dramatically lower packet and flit latency on all of the synthetic traffic

patterns.

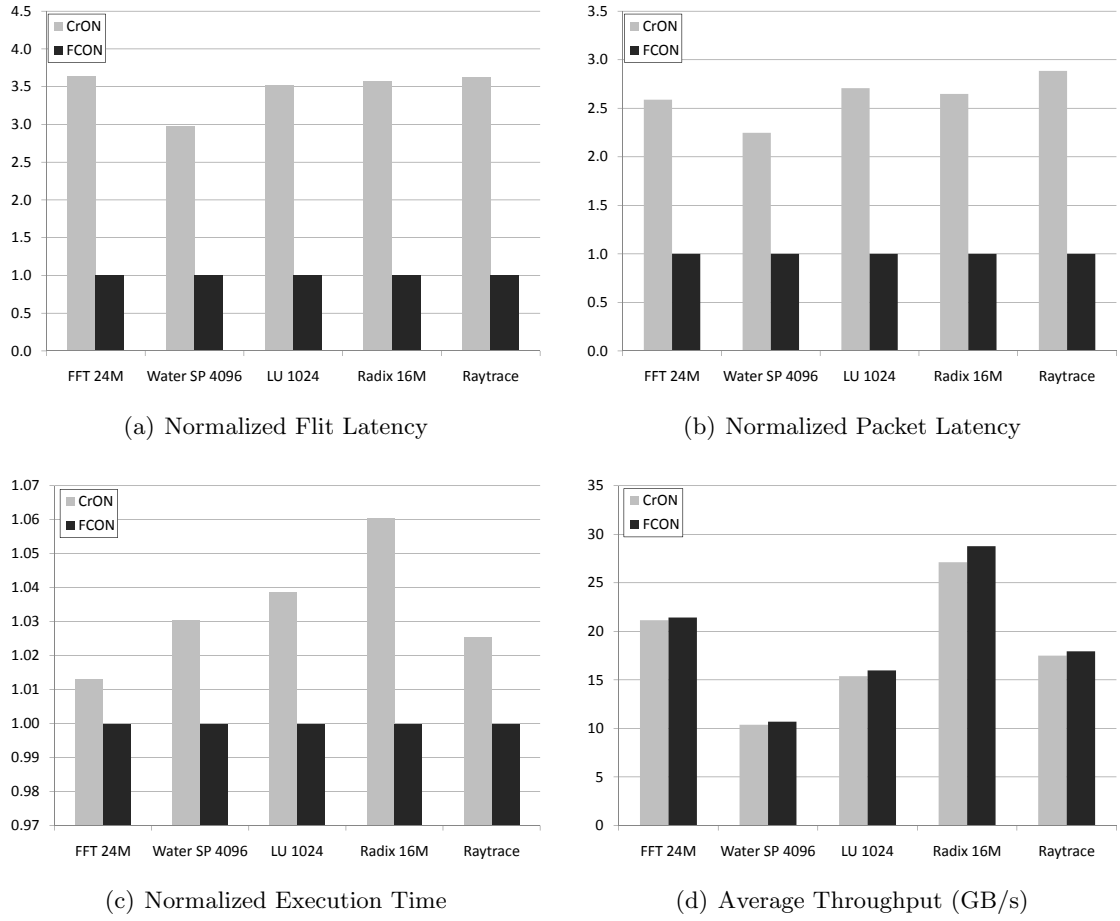
The breakdown of the latency components in cycles for FCON and CrON under low load (< 10% maximum offered load) and high load (> 45% maximum offered load) are shown in Figures 4.6(d), 4.7(d), 4.8(d), 4.9(d), 4.10(d), 4.11(d), and 4.12(d) for Uniform Random, NED, Hotspot, Tornado, Bit Inverse, Nearest Neighbor, and Transpose traffic patterns, respectively. Notice that more time is required for arbitration in CrON than is required for the entire flit transmission in FCON on the Uniform Random, NED, and Hotspot traffic patterns. These results show that arbitration is a limiting factor in the performance of CrON.

From the graphs it appears that FCON exhibits ideal performance on all traffic patterns except for NED and uniform random. In reality, the performance of FCON is slightly lower than the ideal starting at 56GB/s for Hotspot as well. The performance of FCON does match the ideal for Tornado, Nearest Neighbor, Transpose, Bit Inverse, and any other synthetic traffic pattern where each destination can only receive from a single source. This holds because FCON does not require arbitration in order to send a flit, and it is not possible for a single source to trigger the need to disable flow.

Another interesting trend to note is that in Figures 4.9(b), 4.10(b), 4.11(b), and 4.12(b) the latency for CrON actually drops at one point before climbing again. This is also seen in flit latencies in Figures 4.9(c), 4.10(c), 4.11(c), and 4.12(c). This is due to the nature of the arbitration scheme – under low load arbitration on average takes half the time for the token to complete a lap around the serpentine[4] (since only one source will ever request arbitration for a given destination). As the offered load climbs, the probability increases that arbitration has already been granted when a packet is injected (this is what causes the dip in average packet latency). This result is shown in Figures 4.9(d), 4.10(d), 4.11(d), and 4.12(d) where the number of cycles required for arbitration in CrON is lower under high load than under low load.

The performance results of the SPLASH-2 runs are shown in Figure 4.13. Figures 4.13(a) and 4.13(b) show the average flit and packet latencies for FCON and CrON, normalized to the network with the lowest latency (in all cases FCON). The figures show

[4]The serpentine layout for CrON can be seen in Figure 4.1.



(a) Normalized Flit Latency

(b) Normalized Packet Latency

(c) Normalized Execution Time

(d) Average Throughput (GB/s)

Figure 4.13: SPLASH-2 Performance Results

that FCON has dramatically lower average latencies across all the benchmarks; however, the lower latency does not result in a dramatic difference in the overall execution time.

Figure 4.13(c) shows the execution time of each benchmark normalized to the shortest execution time, and the figure shows that FCON executed the benchmarks from 1.3% to 6.0% faster than CrON. Reducing the latency by a factor of two results in such a small performance increase because the network requirements of the benchmarks are quite low.

Figure 4.13(d) shows the average throughput in GB/s for the various benchmarks. The average throughput of the SPLASH-2 benchmarks equates to ∼0.4% of the total network bandwidth *allowed* (FCON has the capability of ∼320TB/s; therefore, the benchmarks actually equate to ∼0.006% of FCON total network bandwidth), leading one to question

why anybody would want to build a network like this. While it may at first appear that the networks are over-designed, it is important to note that the average of the peak throughputs attained on the benchmarks was 1,295GB/s for CrON and 5,120GB/s for FCON. This means that at some point during the execution on FCON the maximum allowed network throughput was obtained on every benchmark, and there are critical points at which all the allowable network bandwidth is utilized. It is possible that if the PDGs contained multicast packets that the peak throughput of FCON would have been even higher. One must be careful not to unwisely restrict the flexibility of tomorrow's on-chip processor network based on the results of running yesterday's parallel processing benchmarks. This point cannot be overemphasized.

4.5 FCON Power

Using the FCON layout configuration described in Section 4.2, the FCON on-chip link losses were calculated to be 8.97dB, requiring approximately one seventh the power of the CrON design per link (CrON has a link loss of 17.3dB). The reduction in attenuation is due primarily to the reduction in the number of off-resonance microrings through which the light must travel. CrON requires that light travel through $D \cdot N - 1$ off-resonance rings, where FCON requires that light travel through $2D - 1$ (again D is the data path width and N is the number of nodes). The linear correlation of attenuation with node count in CrON results in an exponential correlation of link power, leading me to believe that CrON will not scale beyond 64 nodes. Worst case waveguide lengths are also shortened in FCON, since a given waveguide does not need to contact each node in the network as it does in CrON – it only needs to contact the transmitter and receiver nodes involved.

The maximum power consumed under high load was 15.5W and 106.1W for CrON and FCON, respectively. Given the previous statements that FCON requires one seventh the link power of CrON, one might expect that FCON would consume less power overall – however, FCON has 63 times $(N - 1)$ as many links as CrON, and the laser has to be sized to support the potential of simultaneous communication on all of them. This leads to the almost seven-fold increase in overall power consumption.

The simulations also showed energy efficiencies approaching 652fJ/b and 3.04pJ/b for CrON and FCON, respectively. FCON is able to simultaneously transmit to every other node without arbitration, and when operated under low workload (as is done in these simulations) FCON is far less power efficient than CrON. This is not the whole story, however – since there is a fixed amount of energy entering the network and the efficiency is calculated by dividing the energy used by the amount of information transmitted, the energy efficiency of optical networks is highly dependent upon workload. The energy efficiency of FCON under a 100% workload approaches 77fJ/b, which is over a factor of 8 lower than CrON under 100% load (although it is unclear if there would ever be a time in a real system when every node would be simultaneously transmitting to every other node).

The extreme amount of photonic power required for FCON is not justified by the performance improvements observed. The ability to have energy efficient links without the need for arbitration will only be useful if the total power (especially the static portion) can be reduced. In the next chapter I present my investigation into ways to accomplish this goal.

4.6 Related Work

In [92] HP researchers describe a 64x64 WDM based crossbar (called Corona) for a 256-core CMP. Corona uses a multiple-writer single reader crossbar architecture, which requires arbitration (realized using a distributed scheme and additional optical channels). Cornell researchers described a bus-based scheme to connect clusters of processors in [43], and more recently propose a hybrid opto-electronic on-chip network called Phastlane that uses a low complexity nanophotonic crossbar supported by an electrical network for buffering and arbitration. Phastlane uses packets with a single flit and an ARQ based flow control scheme, where packets are allowed to be dropped.

MIT and Berkley researchers [40] propose a multistage Clos network using a mixture of electronic routers that are connected by WDM based photonic links. Clearly, this network has less flexibility and a higher average hop-count than a crossbar. Furthermore, the CMXBar described in the paper requires arbitration, which FCON does not. The au-

thors in [83, 82] propose a photonic 2D torus network that employs an electrical network for arbitration and flow control. The network is evaluated on a variety of synthetic and scientific benchmarks [32] to show that the hybrid photonic torus network can achieve a factor of 37x improvement in performance per energy spent. This paper also points out that many scientific workloads exhibit communication patterns that change over time, which is another reason the fully connected nature of FCON is so attractive.

Firefly [71] is another hybrid opto-electronic network proposal that uses an electrical network for intra-cluster communication and a nanophotonic crossbar for inter-cluster communication. The Single Writer Multiple Reader (SWMR) network discussed in [71] requires a broadcast network in order to send the head flit, and this broadcast network will require arbitration - the timing between the sending of the head flit and transmitting the data flits will also require precise delay. In addition, the broadcast network will require power, which is likely to be nearly equal to that of the SWMR crossbar itself.

The FlexiShare network is a flexible photonic crossbar [70] that is a combination of a Multiple Writer Single Reader (MWSR) and a SWMR design. The FlexiShare network decouples the number of communication channels from the number of nodes, in an attempt to reduce the required photonic power. FlexiShare implements a token stream for arbitration and credit sharing, adopting the reservation assisted scheme from Firefly.

Sun Labs/Oracle researchers [46] recently investigated using silicon photonics for the interconnection network of a multi-chip system or "Macrochip". They analyzed three different photonic networks in the multi-die system that used mirrors to couple light between dies, and concluded that a statically routed point-to-point network outperformed the other networks analyzed. The point-to-point networks analyzed in [46] were limited to 2-bit site-to-site connections, which the authors admit "is a potential performance limiter". The inter-layer coupler assumed in [46] differs from our photonic vias in that the inter-layer coupler connects signals between two dies, where our photonic via couples between layers of the same die.

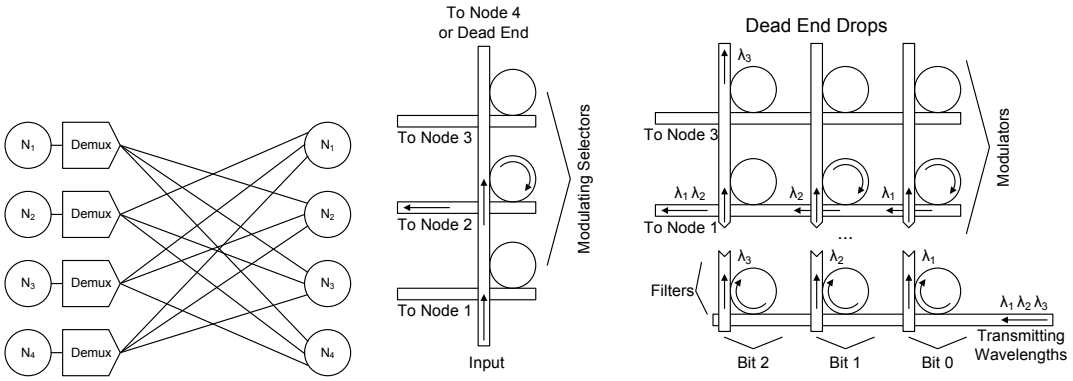
Chapter 5

Directly Connected Arbitration Free Network

As shown in the previous chapter, the FCON design has 63 and 32 times the total bandwidth and bisectional bandwidth of CrON, respectively. Providing this extraordinary bandwidth requires a large amount of power (most of it photonic), and since the bandwidth is underutilized FCON has a lower average energy efficiency than CrON. In my simulations I had only allowed a single flit per cycle to be injected into the network, so I decided to explore what would happen if FCON was modified such that it could only transmit to a single destination at a time. The investigation led to the development of the family of Directly-Connected Arbitration Free (DCAF) networks, which will be described in detail in this chapter.

5.1 DCAF Topology

Like FCON, the DCAF design features waveguides which directly connect each source/destination pair, creating a fully-connected backbone; however, DCAF incorporates additional microring resonators in the transmitter section of each node which are used to limit the number of destination nodes (denoted by k) that can simultaneously have information sent to them. If k is 1, DCAF is in essence a many-to-one crossbar – a single node can simultaneously receive from multiple sources, but can send to only one. This



(a) Arbitration Free Crossbar (Demux ensures that a node can send to only one other node at a given time.)

(b) 1:4 Optical Demux

(c) TX Section for Node 4

Figure 5.1: DCAF: 4 Node Network Equivalent (a), 1:4 Optical Demultiplexer (b), and Transmit Section (c)

differs from the SWMR crossbar described in [71], which requires the target receiver enable its receiving microrings at the correct time in order to receive the flit and prevent "stealing" of other flits destined to other nodes. DCAF does not require this coordination since there is dedicated hardware for each source/destination pair.

Figure 5.1(a) shows the equivalent network connectivity for a four node DCAF with k equal to 1. Since the dedicated links make it possible for each node to receive messages from all other nodes simultaneously, no arbitration is required. DCAF essentially has a locally controlled demultiplexer in its transmit section, while CrON has the equivalent of a receive multiplexer which must be globally arbitrated. Figure 5.1(b) is an example of how a 1:4 optical demultiplexer can be constructed using microring resonators. Figure 5.1(c) illustrates the DCAF transmitter section – in this figure λ_1 and λ_2 are being transmitted to node 2, while λ_3 is not (in other words, node 4 is transmitting a binary 011 to node 2).

Figure 5.1(c) shows that the DCAF design is not technically limited to transmitting to a single receiver at a time; the actual limitation is that each individual wavelength can only go to one receiver. It would be possible to send wavelength 3 to either node 1 or 3 (but not both), for example. This limit of 1 wavelength per receiver effectively prevents bus-width sized messages from being transmitted to multiple receivers, since one of the wavelengths is assigned to be the clock wavelength and can only go to a single receiver at

a time. It would be possible to add a clock wavelength for every grouping (e.g. 8 bits) in order to send smaller transmissions to multiple destinations, but still support full sized transmissions to a single destination – if a clock wavelength is used for each grouping of data bits, then the limit of simultaneous transmissions is bound by the granularity of the grouping.

For k to be greater than 1 there must be multiple sets of power waveguides and filtering rings. Each node in such a design would be limited to a single transmission per receiver group, but would be capable of k simultaneous transmissions. DCAF designs where ($1 < k < N$-1) result in a network that lies somewhere between a crossbar and FCON – a detailed investigation of these types of configurations is beyond the scope of this work, although the potential is intriguing and warrants further study in the future.

DCAF does not require arbitration in order to transmit a flit, and therefore it will not be subject to the limitations imposed by systems which require global clock synchronization. However, even though DCAF is arbitration-free, it does require flow control. It is not feasible to add another signal per potential source for flow control, as was done in FCON, because doing so would double the required photonic power of a 64-bit 64 node DCAF when k is 1. Instead, DCAF uses an ACK based ARQ scheme for flow control. If a flit arrives at a node and there is no available space in the buffer, the flit is dropped and the ACK is not sent back. A Go-Back-N (GBN) ARQ scheme was chosen over a conventional credit based flow control approach since multiple flits can be in flight simultaneously on a single waveguide – or, to put it another way, the round trip of a single link can be much greater than 2 cycles. The ARQ scheme allows for efficient flow control without the need for excessive buffering. As discussed in Section 3.3 increasing the reliability of communication is another benefit of using an ARQ scheme for flow control, since lost or potentially corrupted flits can be retransmitted.

In order to support uninterrupted flow in a GBN ACK based protocol, the size of the sequence number of each flit must be able to accommodate the maximum number of outstanding flits. In the 64 node DCAF the sequence number was chosen to be 5 bits in size, which is large enough to account for worst case round trip propagation delay. It should be noted that the 5 bit sequence number per flit is not additional overhead that DCAF will

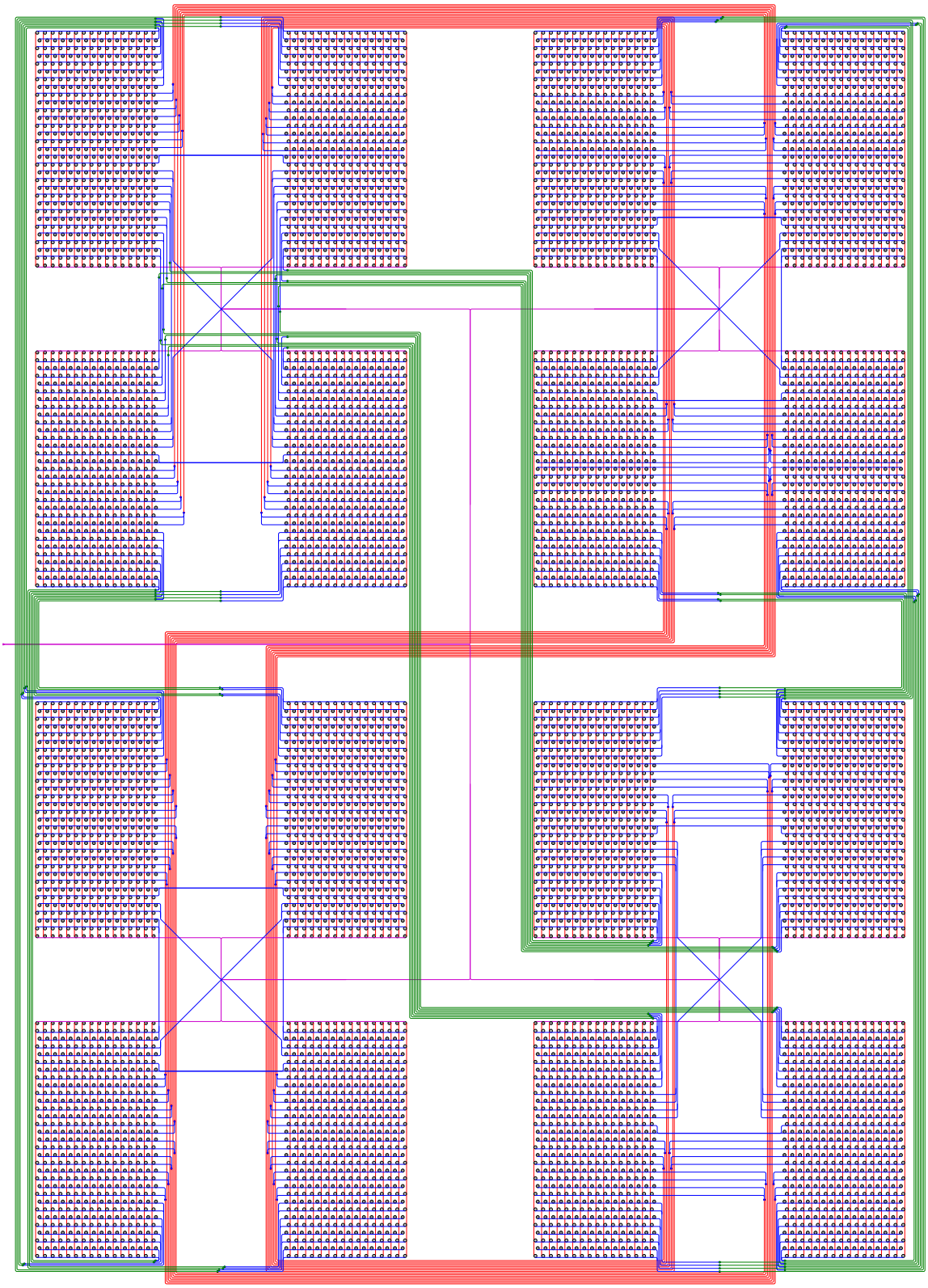


Figure 5.2: DCAF Layout 16 Node 16-bit

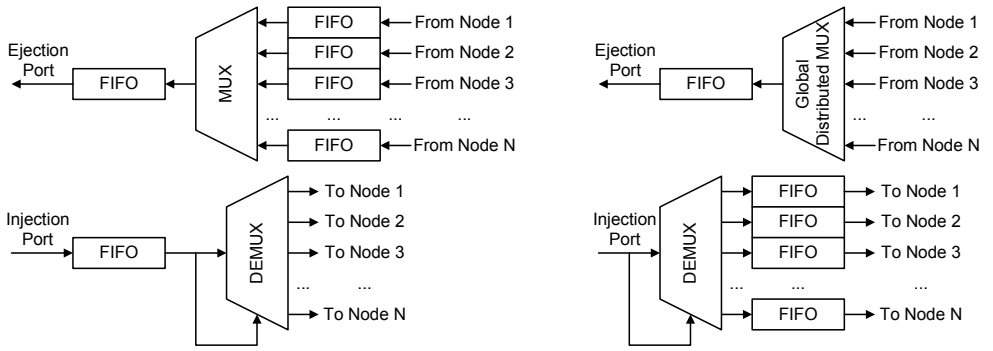
Table 5.1: CrON/DCAF Network Parameters

Network	Tech	WGs	Microrings		Bandwidth		
			Active	Passive	Total	Bisection	Link
CrON	16nm	75	∼292K	∼4K	5TB/s	5TB/s	80GB/s
DCAF	16nm	∼4K	∼276K	∼280K	5TB/s	5TB/s	80GB/s

incur when compared to CrON – CrON requires 6 bits to designate the flit source, which DCAF does not need to provide since DCAF has a dedicated receiver for each source.

As in Figure 4.3, Figure 5.2 presents the entire layout for a 16 node DCAF using a 16-bit data path. The same values for ring and waveguide pitch are assumed – 8μm ring pitch (3μm ring and 5μm ring spacing), and a 1.5μm waveguide pitch (0.5μm waveguide and 1μm waveguide spacing), but the network takes slightly more area (∼1.15mm^2). In Figure 5.2 each color of waveguide designates a different layer; green waveguides connect node groups in the vertical direction, while red waveguides connect node groups in the horizontal. The purple waveguides are the photonic feeds – notice that the main feed enters on the center left of the network splitting in an H-tree pattern until it reaches the node, but unlike FCON the feed does not need to fan out into a tree structure, it just needs to connect to the k sets of filter rings. Another difference between FCON and DCAF is the set of vertical red waveguides that connect the filter rings to the modulators within each DCAF node. A 64 node DCAF would be constructed in the same fashion as a 64 node FCON, and if k=1 would take 5% more space because of the microrings required for filtering and the ACK.

Table 5.1 illustrates the structural differences between CrON and DCAF. Note that the number of waveguides in CrON is somewhat misleading – if one considers a single loop around the chip as just one waveguide, then the number is 75; however, if you consider each segment between nodes to be a separate waveguide then there are actually ∼4.6K, which is more than is used by DCAF. DCAF also requires ∼88% more microrings than CrON, although there are in fact fewer *active* (power-consuming) microrings required in DCAF than in CrON.



(a) DCAF Node FIFO Configuration (b) CrON Node FIFO Configuration

Figure 5.3: Assumed Node TX and RX FIFO Configuration for DCAF (a) and CrON (b)

5.1.1 Buffering Analysis

The amount and configuration of network buffering is an important factor in analyzing the performance and power consumption of on-chip networks. The amount of transmit and receive buffering (in the form of First In First Outs (FIFOs)) at a given node alone is not enough to determine the power/performance of the network, however – for example, one cannot assume shared buffering for all transmitters at a node in CrON, since multiple flits can be simultaneously transmitted. For the buffers to be shared, one must also include an electrical crossbar to connect the buffers to the transmitters. The same is true on the receive side in DCAF – sharing the receive buffer requires a crossbar to connect the receivers to the shared buffer. These local crossbars require N-1 input and output ports, and including the power consumed by these crossbars is necessary when trying to evaluate the actual power used by photonics.

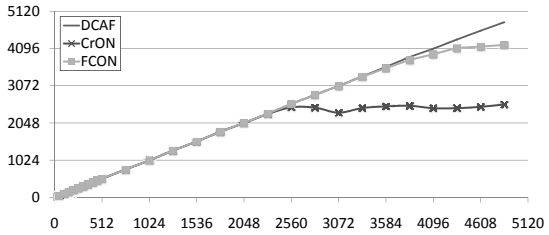
It is possible for DCAF to use a smaller local crossbar, with N-1 input ports and less than N-1 output ports, which would allow the same number of flits as output ports to be simultaneously transferred from the private buffers to the shared buffer. This approach cannot be used in the transmit side of CrON, though, since flits must be sent sequentially once arbitration has been obtained. (DCAF can drop an incoming flit if the private buffers are full.) In my analysis I assume DCAF uses a small shared receive buffer, connected to the N-1 private receive buffers. Figure 5.3 shows the assumed FIFO configurations at each node for both DCAF and CrON.

In CrON I assume each node has a shared receive buffer, since there is only one receiver per node. The amount of buffering must match the token size, so in order to avoid wasting photonic power the receive buffer size was chosen to be 16 flits since it evenly divides into the 64 wavelengths – this was also the approach used in [91]. DCAF does not require a private buffer for each transmitter, since only k simultaneous transmissions are possible. I assume a single shared transmit buffer for DCAF since k is 1; the shared buffer was chosen to be 32 flits since it works well with the ARQ scheme chosen. The small shared receive buffer also stores 32 flits, to match the size of the transmit buffer.

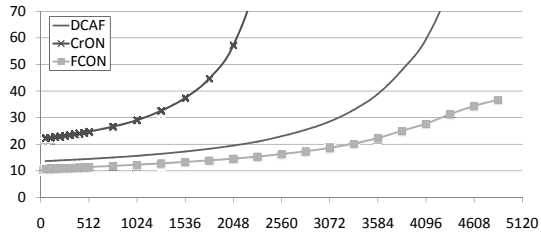
In order to determine the optimal amount of buffering for CrON and DCAF, the throughput of the networks with various buffering configurations was compared to that of an equivalent network with infinitely large buffers. The NED traffic pattern was used because its behavior is similar to real traces. The results of the buffering analysis showed that CrON had degraded throughput when only 4 flit buffers were employed, and had no loss in throughput when 8 flit buffers per transmitter were available. The performance of DCAF was diminished when only 2 flit buffers were used (even assuming a 2-output port local crossbar), but using 4 flit buffers per receiver resulted in maximal throughput for the topology. Thus, the performance and power results presented in the remainder of this work assume 8 flit buffers per transmitter and 16 flit buffers per receiver for CrON, and 32 flit transmit buffers, 4 flit receive buffers and a 32 flit shared receive buffer for DCAF. This results in a total of 520 and 316 flit buffers per node for CrON and DCAF, respectively.

5.2 DCAF Performance

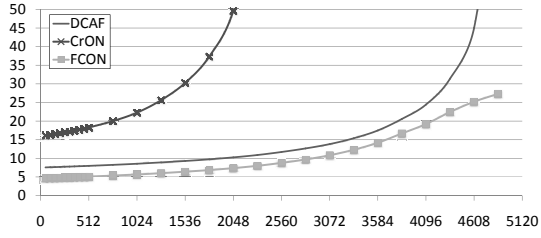
The same set of synthetic traffic PDGs were used for the DCAF/CrON/FCON comparison as were done in Section 4.4 when evaluating FCON. The throughput in GB/s is shown as a function of offered load in GB/s for DCAF, CrON and FCON in subfigure (a) of Figures 5.4 through 5.10 for all the synthetic traffic patterns. The packet and flit latencies in cycles is shown as a function of offered load in GB/s for DCAF, CrON and FCON in subfigures (b) and (c), respectively, of Figures 5.4 through 5.10. It is clear looking at these figures that DCAF outperforms CrON on every one of the synthetic traffic patterns



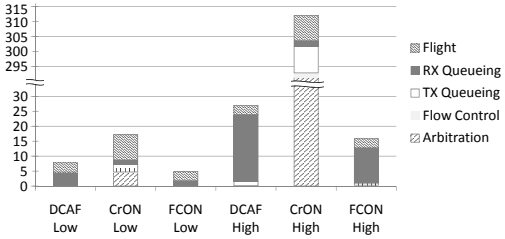
(a) Throughput (GB/s) vs. Offered Load (GB/s)



(b) Packet Latency (cycles) vs. Offered Load (GB/s)

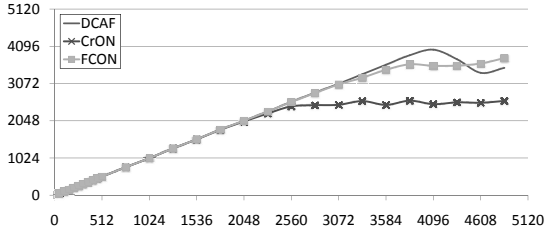


(c) Flit Latency (cycles) vs. Offered Load (GB/s)

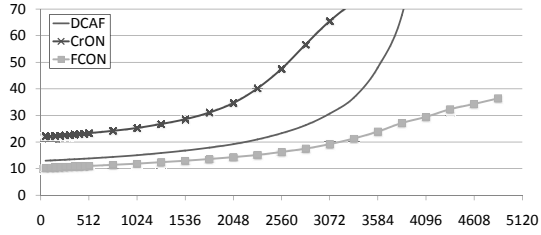


(d) Flit Latency Component (cycles) vs. Network & Load

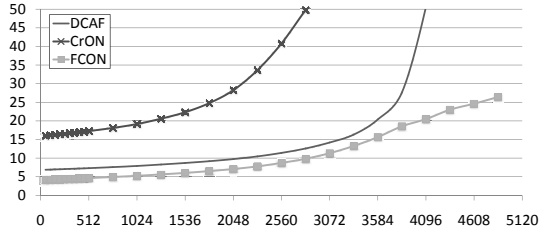
Figure 5.4: All Networks Throughput (a), Packet Latency (b) & Flit Latency (b) vs. Offered Load (GB/s) Flit Latency Components (d) vs. Network for Random Traffic Pattern



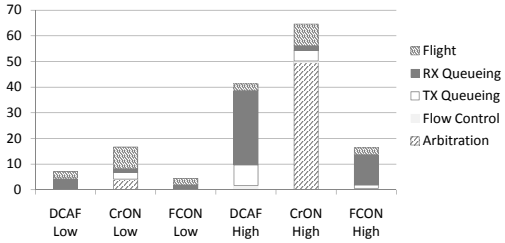
(a) Throughput (GB/s) vs. Offered Load (GB/s)



(b) Packet Latency (cycles) vs. Offered Load (GB/s)



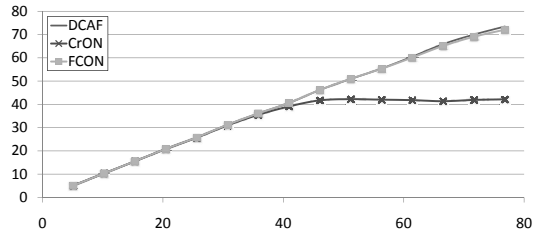
(c) Flit Latency (cycles) vs. Offered Load (GB/s)



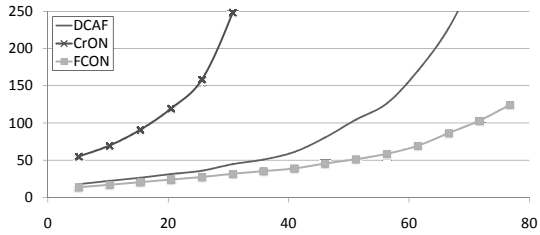
(d) Flit Latency Component (cycles) vs. Network & Load

Figure 5.5: All Networks Throughput (a), Packet Latency (b) & Flit Latency (b) vs. Offered Load (GB/s) & Flit Latency Components (d) vs. Network for NED Traffic Pattern

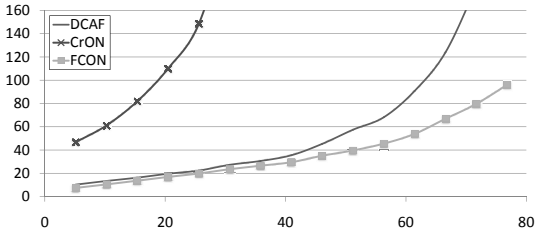
in terms of throughput and latency. Note also that the throughput for DCAF with the NED traffic pattern does not maintain a maximum level, but actually tapers off as a higher load is offered. This is due to the ARQ flow control – as the offered load increases, more flits



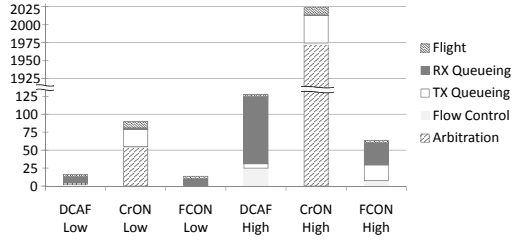
(a) Throughput (GB/s) vs. Offered Load (GB/s)



(b) Packet Latency (cycles) vs. Offered Load (GB/s)

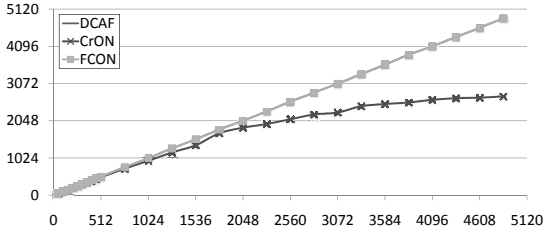


(c) Flit Latency (cycles) vs. Offered Load (GB/s)

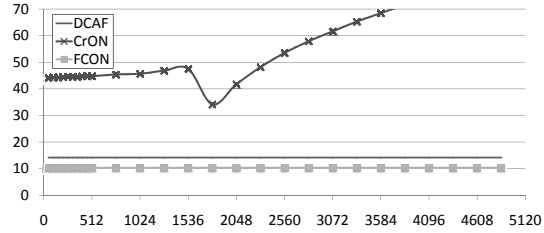


(d) Flit Latency Component (cycles) vs. Network & Load

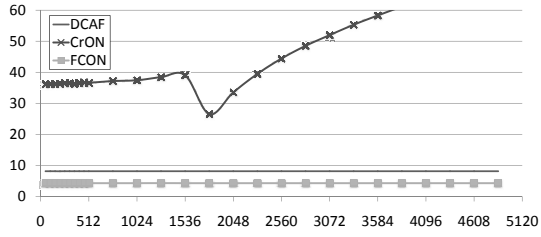
Figure 5.6: All Networks Throughput (a), Packet Latency (b) & Flit Latency (b) vs. Offered Load (GB/s) & Flit Latency Components (d) vs. Network for Hot-Spot Traffic Pattern



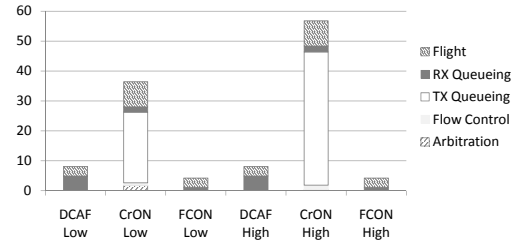
(a) Throughput (GB/s) vs. Offered Load (GB/s)



(b) Packet Latency (cycles) vs. Offered Load (GB/s)



(c) Flit Latency (cycles) vs. Offered Load (GB/s)

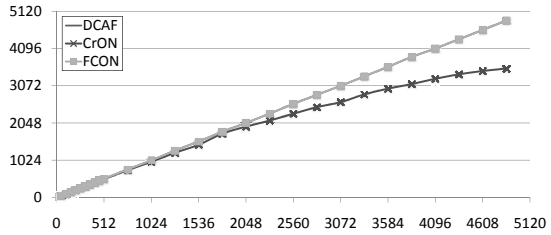


(d) Flit Latency Component (cycles) vs. Network & Load

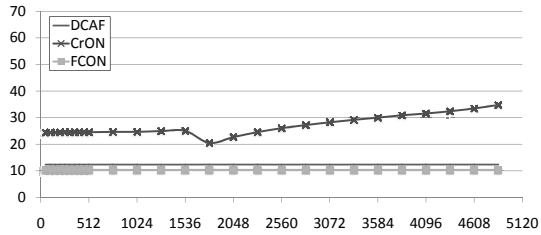
Figure 5.7: All Networks Throughput (a), Packet Latency (b) & Flit Latency (b) vs. Offered Load (GB/s) & Flit Latency Components (d) vs. Network for Tornado Traffic Pattern

are dropped and must be retransmitted.

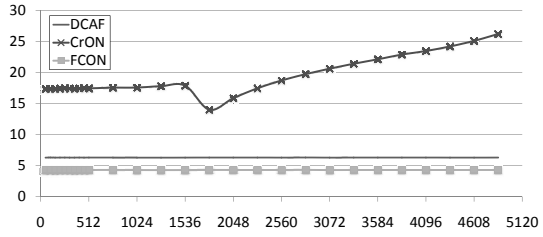
The reader may notice that the results for DCAF and FCON on the uniform random and NED traffic patterns (seen in Figure 5.4(a) and 5.5(a)) may seem incorrect



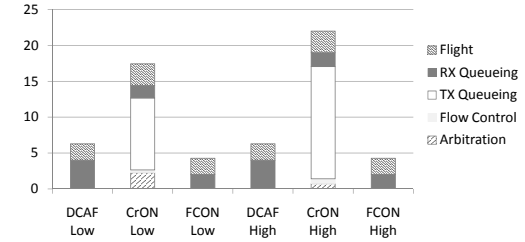
(a) Throughput (GB/s) vs. Offered Load (GB/s)



(b) Packet Latency (cycles) vs. Offered Load (GB/s)

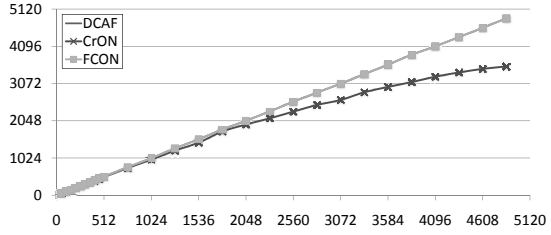


(c) Flit Latency (cycles) vs. Offered Load (GB/s)

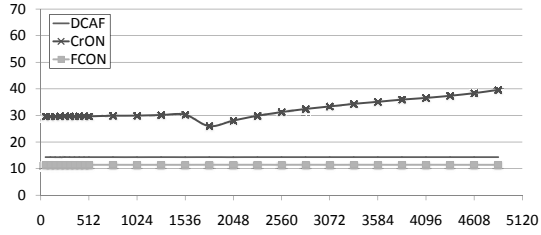


(d) Flit Latency Component (cycles) vs. Network & Load

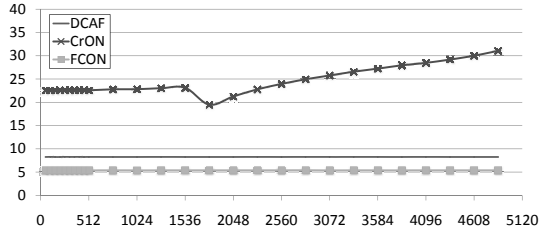
Figure 5.8: All Networks Throughput (a), Packet Latency (b) & Flit Latency (b) vs. Offered Load (GB/s) & Flit Latency Components (d) vs. Network for Nearest Neighbor Traffic Pattern



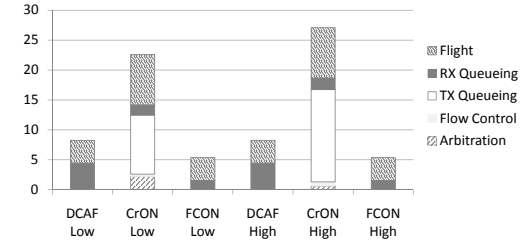
(a) Throughput (GB/s) vs. Offered Load (GB/s)



(b) Packet Latency (cycles) vs. Offered Load (GB/s)



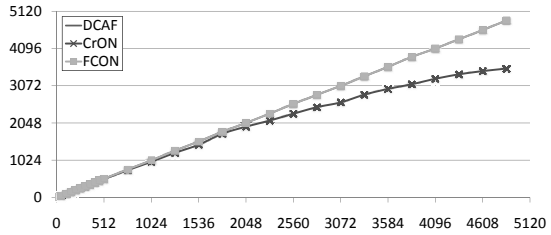
(c) Flit Latency (cycles) vs. Offered Load (GB/s)



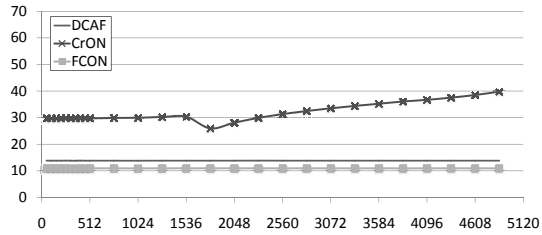
(d) Flit Latency Component (cycles) vs. Network & Load

Figure 5.9: All Networks Throughput (a), Packet Latency (b) & Flit Latency (b) vs. Offered Load (GB/s) & Flit Latency Components (d) vs. Network for Bit Inverse Traffic Pattern

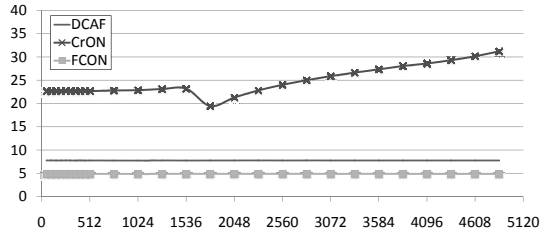
since DCAF has a higher throughput than FCON under high loads – this result is due to the re-enabling of the flow signal that is used in FCON. FCON could perform as well as DCAF if the same flow control scheme were employed, or if it used a re-enabling algorithm



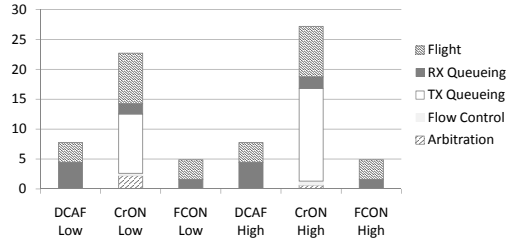
(a) Throughput (GB/s) vs. Offered Load (GB/s)



(b) Packet Latency (cycles) vs. Offered Load (GB/s)



(c) Flit Latency (cycles) vs. Offered Load (GB/s)



(d) Flit Latency Component (cycles) vs. Network & Load

Figure 5.10: All Networks Throughput (a), Packet Latency (b) & Flit Latency (b) vs. Offered Load (GB/s) & Flit Latency Components (d) vs. Network for Transpose Traffic Pattern

other than round robin. In Figures 5.7(a), 5.8(a), 5.9(a), and 5.10(a) the DCAF results are indistinguishable from the FCON results due to the fact that both networks perform ideally whenever the traffic pattern has a single source for each possible destination.

Since DCAF and FCON share the same topology, they have the same performance results for Tornado, Nearest Neighbor, Transpose, and Bit Inverse. From the graphs it appears that DCAF performs ideally on all traffic patterns except for NED. In reality, the performance of DCAF is slightly lower than the ideal starting at 56GB/s for Hotspot and 4096GB/s for Uniform Random.

In addition to the average flit and packet latencies, the breakdown of the latency components in cycles for the networks under low load ($< 10\%$ maximum offered load) and high load ($> 45\%$ maximum offered load) are shown in subfigure (d) of Figures 5.4 through 5.10. Notice in Figure 5.6(d) that the latency attributed to arbitration in CrON for Hotspot is greater than the entire flit latency for DCAF. In order to evaluate the contribution of arbitrating/flow control to latency, Figure 5.11 shows the average flit latency component due to arbitration in CrON and flow control in DCAF as a function of offered load when

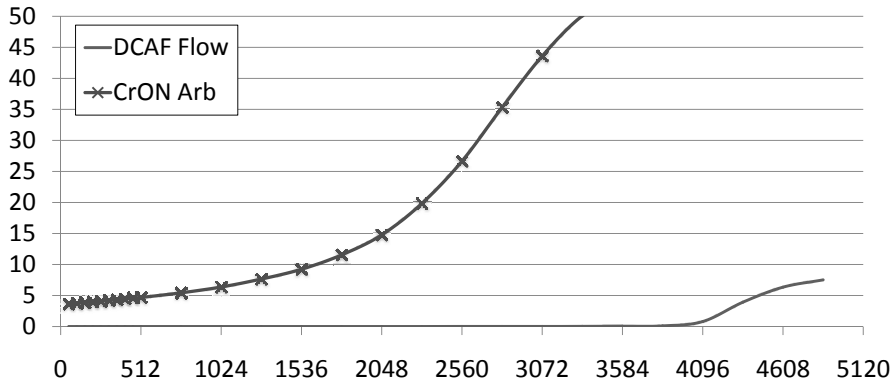


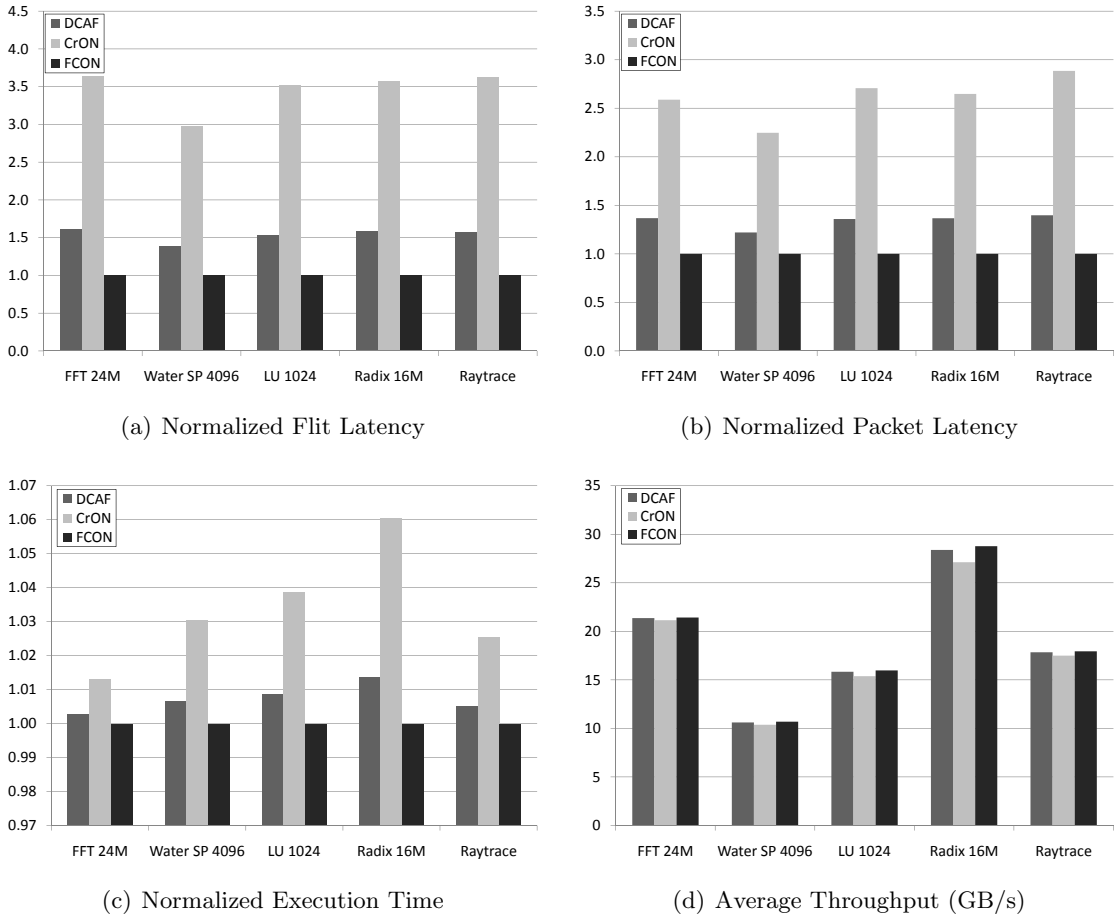
Figure 5.11: Latency (cycles) vs. Offered Load (GB/s) for NED Traffic Pattern

using the NED traffic pattern[1]. Notice that arbitration in CrON adds latency to each flit even under low loads, but the ARQ flow control in DCAF only adds latency when the network has become overwhelmed. Arbitration is an overhead that must be paid for each communication, while the ARQ flow control is an "on-demand" type of penalty that is only paid when the network is overwhelmed.

The performance results of the SPLASH-2 runs are shown in Figure 5.12. Figures 5.12(a) and 5.12(b) show the average flit and packet latencies for DCAF, CrON, and FCON, normalized to the network with the lowest latency (in all cases FCON). The figures show that DCAF and FCON have dramatically lower average latencies than CrON across all the benchmarks; however, the lower latency does not result in as dramatic a difference in the overall execution time as it did for the synthetic traces.

Figure 5.12(c) shows the execution time of each benchmark normalized to the shortest execution time, and the figure shows that DCAF executed the benchmarks from 1% to 4.6% faster than CrON, and on average less than 0.8% slower than FCON. Figure 5.12(d) shows the average throughput in GB/s for the various benchmarks. The average throughput (or average required bandwidth) of the SPLASH-2 benchmarks equates to ∼0.4% of the total network bandwidth for DCAF, but the average of the peak throughputs attained on the benchmarks was 5,104GB/s (approximately 99.7% of total network bandwidth) for DCAF.

[1]NED was chosen because the flow control component in DCAF is by far the highest in NED – it is negligible in the other traffic patterns.



(a) Normalized Flit Latency

(b) Normalized Packet Latency

(c) Normalized Execution Time

(d) Average Throughput (GB/s)

Figure 5.12: All SPLASH-2 Performance Results

5.3 DCAF Power

The minimum and maximum power consumption for DCAF, CrON, and FCON is shown in Figure 5.13. The minimum power consumption is the minimum power that is used even when the network is idle and at its lowest ambient temperature, while the maximum power is the maximum observed across all the simulations. The dominant factor for all networks is the laser power, which is expended regardless of activity. The reader may notice that CrON also consumes dynamic electrical power even when idle; this is due to the fact that arbitration tokens must be replenished every loop, requiring modulation of the arbitration microrings.

As one might expect, the overall maximum trimming power required for DCAF and FCON is higher than for CrON, since DCAF and FCON have ∼88% and ∼76% more

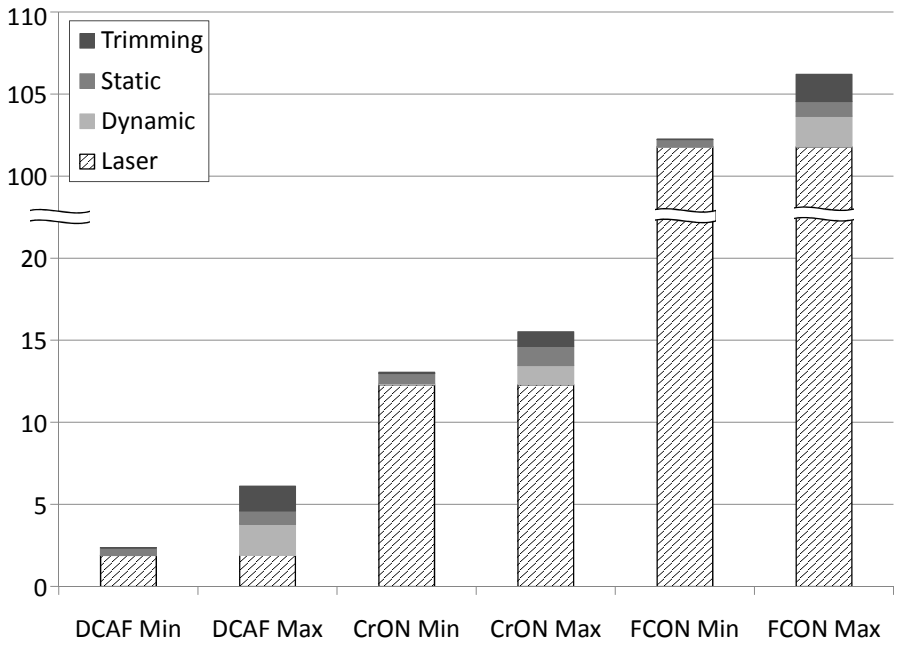
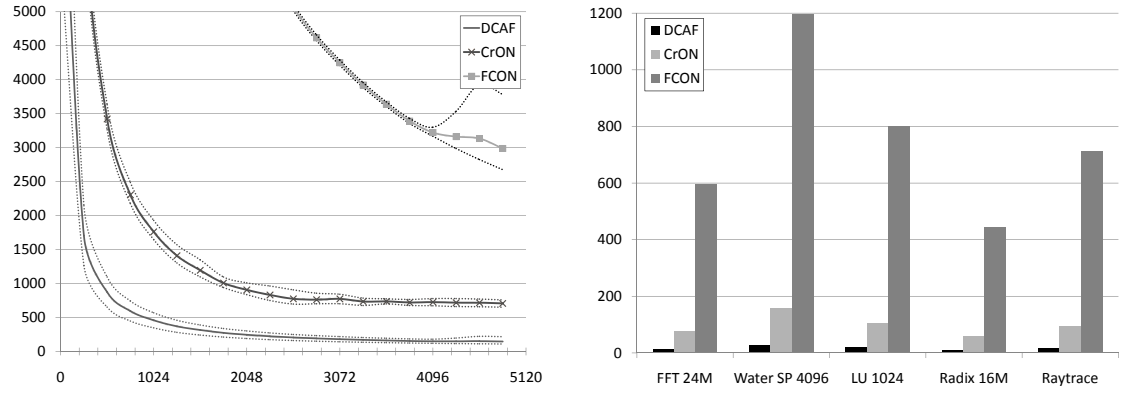


Figure 5.13: Power (W) vs. Network (Min/Max Load)

microrings, respectively. However, the average trimming power *per microring* is actually 18% higher for CrON when compared to DCAF. Previously I observed in Section 3.2 that the heating power required for trimming has a non-linear relationship with microring count, and my findings show that current injection has a non-linear relationship as well. CrON requires more trimming power per microring since the network operates at a higher temperature due to the greater total power consumption (static plus dynamic power) when compared to DCAF.

The maximum amount of dynamic power consumed by DCAF and FCON is much higher than that of CrON, but DCAF and FCON both greatly outperform CrON in the maximal case as well. Figure 5.14(a) shows the energy efficiency in fJ/b as a function of offered load in GB/s. The energy efficiency shown in Figure 5.14(a) is calculated by taking the power consumed divided by the actual network throughput (not the theoretical maximum throughput). The solid lines for DCAF, CrON, and FCON are the average energy efficiencies (the average power consumed divided by average throughput). The dotted lines show the minimum and maximum energy efficiencies for the two networks; the efficiency varies with achieved throughput and ambient temperature. DCAF is clearly more energy



(a) Offered Load (GB/s)

(b) SPLASH-2 Benchmark

Figure 5.14: Energy Efficiency in (fJ/b) vs. Offered Load (GB/s) (a) and in (pJ/b) vs. SPLASH-2 Benchmark (b)

efficient than CrON, and both DCAF and CrON are more energy efficient than FCON. For the synthetic traffic patterns run DCAF, CrON, and FCON approach 109, 652, and 2,675 fJ/b respectively, though this only occurs under high load. FCON could approach energy efficiencies below 100fJ/b if each node was simultaneously communicating to every other node, a situation that was not simulated in the current set of synthetic traffic patterns.

The energy efficiencies that can be obtained by DCAF, CrON, and FCON under high load are not observed when the networks execute the SPLASH-2 benchmarks, which can be seen in Figure 5.14(b). The average energy efficiency for DCAF, CrON, and FCON on the SPLASH-2 benchmarks were 24.1, 104, and 750 thousand fJ/b. The lower energy efficiency observed in these photonic networks under low load is a problem that will likely be shared with future on-chip electrical networks; while electric networks will not have the static laser overhead, the static electrical leakage is of greater and greater concern as we move from deep submicron into nanoscale technologies.

A network whose design offers limited performance may have the potential for higher energy efficiency, but a lower performing network will also impact the energy efficiency of the cores and caches due to the increased number of stalled cycles. Examining the impact of network performance on the energy efficiency of the cores is beyond the scope of this work.

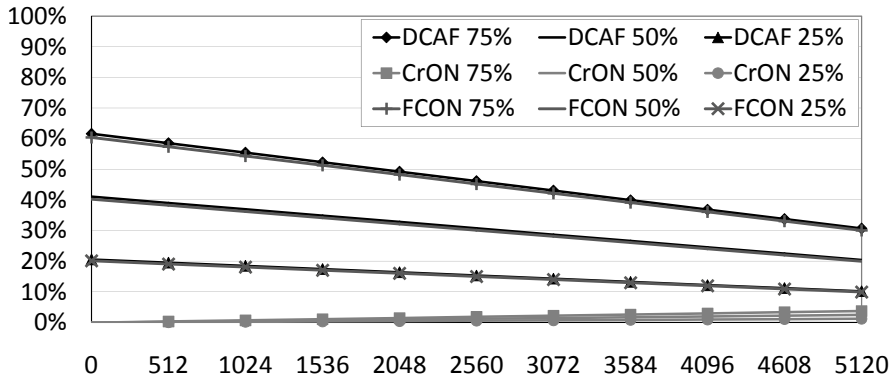


Figure 5.15: Percent of Laser Power Recaptured vs. Offered Load (GB/s)

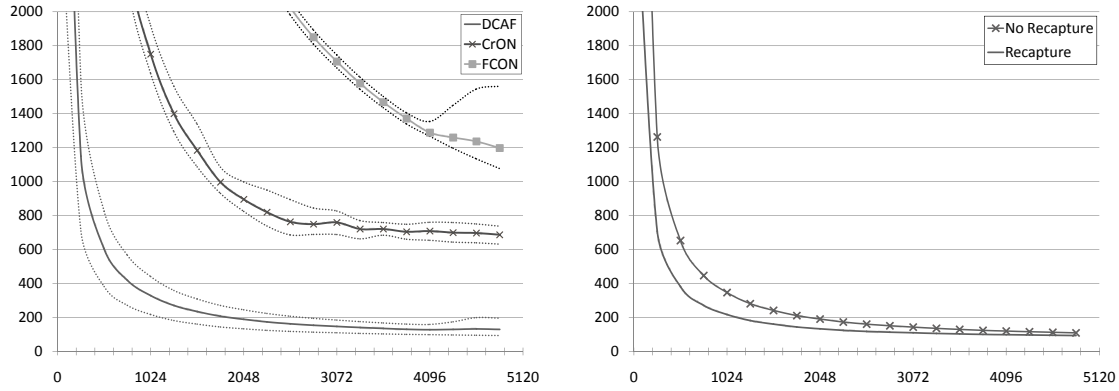
5.4 DCAF Discussion

This section presents a discussion of how the DCAF energy efficiency could possibly be improved under low workloads. A discussion of the scalability issues of DCAF, and how DCAF can be scaled beyond 64 nodes is also presented in this section.

5.4.1 Improving DCAF Energy Efficiency

Average energy efficiency is a common concern among computer architects. As was shown in the previous section, the average throughput of the SPLASH-2 benchmarks is very low compared to the total network bandwidth, and this low average throughput leads to a low average energy efficiency of the network. However, reducing the capabilities of the network is not necessarily desirable, since the entire network bandwidth *is* utilized at certain points in the benchmarks. The main reason for the energy inefficiency at low load is the large amount of static power required (the static leakage and fixed laser power). Reducing the static leakage power in electronic circuits is a well-studied area, but the approach of reducing the laser power or adjusting it to match the workload has not yet been examined.

At this point scaling the laser power is not a viable option, since lowering the incoming laser energy uniformly drops the power on all links, which means that an insufficient number of photons will arrive at the receivers. However, it is possible the unused energy could be recaptured – the photons not used to communicate could be harvested and turned into electricity. Converting the unused photons to electrons would be relatively straight-



(a) DCAF, CrON, & FCON with Recapture

(b) DCAF with/out Recapture

Figure 5.16: Energy Efficiency in (fJ/b) vs. Offered Load (GB/s) for DCAF, CrON, & FCON with 75% Efficient Recapture (a) and DCAF with and without 75% Efficient Recapture (b)

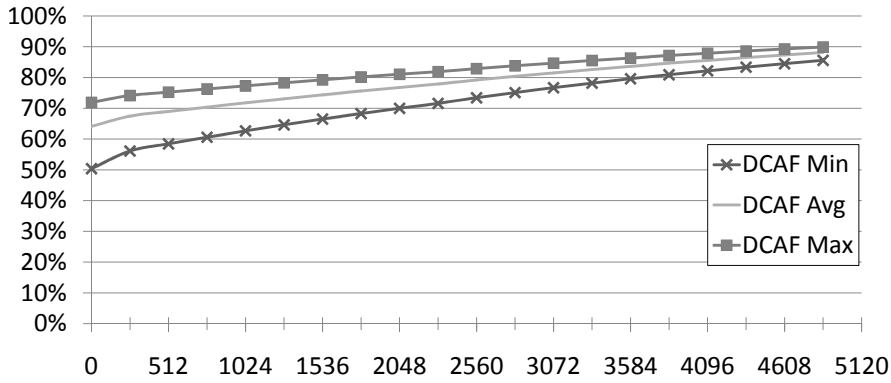


Figure 5.17: Percent of DCAF Total Power with 75% Efficient Recapture Compared to No Recapture vs. Offered Load (GB/s)

forward, requiring only the modification of existing photodiode structures. The number of photons available for recapture is a function of the activity occurring on each wavelength, which is related to the workload and the distribution of ones and zeros.

The theoretical limits of photonic energy recapture efficiency can be established using thermodynamic arguments, in the same way it was done for solar cells in [65]. Assuming a recapture area equivalent to 10k n-i-p photodiodes (440 μm^2 ea. [98]) and a liquid nitrogen ambient of 77K, and allowing a maximum photodiode temperature of 373K (100°C), we obtain a peak efficiency of ∼79%. This provides a definite theoretical bound for recapture efficiency.

Figure 5.15 shows the percent of laser power that potentially could be recaptured

for DCAF, CrON, and FCON assuming a 25%, 50% and 75% conversion efficiency. Notice that the slope of the lines for DCAF and FCON are opposite of CrON – DCAF and FCON can recapture the most photonic power under low load, while CrON recaptures the most under high load. This fact is due to the structure of the networks; CrON can only recapture a wavelength when a zero is being sent on that bit, while DCAF and FCON can recapture a wavelength whenever a one is *not* being sent on that bit (recapture occurs when there is no transmit or a transmit of a zero). Another reason DCAF and FCON have a higher recapture percentage is that recapture always occurs closer to the photonic source, where CrON would recapture potentially anywhere along the serpentine. These projections show that photonic recapture has the potential for substantially improving the energy efficiency of DCAF and FCON under low load, but the energy efficiency of CrON will only improve under high load (which unfortunately is the opposite of what is desired).

Figure 5.16(a) shows the energy efficiency in fJ/b of DCAF, CrON, and FCON vs. offered load in GB/s with 75% efficient recapture. The values in Figure 5.14(a) are almost identical for CrON, but the energy efficiency for DCAF and FCON is noticeably improved, especially for low offered load. Figure 5.16(b) shows the maximum energy efficiency in fJ/b vs. offered load for DCAF assuming no recapture and a 75% efficient recapture. Notice that the recapture has the greatest impact at low load. Figure 5.17 shows the percentage of total power needed for DCAF with 75% efficient recapture compared to DCAF without recapture – note that the idle power for DCAF is almost cut in half in the best case.

5.4.2 DCAF Scalability

Another common concern of architects is the scalability of network topologies. A 64-bit DCAF with 128 nodes will require an area of $\sim$293mm^2, but a 256 node DCAF would require $\sim$1,650mm^2. The photonic power of DCAF does not scale linearly either, although there is a less than 5% increase in required channel power scaling from 64 to 128 nodes. A 64-bit CrON with 256 nodes will require a smaller area ($\sim$323mm^2), but the photonic power of CrON will likely not scale to even 128 nodes. The number of off-resonance rings which light must pass through will roughly double when scaling CrON from 64 to 128 nodes, and this fact alone will increase the path attenuation by over 6dB. My estimates show that a

128 node CrON would require over 100W of photonic power. While the scalability of DCAF is limited to 128 nodes, CrON is limited to half that.

The bandwidth capability of DCAF is likely sufficient to support multiple cores per network node. As was shown by the SPLASH-2 benchmark performance results, the average network utilization of modern benchmarks is quite low. It is probable that an architect would choose to cluster multiple cores per node, as was done in [92], and then use DCAF to connect those clusters. The number of clustered cores which DCAF could support could also be increased by increasing the degree of simultaneous communication k. Increasing k would not only increase the total network bandwidth, but would also support multicasting at the hardware level.

Clustering processors electronically in this fashion would create a hierarchical network. If creating a hierarchical network is the chosen method for scaling, then connecting multiple smaller DCAF networks in a hierarchy may be a better solution. Assuming the goal is to support 256 nodes, this could be accomplished by using a DCAF network among 16 nodes and connecting these larger nodes using another level DCAF network. The local networks would have 17 nodes (16 cores plus one connection to the global network). Table 5.2 shows the breakdown and overall requirements of each of the subcomponents and the overall network. Notice that the required photonic power is less than 4x that of the 64 node DCAF – this is due to the reduction of off-resonance rings though which the light must travel in the smaller networks, plus the worst case paths are also shortened by creating the network hierarchy. Another counter intuitive result is that the required area is reduced while the microring count increases – this is due to the fact that the area calculation takes into account the waveguides surrounding the perimeter of each node, and the number of waveguides that must surround each node in the hierarchical is much smaller than in the 64 node case.

When comparing the average hop count (potential performance impact) and the energy efficiency of the two configurations, the hierarchy of smaller DCAF networks appears to win out over the hybrid network with electronic clustering as well. The average hop count is 2.88 and 2.99 for the 16x16 node hierarchical DCAF and four node electronically clustered 64 node DCAF, respectively. The energy efficiency for the 16x16 will approach 259fJ/b

Table 5.2: Hierarchical DCAF Network Parameters

Component	WGs	Microrings		Area (mm^2)	Bandwidth	Photonic
		Active	Passive		Total	Power (W)
Local Node	N/A	1,120	1,190	0.177	80GB/s	0.016
Local Network	272	∼20K	∼19K	3.01	∼1.3TB/s	0.277
Global Node	N/A	1,050	1,120	0.165	80GB/s	0.017
Global Network	240	∼16K	∼18K	2.65	1.25TB/s	0.277
Entire Network	∼4.5K	∼314K	∼334K	55.2	20TB/s	4.71

while the 4x64 would approach 264fJ/b; furthermore, the electrically clustered network value only accounted for the wiring efficiency and does not take into account the energy for the required repeaters (repeaters would be necessary to send the signal to the optical interface considering that the furthest a 10GHz signal can be sent in 16nm is ∼600μm according to the equations in [64]). In fact, the need to get the electrical signals to the optics is a significant challenge, one that has not been addressed in the literature so far – in the next chapter I present the results of some preliminary work on this matter.

Chapter 6

Electrical/Optical and Optical/Electrical Interface

As discussed in Section 2.1, microring resonators are an enabling technology which can be used modulate and filter the high quantity of wavelengths per waveguide needed for Dense Wavelength Division Multiplexing (DWDM). As shown in Table 3.1 these microring resonators are relatively large, typically ranging from 3μm to 10μm in diameter. Most of the photonic networks in the current literature [92, 40, 15, 71, 70, 42] assume the ability to use 64 wavelengths to create 64-bit data paths, which are modulated at 10GHz. Researchers thus far have focused on the photonic power required by the network, but details such as the power required to move data to and from the microrings has been largely ignored. When the overhead of Electrical-Optical-Electrical (E-O-E) interfaces are taken into account, it is not clear if 64 wavelengths per waveguide is the most energy efficient approach. In this chapter the trade-offs between the number of parallel bits and the signaling rate used to attain a target link bandwidth is explored.

The rest of this chapter is organized as follows; Section 6.1 discusses the trade-offs using varying number of parallel bits in WDM to meet a target bandwidth, while in Section 6.2 the description of the experimental setup is presented along with the results of the parallel bit trade-off experiment. Section 6.3 discusses the ramifications of changing the number of parallel bits for the three networks analyzed in this work, and the conclusions

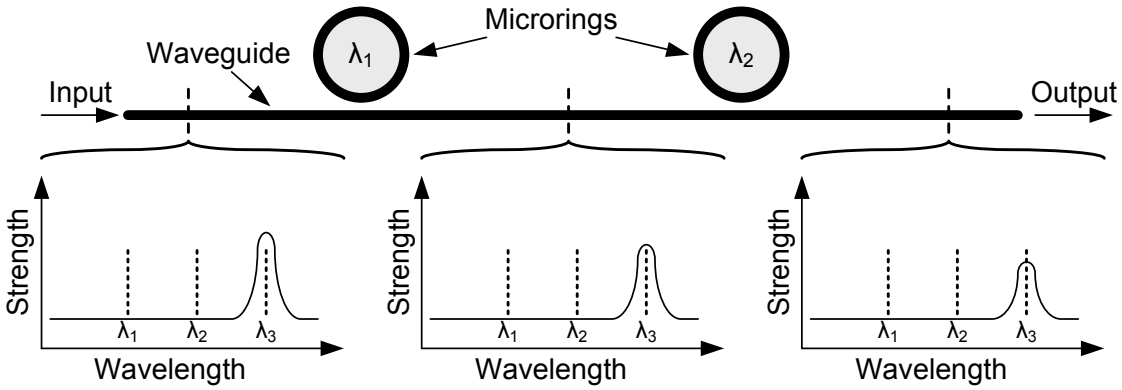


Figure 6.1: Signal deterioration due to off-resonance microrings

are presented in Section 6.4.

6.1 Data Path Width/Switching Speed Trade-off

Bandwidth is the amount of information transferred per unit time[1], so it is a function of both the number of bits being sent on each transmission and the number of transmissions done per second. The number of parallel bits (bits per transaction) and the signaling rate (transactions per second) used to attain a target link bandwidth involve trade-offs in both power and area - the more bits used the lower the signaling rate needed to provide the target bandwidth, and vice versa. Many of the terms in the power equation favor a lower switching frequency, while others favor reducing the number of hardware components. In the next two sections I will examine in more detail the photonic and electrical contributions to the total power used.

6.1.1 Photonic Power Requirements

The required laser power per wavelength is a simple calculation of $P_{PD}10^{\frac{A}{10}}$(where P_{PD} is the required power at the photodetector and A is the attenuation of the path). Assuming that Transimpedance Amplifiers (TIA) are not being used to compensate[2], the

[1]Bandwidth is defined differently in the signal processing realm – in this document the word "bandwidth" by itself will refer to this definition, while "3dB bandwidth" will indicate the signal processing version.

[2]Miller in [60] discusses the possibility of running an optical interconnect "receiverless", and that the advantages would include very low latency and high energy efficiency. This work assumes a receiverless design.

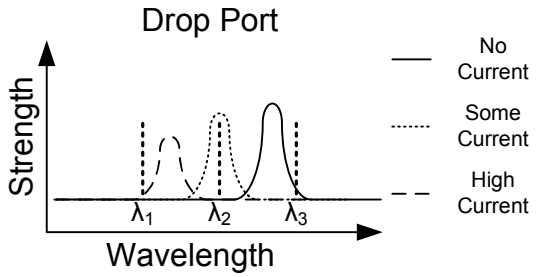
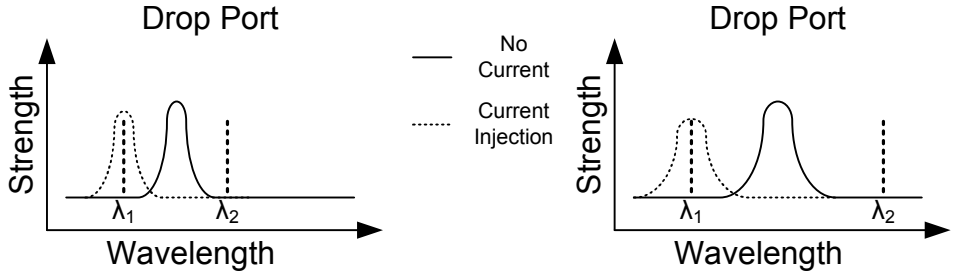


Figure 6.2: Degradation in signal quality



(a) Low Bandwidth (Narrower 3dB Bandwidth) (b) High Bandwidth (Wider 3dB Bandwidth)

Figure 6.3: Signal quality for modulation low bandwidth (a) and high bandwidth (b)

required power at the photodetector P_{PD} is based upon two factors: the amount of energy needed to switch the photodetector from a zero to a one and the switching frequency. When considering the data width/frequency trade-off, a narrower/faster link will require the photonic signal to pass through fewer off-resonance microrings; however, switching at a higher frequency requires a microring with an accordingly larger 3dB bandwidth[3]. This will in turn require larger resonance shift for modulation, and since resonance shift due to current injection increases the microring attenuation (as discussed in Section 3.3.1), there will be an increased path attenuation.

Figure 6.1 shows how the signal deteriorates as it passes through off-resonance rings. In the figure, one can observe that wavelength λ_3 is diminished in strength, even though it is passing microrings that are tuned to resonate to a different value (λ_1 and λ_2, in the figure.) Figure 6.2 illustrates how the signal deteriorates with increased current injection while the signal shifts towards the blue – the more the resonance needs to be shifted, the

[3]The Nyquist rate sets the limit on the maximum number of code elements per second that can unambiguously be resolved over a 3dB passband channel. A detailed discussion of the 3dB bandwidth requirements of microring resonators is beyond the scope of this work.

more the signal attenuates.

Figure 6.3(a) illustrates a low 3dB bandwidth modulation filter, while Figure 6.3(a) shows a higher 3dB bandwidth modulation filter. Increasing the signaling rate requires a correspondingly larger 3dB bandwidth filter (a wider filter, in other words). In order to maintain signal quality, the higher bandwidth filter must be shifted further for modulation, which is illustrated in Figure 6.3 – if the higher bandwidth filter is not shifted correspondingly further, the signal level difference of a zero and a one would be so small that it could become indistinguishable. Since the higher bandwidth filter needs to be shifted further, it also experiences a higher signal attenuation.

6.1.2 Electronic Power Requirements

The electrical power required by a microring based photonic link depends upon the power needed for trimming, microring modulation, the SERializer/DESerializer (SERDES), and local transport (the power required to drive the wiring from the network interface to the microring drivers). The power needed by receiver amplifiers such as TIAs is another potential factor in power consumption – TIAs can lower the photonic power required at the receiver at a cost of increasing the TIA electrical overhead. This additional degree of freedom would greatly increase the exploration space; therefore, they will not be considered in favor of a receiverless photodetection scheme. The impact of each of these factors will now be examined in more detail:

- **Trimming** – The trimming power for microring resonators was shown in Section 3.2 to have a non-linear relationship with microring count, and that the entire power/thermal analysis must be done to accurately estimate the required trimming power.
- **Modulation** – The power required for microring modulation is dominated by the microring capacitance and the switching frequency; the modulation power remains somewhat constant across various data width/frequency pairings for a given target bandwidth. Narrow, fast links will be negatively impacted by the fact that transistors will need to be increased in size as the switching limits of a technology point are approached.
- **SERDES** – Like microring modulation, SERDES will also favor a wider, slower link – the complexity of the serialization structure (whether using a serial shift register or a multiplexer) increases with the degree of serialization, as well as when the frequency of switching is increased.

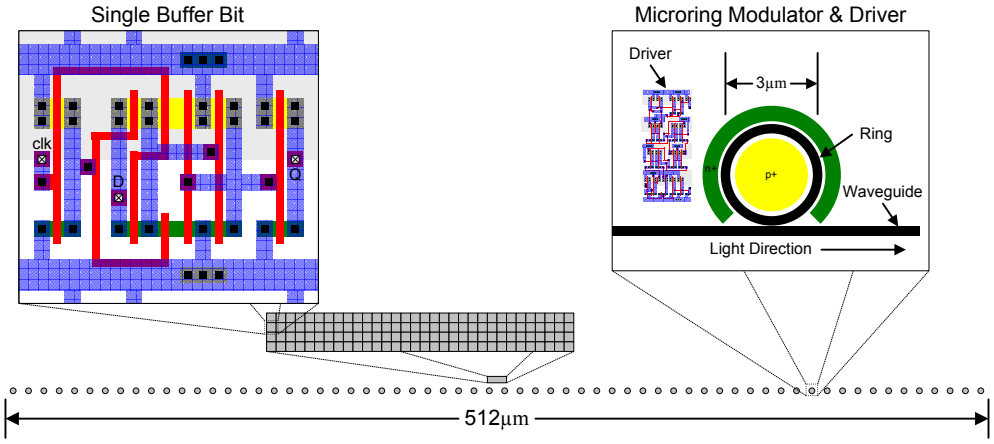


Figure 6.4: Electrical buffer comparison to 64 microring resonators

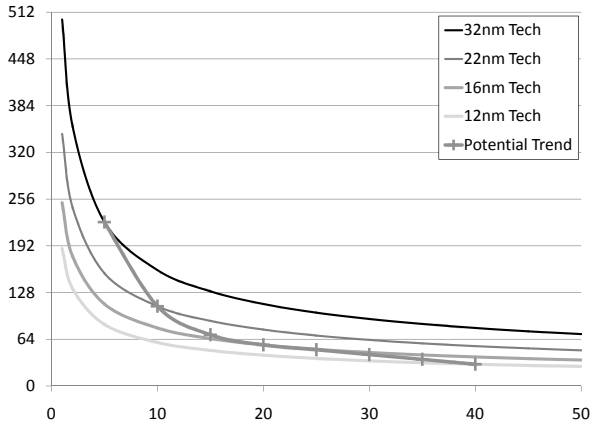


Figure 6.5: Maximum 8μm Pitch Microrings That Can Be Supported Without the Use of Repeaters by Wire Technology vs. Switching Speed (GHz)

- **Local Transport** – The length of the wires required to support a local transport within a node is typically ignored – however, the length of the wires required to support the microrings should be considered in the detailed power model since the distances are relatively large. For example, a 64 bit link with 8μm ring pitch (3μm diameter[4], 5μm spacing) is 512μm long, which is approaching the same order of magnitude as that assumed for inter-node links in an electrical mesh (for a 2 - 3mm wide tile).

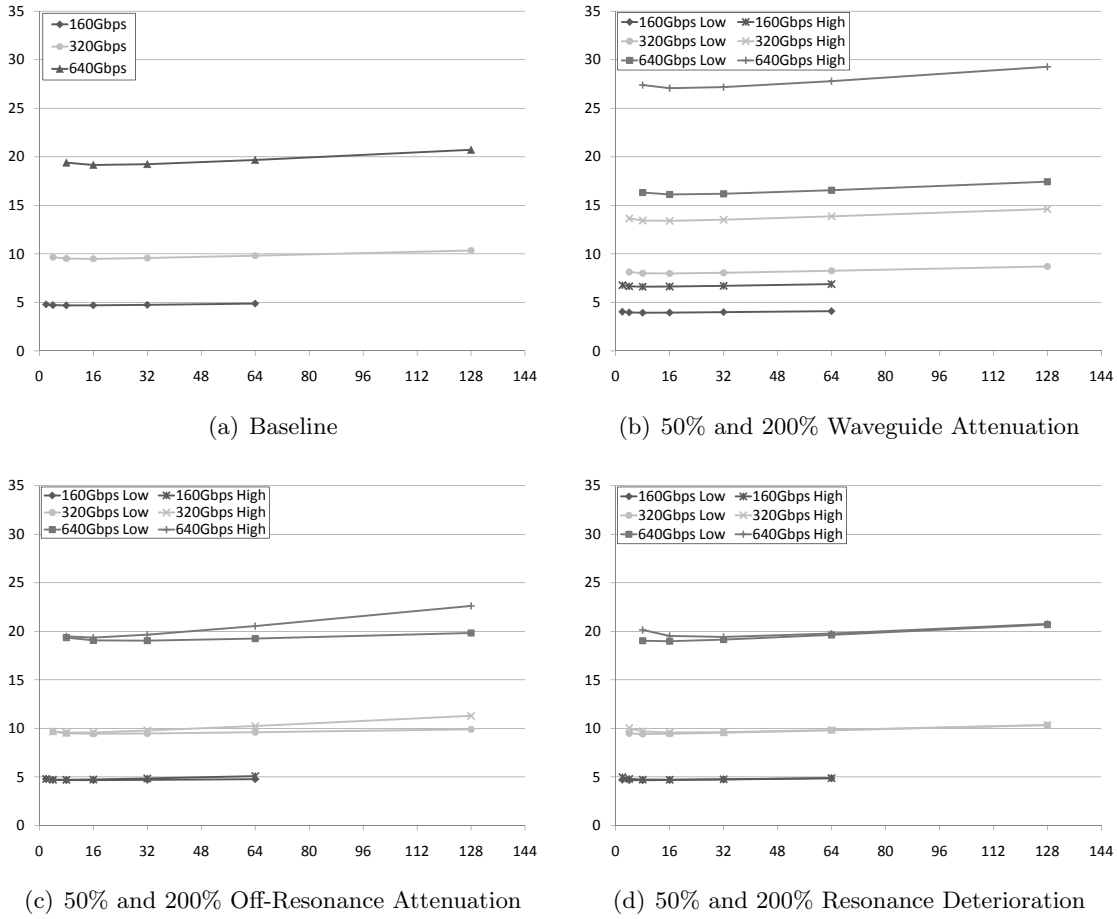
[4]According to Xu, Fattal and Beausoleil [96], "For a modulator working at 10 - 20 Gbit/s, a moderately high operating Q on the order of 10,000, which corresponds to an optical bandwidth of ~20 GHz, is appropriate for the critically coupled resonator, which requires an intrinsic Q of 20,000. ... one can conclude that the minimal radius to obtain an intrinsic Q of 20,000 around the wavelength of 1.55μm is 1.37μm." The Q factor is the quality factor for a resonator, and is written $\frac{f_0}{\Delta f}$ (where f_0 is the center frequency and Δf is the 3dB bandwidth of the resonator.) In other words, for a 193.5THz f_0 (1550nm laser), a microring diameter of 2.74μm is the minimum that can be used while maintaining an acceptable Q factor. Therefore, microring resonators do not scale with Moore's Law.

Figure 6.4 illustrates the relative distance for 64 microrings with a 8μm ring pitch and a 16nm electrical technology point. The grey rectangle in the center of the figure above the 64 rings is equivalent to the area needed for a 128 bit flit buffer - note the relative size of a standard cell of a D flip-flop (commonly used in buffers) compared to the size of the microring. The distance that the electrical signals must travel will become of greater and greater concern as feature sizes decrease, since Moore's Law does not apply to photonics. In essence the photonics will be growing in size relative to the electronics, in the same way the I/O pads on a chip have increased in size relative to the minimum feature size of devices.

Figure 6.5 shows the maximum number of microrings that can be supported without the use of repeaters on the y-axis vs. the switching speed on the x-axis (the maximum distance a wire can carry a signal was determined using the bandwidth equation from [64] and wire technology data from [34]). The figure shows that each successive technology point can support fewer microrings at the same switching speed, as shown by the line marked *potential trend*. This implies that to achieve a given link bandwidth, the electronics has to operate at a much higher clock rate (because of the narrower datawidth) as the technology shrinks. Accomplishing this could become exceedingly challenging or result in much higher dynamic power consumption, which implies that the bandwidth of a photonic link may be limited as the electrical geometries shrink.

The reader may be wondering why the microrings must be in a linear layout – why can't the waveguide just weave back and forth through a grid of microrings? The problem with laying out a single link in a grid is the number of tight bends that would be required. For the case of 64 microrings (in an 8x8 grid), this could increase the photonic signal attenuation by as much as 2.8dB if 0.1dB per 90° bend is assumed[5], which would almost double the required photonic power. In addition, it may not even be possible – many of the proposed on-chip photonic networks [92, 40, 70] assume the waveguides are laid out in lanes around the chip, and putting the microrings in a grid would require a dramatic increase in the complexity of the entire network.

[5]A 0.1dB loss is a reasonable assumption considering that a 0.32dB 90° bends have been demonstrated in [76] and losses of 0.0043dB per 180° 6.5μm radii bend have been demonstrated by Vlasov et al., IBM [45]



(a) Baseline

(b) 50% and 200% Waveguide Attenuation

(c) 50% and 200% Off-Resonance Attenuation

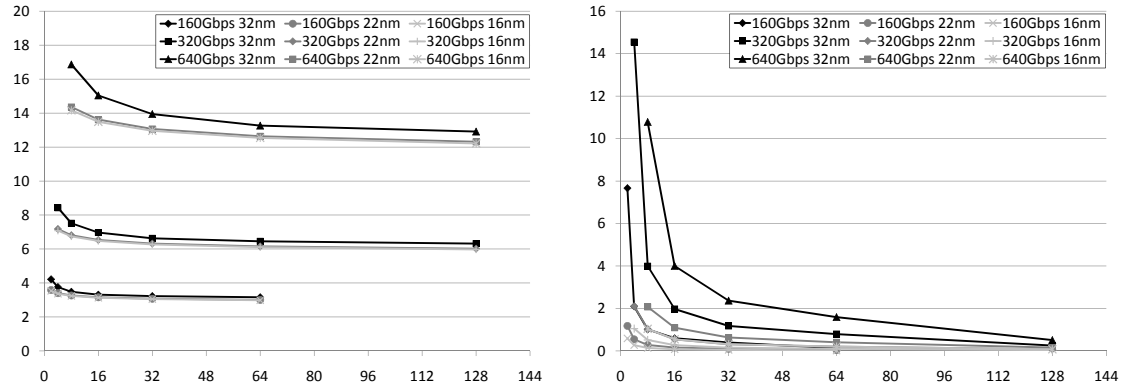
(d) 50% and 200% Resonance Deterioration

Figure 6.6: Link Photonic Power (mW) vs. Parallel Bits for Baseline (a), 50% and 200% Waveguide Attenuation (b), 50% and 200% Off-Resonance Attenuation (c), and 50% and 200% Resonance Deterioration (d)

6.2 Experimental Setup and Results

The power for each target link bandwidth, the number of parallel bits, and the technology point were calculated by determining the dynamic power components and photonic power individually. The total dynamic and photonic power, along with the component area, was then used to solve for the final resting temperature and static power consumption (which includes transistor static leakage and microring trimming costs). The photonic power and dynamic electrical power were calculated using Mintaka as was done for the work in the previous chapters.

I show results for three different target bandwidths. The baseline is 640Gbps (80GB/s) which is the assumed link bandwidth of the networks analyzed in the previous



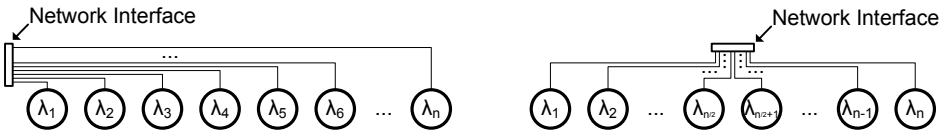
(a) Modulation (b) SERDES

Figure 6.7: Electrical Power (mW) vs. Parallel Bits for Modulation (a) and SERDES (b)

chapters – this link bandwidth is also the assumed "default" link bandwidth for many of the other published on-chip microring based networks. I also show the results for 320Gbps and 160Gbps for comparison.

Figure 6.6(a) shows the photonic power required in mW on the Y-axis versus the number of parallel bits on the X-axis. The lines in the figure show the link bandwidth – note that while the photonic power required increases with bandwidth, the energy efficiency (measured in energy per bit) is staying relatively the same (as the bandwidth doubles, so does the photonic power required). The sensitivity to varying the waveguide attenuation, off-resonance attenuation, and attenuation due to current injection resonance deterioration is shown in Figures 6.6(b), 6.6(c), and 6.6(d), respectively. The "High" values assumed double the attenuation (200%), while the "Low" values assume half the attenuation (50%). All the power values shown in Figure 6.6 are the photonic power required on-chip – they do not include the power lost upon entering the chip or the power required to create the photons in the laser. Looking at the figures it should be clear that for a given link configuration the required photonic power is much more sensitive to waveguide attenuation than it is to off-resonance attenuation or resonance deterioration.

Figure 6.7 shows the electrical power required in mW on the Y-axis versus the number of parallel bits on the X-axis for the modulation of the microrings and SERDES. The power required for modulation goes down as the number of parallel bits increases, since the switching speed decreases (illustrated in Figure 6.7(a)). The modulation power is



(a) Network Interface at End of Microrings (b) Network Interface at Middle of Microrings

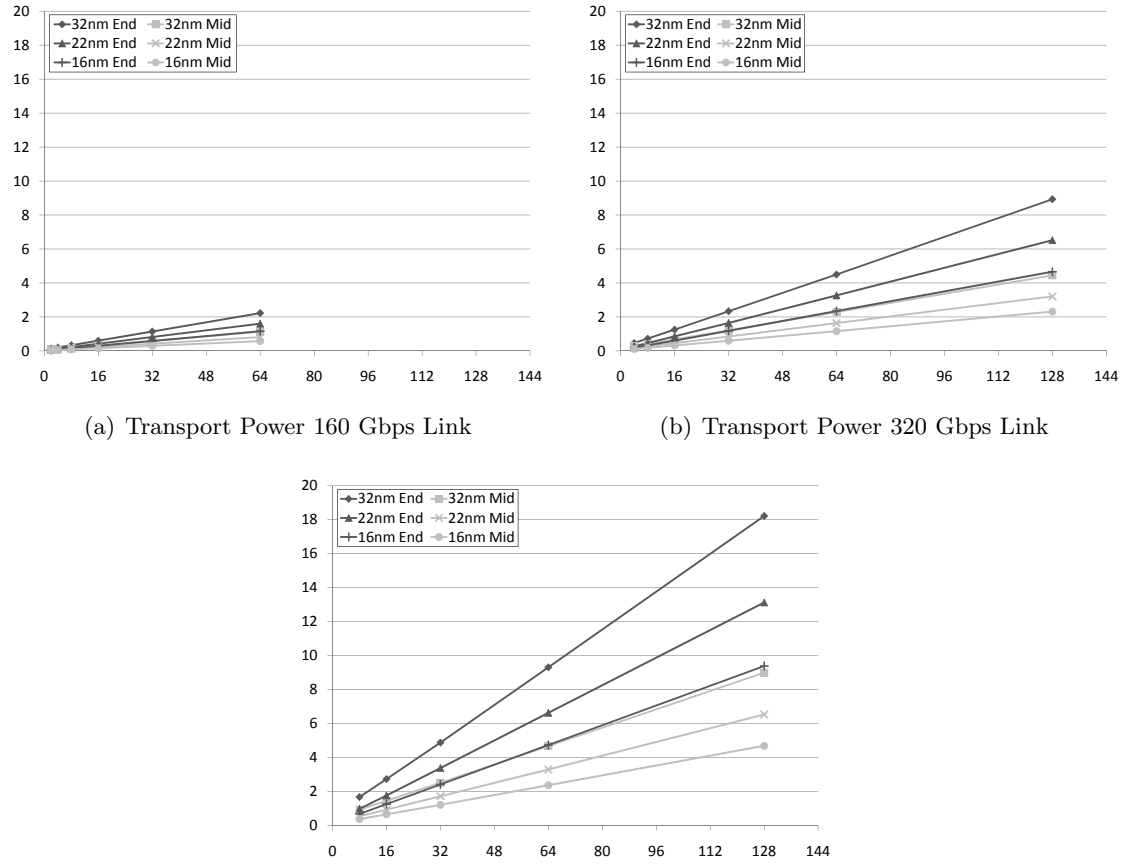
Figure 6.8: Microring Wiring with Network Interface at End (a) and Middle (b) of Row

also reduced with each successive technology point, due to the reduced capacitance of the driver circuit; however, the capacitance of the microring resonator limits the amount the modulator power can be reduced.

Figure 6.7(b) shows the power required for SERDES, and as one might expect, the power dramatically decreases as the number of parallel bits increases, since the switching speed and SERDES structure capacitance are decreasing simultaneously. Each successive technology point also shows a dramatic reduction in power. The modulation and SERDES power requirements clearly favor the use of a higher number of parallel bits.

Figure 6.8 shows two possible configurations for wiring a row of microring resonators. The microrings with the network interface at the end as in Figure 6.8(a) will obviously require longer wires than a network where the interface is in the middle of the row as in Figure 6.8(b). Figure 6.9 shows the electrical power required for the local transport in mW on the Y-axis versus the number of parallel bits on the X-axis. The transport power shown assumes the wires must run from one end of the row to the microrings (denoted as "End", see Figure 6.8(a)) or that the wires are run outward from the middle of the row to the microrings (denoted as "Mid", see Figure 6.8(b)). The wire to each microring was properly sized (local, semi-global, global) using the bandwidth equation from [64] and wire technology data from [34]. A few of the extreme cases (e.g. 16nm "End" for 128 bits) required repeaters in order to provide the proper bandwidth to the furthest microrings, and the repeater power is included for those cases.

As one would expect, the power increases with the increase in wire length – the "End" values require higher power than the corresponding "Mid" values. The total capacitance of the wires is roughly bound by $\frac{N(N+1)}{2}$ where N is the number of parallel bits, leading one to expect the power to increase quadratically; however, as the number of parallel



(a) Transport Power 160 Gbps Link

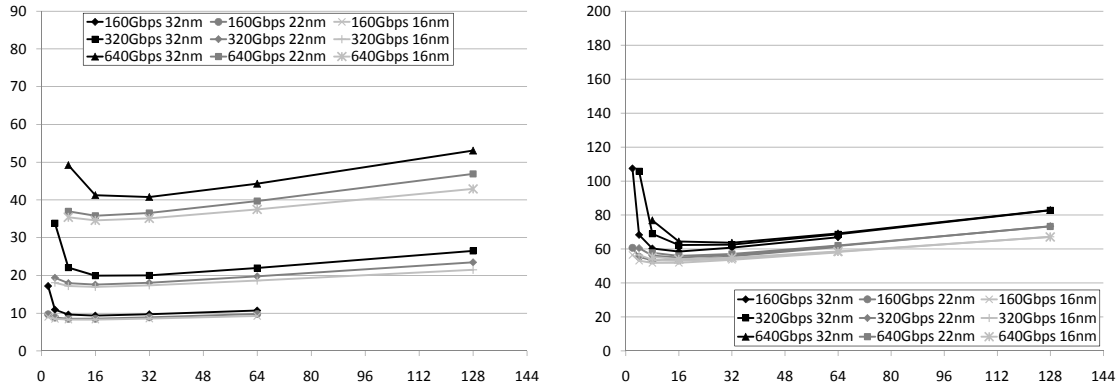
(b) Transport Power 320 Gbps Link

(c) Transport Power 640 Gbps Link

Figure 6.9: Electrical Power (mW) vs. Parallel Bits for Local Transport at 160Gbps Link (a), 320Gbps Link (b), and 640Gbps Link (c)

bits increases the switching rate decreases, resulting in an approximately linear growth. As discussed in the previous section, the number of microrings that can be supported without needing repeaters decreases with each successive technology point, as does the transport power (since the capacitance of the wires decreases). It is clear that unlike the dynamic electrical power components, the local transport power consumption favors a reduction in the number of parallel bits.

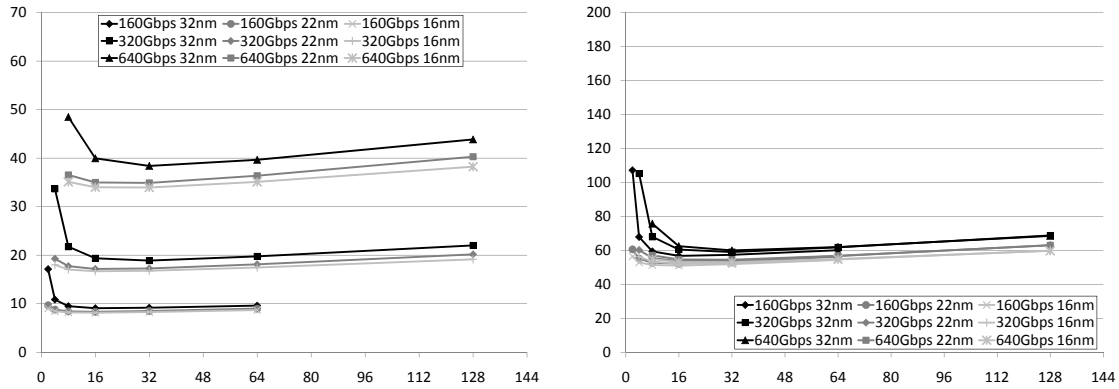
Figure 6.10 shows the total link power in mW (Figure 6.10(a)) and the total link energy efficiency in fJ/b (Figure 6.10(b)) for links assuming local transport starts at one "End" of the row of microrings. The results show that 64 parallel bits, the default value used by most researchers in this area, is never the most energy efficient for the configurations analyzed. Sixteen parallel bits is the most energy efficient across all the configurations, with



(a) Total Link Power

(b) Total Link Energy Efficiency

Figure 6.10: Total Link Power (mW) (a) and Total Energy Efficiency (fJ/b) (b) vs. Parallel Bits for Local Transport Starting at "End"



(a) Total Link Power

(b) Total Link Energy Efficiency

Figure 6.11: Total Link Power (mW) (a) and Total Energy Efficiency (fJ/b) (b) vs. Parallel Bits for Local Transport Starting at "Mid"

the exception of 22nm and 16nm technology points at 160Gbps (where 8-bits is optimal), and 32nm technology point at 640Gbps (where 32-bits is optimal).

Like Figure 6.10, Figure 6.11 shows the total link power in mW and the total link energy efficiency in fJ/b when the local transport wires run outward from the center of the row of microrings. Again the totals show that 64 parallel bits is never the most energy efficient for the configurations analyzed; furthermore, 16 parallel bits is the most energy efficient across all the configurations, with the exception of all technology points at 640Gbps (where 32-bits is optimal).

The power and energy efficiency values shown in Figures 6.10 and 6.11 assume

Table 6.1: Parallel Bits & Convergence Point for Most Energy Efficient Configuration

Tech Point	**160Gbps**	**320Gbps**	**640Gbps**
32nm "Mid"	8/16 (7.3%)	16/32 (21%)	16/32 (6.8%)
22nm "Mid"	8/16 (29%)	16	16/32 (60%)
16nm "Mid"	8/16 (39%)	16	16/32 (70%)
32nm "End"	8/16 (9.6%)	16	16/32 (18%)
22nm "End"	8	16	16
16nm "End"	8	16	16

a 100% link utilization, which is unrealistically high (at least in the steady state.) As discussed in the previous chapter, the actual energy efficiency is highly dependent on the link utilization. Therefore, I analyzed the energy efficiency of each of the configurations while varying the link utilization. Table 6.1 shows the results of this analysis. The single value results indicate the number of parallel bits that are the most energy efficient regardless of link utilization, while the split values show the most energy efficient number of parallel bits for low link utilization (left), and for higher link utilization (right). The percentage in the parentheses is the link utilization point where the most energy efficient configurations converge (both configurations are equally energy efficient). The results show a trend that as the technology point shrinks, there is a higher convergence point – this occurs because the static power becomes a larger portion of the total link power.

6.3 Network Ramifications

Based on the results from the previous section, it appears that FCON, CrON, and DCAF should have been evaluated with either 16 or 32-bit data paths (given that the target link bandwidth is 640Gbps in a 16nm technology). Unfortunately, it is not that simple – the results presented in this chapter can not be used directly to determine the most energy efficient configuration for any arbitrary network. The additional wavelengths required in order to support each network link (such as arbitration in CrON and flow control in DCAF), for example, impact the overall results. Furthermore, some networks benefit more than others from a reduction in the number of parallel bits, due to the configuration of the photonic link.

The CrON architecture has the potential to significantly lower power consumption by using a 16-bit instead of 64-bit data path, because the path attenuation is dramatically reduced due to the many fewer off-resonance microrings through which the light must pass. A 16-bit, 40GHz CrON would remove 3,072 off-resonance microrings (64 nodes times 48 microrings/node) from the worst case path, which would reduce the worst case path attenuation by over 3dB. The clock, arbitration token, and fast forward wavelengths prevent the laser power from being cut in half, since they must continue to operate at the speed of the network, but the laser power is reduced from 12.2W to 7.45W. The primary challenge to creating a 16-bit CrON operating at 40GHz is the extremely tight timing required between the reception of the arbitration token and the transmission of the next flit. The furthest a 40GHz signal in can travel in 16nm technology is just over 300μm – depending upon the layout of the data waveguides to the arbitration waveguides, the wiring may be the limiting factor.

Like CrON, FCON has a reduction in path attenuation when moving from a 64-bit to a 16-bit data path (although the reduction is much less dramatic). This does not result in a reduction in required laser power, however – the clock and flow control wavelengths do not scale down like the data path does, and therefore there is an actual increase in laser power by 7.9W since four times as many transmissions must occur to move the same amount of data. Since the wavelengths used for clock and flow control must be different than the ones used for data transmission, reducing the power consumed by these wavelengths will not be straightforward. The clock wavelength must operate at the speed of the network so that phits can be detected at the receiver, although the flow control signal does not necessarily need to operate at the same speed. It may be possible to reduce the power required by the flow control wavelength using a splitter and recapturing the excess power, as was suggested in Section 5.4.1. This will be left for future work.

DCAF also experiences a reduction in path attenuation but an increase in photonic power of approximately 24% when moving from a 64-bit to a 16-bit data path, due to the support signals (the clock and ARQ ACK wavelengths). Like FCON, DCAF cannot reduce the power of the clock wavelength; however, the ARQ ACK potentially could be serialized, though this would reduce the capability of DCAF to acknowledge received flits. Even with

serialization, there would be a 0.7% increase in overall laser power.

As was shown in this section, the results from the previous section cannot directly be applied to all network configurations. FCON and DCAF would actually see significant increases in photonic power (greatly outweighing the electrical gains) if a narrower data path were to be employed, due to the fact that the control and data wavelengths are not completely decoupled. CrON, on the other hand, experiences a dramatic reduction in photonic power and could greatly benefit from a narrower data path – though nuances of the arbitration scheme may make a 40GHz CrON impractical.

6.4 EO/OE Conclusions

This chapter explored the trade-offs between power consumption, the number of parallel bits and the signaling rate used to attain a target link bandwidth. It was shown that when the overhead of E-O-E interfaces are taken into account, using the maximum number of wavelengths available may not result in the most energy efficient network. This is primarily because Moore's Law does not apply to photonics – the electronics necessary to move data to and from a photonic network shrink as technology advances, while the photonic aspects do not. This increasing size disparity has a significant impact on overall energy consumption.

As was discussed in Section 6.3, the results from Section 6.2 cannot directly be applied to all network configurations. FCON and DCAF actually see significant increases in photonic power if a narrower data path were to be employed, due to the fact that the support and data wavelengths are not completely decoupled. CrON, on the other hand, sees a dramatic reduction in photonic power and could greatly benefit from a narrower data path if the electrical components can support the arbitration scheme at 40GHz. The overhead of the E-O-E interfaces must be taken into account along with the network topology and average link utilization in order to accurately determine the most energy efficient configuration that meets the target link bandwidth.

This chapter again shows that energy efficiency is based on workload, and while trimming power may be eliminated if athermalized microrings can be constructed, the

static power of the external laser will continue to be a problem (though photonic energy recapture as discussed in Section 5.4 could help remedy this situation). The potential use of on-board nanolasers and surface plasmon antenna as discussed by Miller in [87] to decouple the energy efficiency from the workload warrants further investigation, since these structures (when combined with microring resonators) may allow DWDM while reducing the static power overhead and bringing electrical/optical and optical/electrical interface bits physically closer together.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

Future multicore systems will require high bandwidth communication networks, and electrical networks are not likely to scale up well (primarily for latency and power consumption reasons). Fortunately, optical interconnects promise to provide higher bandwidth while consuming less energy than electrical interconnects, and the unique properties of optics can be exploited to create topologies that are impractical using only electronics. Unfortunately, microring resonators are known to be extremely sensitive to thermal variations and fabrication inaccuracies.

The thermal sensitivity of microring resonators motivated me to investigate if large networks based on microring resonators are feasible. In order to conduct this investigation I developed a power and floor-plan simulation library for use in photonic research, which I have named Mintaka. Mintaka was integrated with Hot-Spot to provide a closed loop power/thermal solver that is capable of calculating the system level power requirements of microring based on-chip networks. I first used Mintaka to verify that the network itself was thermally stable, and then began investigating the impact of trimming. I showed that trimming has a non-linear relationship with microring count, a finding which invalidates an assumption commonly made by researchers in this area. I also discovered that using current injection to trim microrings leads to thermal runaway very quickly (within 1°C). In order to deal with this problem I proposed and analyzed the Sliding Ring Window technique,

and showed that it would increase the range of temperatures over which the rings remain usable/stable. The trimming study also showed that rings can be trimmed as a co-located group, but this requires that all rings in the group be uniformly spaced spectrally. The uniformity of the spectral spacing will be degraded by fabrication defects that cannot cost effectively be fixed though post fabrication techniques, and these defects will reduce the reliability of microring based photonic links.

The observation that fabrication defects would reduce reliability led me to investigate methods to improve the reliability of microring based links. In order to accomplish this I developed the first microring based fault model. Using this fault model I demonstrated that the fault rates for photonic microrings must be extraordinarily low before optical networks can be implemented without using any error correction or error detection schemes, particularly if they want to meet a 1M hour Mean Time Between Failure (MTBF). I also showed that by choosing the correct structures, links can be constructed such that the errors resulting from non-interfering faults will be asymmetric – this is important because any arbitrary number of errors can be detected if the errors are asymmetric. Even in the presence of interfering faults, and highly unreliable microrings, I show that a 1M hour MTBF can be achieved using an N choose K (NcK) encoding.

Having established that large nanophotonic on-chip networks face no insurmountable hurdles, I decided to explore topologies that are implementable in photonics but not realizable using electronics (a fully connected network, for example). Analyzing these photonic topologies required a network simulator, and I realized while developing it that ignoring dependencies within network traces can lead to significant errors. Therefore, I developed a dependency tracking network performance simulator that was used in [68] to prove that dependencies must be tracked in trace based simulation and that the common methodology (non-dependency based trace simulation) used by researchers should be altered. The dependency tracking simulator was integrated with Mintaka to provide performance, power, and thermal results for a Fully-Connected Optical Network (FCON), and these results were compared to the baseline a Crossbar Optical Network (CrON). My results show that FCON greatly outperforms CrON, as one might expect a fully connected network to do, but the results also showed that FCON requires a tremendous amount of photonic power.

The performance advantages coupled with the extreme power requirements of FCON led me to investigate methods to reduce the photonic power while maintaining the network performance. What resulted was the Directly-Connected Arbitration Free (DCAF) family of networks, which are capable of being designed with a degree of simultaneous communication k from 1 to $N-1$. My simulations of DCAF with k=1 shows it outperforms CrON under all traffic patterns while consuming significantly less power. I showed that on some traffic patterns the flow control scheme used in DCAF outperformed the simplistic scheme used in FCON. In fact, I showed that ACK based ARQ protocol used in DCAF for flow control can be expanded with forward error correction to make a resilient HARQ protocol. My power results also show that energy efficiency is highly related to workload – this is due to the relatively large amount of static power overhead involved, primarily in the form of the external laser and microring trimming.

During the detailed power study, I realized that significant power was consumed in getting data to and from the group of microring resonators. This observation led me to investigate the impact of Electrical-Optical-Electrical (E-O-E) interfaces on the total link power. The underlying problem is that photonics do not scale with Moore's Law, meaning the photonics are in essence growing in size relative to the electronics. My E-O-E researched involved exploring the trade-off between the data path width and the modulation frequency in order to maintain a given bandwidth, and showed that the default of 64-bit data paths assumed by most photonic researchers is not the most energy efficient. I also found that one cannot create a universal data width/frequency chart, because the nuances of the link configurations and the control plane make a narrow data path highly beneficial for some networks while actually being detrimental for others. The ramifications of this result reinforces my findings from the trimming study: computer architects must take a holistic approach when designing microring based photonic on-chip networks.

7.2 Future Work

The field of photonics has great potential, but there are still many questions that need addressing. In this work, I proposed and performed an initial analysis of the DCAF

family of networks, but further exploration of DCAF should be conducted. The trade-offs of varying the degree of simultaneous communication k in DCAF should be investigated – included in this investigation should be the determination of the appropriate value of k and the optimal partitioning of nodes among the k groups, given a particular target application. In Section 5.1 I discuss the possibility of adding a clock wavelength for every grouping of data bits (e.g. 8 bits) in order to send smaller transmissions to multiple destinations, but also support full sized transmissions to a single destination. A DCAF topology that implemented the data grouping mechanism could multicast short messages (such as cache coherence invalidates), or could use the mechanism for ACK messages removing the need for dedicated ACK wavelengths – the potential advantages of grouping of data bits should be studied. Further investigation into the use of DCAF in a hierarchical network is also warranted; the investigation into the optimal k value and partitioning could be folded into the study of the performance, power, and area trade-offs of various hierarchical DCAF designs.

Considering that DCAF already has an ACK based ARQ protocol, the performance/power impact of various error detecting/correcting schemes should be conducted. The techniques discussed in Section 3.3 focus on improving the reliability of data at the phit level. Given the potential bandwidth of DCAF it is likely that protocols can be developed at a higher protocol layer to correct for or even avoid failed links. For example, since DCAF does not require arbitration, it is entirely possible that periodic link testing can be implemented to identify functional and non-functional links. Completely failed or severely degraded links could be avoided entirely by routing traffic through any other node (at a cost of only one addition hop). Self-healing networks and path restoration algorithms have been well studied for long haul communications [95, 63, 39, 38], and one would expect that some of these techniques could be applied to on-chip networks as well. The potential to create a highly fault tolerant DCAF and the impact on power/performance of these resilience techniques should be further investigated.

This work also shows that energy efficiency is highly based on workload, and while trimming power will decrease as more and more thermally resistant microrings are created, the static power of the external laser will continue to be a problem. Even if highly

efficient photonic recapture (as suggested in Section 5.4.1) can be implemented the energy efficiency will remain dependent upon workload. Dynamic frequency scaling is widely used by computer architects to reduce power in processors, and it may be possible to use a similar approach for the network by reducing the photonic power on all wavelengths and clock the network slower (especially if using a receiverless approach). The power savings will likely have diminishing returns as the frequency is scaled down due to the fact that photodetectors leak and that the laser efficiency will not be constant across all power outputs; however, this approach to improving the network energy efficiency could very easily be folded into any dynamic frequency scaling regimen and should be investigated further. Another potential approach to scaling the photonic power that should be studied is to use multiple lasers in combination with the grouping of data bits discussed earlier – in this approach each grouping of data bits would have a dedicated laser to provide the wavelengths. The full width data path could be used under high load, and a narrower data path could be used under low and moderate loads allowing the lasers that provide the unused wavelengths to be turned off.

As explained in Chapter 6, there will also be the continuing challenges of getting the electrical signals to the microrings, as the electronics shrink and the photonics do not. Therefore, the potential use of on-board nanolasers and surface plasmon antenna as discussed by Miller in [87] to decouple the energy efficiency from the workload needs more attention – nanolasers and surface plasmon antennas could be combined with microring resonators to provide DWDM. In addition, recent work [94] that uses passive microring resonators in combination with plasmonics to modulate the signal is encouraging. These structures may significantly reduce the static power overhead and help solve the scaling problem by bringing electrical/optical and optical/electrical interface bits physically closer together.

References

[1] J. Ahn, M. Fiorentino, R. G. Beausoleil, N. Binkert, A. Davis, D. Fattal, N. P. Jouppi, M. McLaren, C. M. Santori, R. S. Schreiber, S. M. Spillane, D. Vantrease, and Q. Xu. Devices and architectures for photonic chip-scale integration. *Applied Physics A: Materials Science & Processing*, 95:989–997, June 2009.

[2] G. M. Amdahl. Validity of the single processor approach to achieving large scale computing capabilities. In *Proceedings of the April 18-20, 1967, spring joint computer conference*, AFIPS '67 (Spring), pages 483–485, New York, NY, USA, 1967. ACM.

[3] B. Arimilli, R. Arimilli, V. Chung, S. Clark, W. Denzel, B. Drerup, T. Hoefler, J. Joyner, J. Lewis, J. Li, N. Ni, and R. Rajamony. The percs high-performance interconnect. *High-Performance Interconnects, Symposium on*, 0:75–82, 2010.

[4] ASHRAE. 2008 ashrae environmental guidelines for datacom equipment. Technical report, American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc., Atlanta, GA, USA, 2008.

[5] T. I. Association. Tia/eia-644-a electrical characteristics of low voltage differential signaling (lvds) interface circuits. Technical report, Telecommunications Industry Association, Arlington, VA, USA, Feb. 2001.

[6] C. Batten, A. Joshi, J. Orcutt, A. Khilo, B. Moss, C. Holzwarth, M. Popovic, H. Li, H. Smith, J. Hoyt, F. Kartner, R. Ram, V. Stojanovic, and K. Asanovic. Building manycore processor-to-dram networks with monolithic silicon photonics. In *HOTI '08: Proceedings of the 2008 16th IEEE Symposium on High Performance Interconnects*, pages 21–30, Washington, DC, USA, 2008. IEEE Computer Society.

[7] G. Bell, J. Gray, and A. S. Szalay. Petascale computational systems. *IEEE Computer*, 39:110–112, 2006.

[8] J. Berger. A note on error detection codes for asymmetric channels. *Information and Control*, 4(1):68 – 73, 1961.

[9] S. Y. Borkar, P. Dubey, K. C. Kahn, D. J. Kuck, H. Mulder, S. S. Pawlowski, and J. R. Rattner. Platform 2015: Intel processor and platform evolution for the next decade. White paper, Technology Intel Magazine, April 2005. Available online (12 pages).

[10] F. Caignet, S. Delmas-Bendhia, and E. Sicard. The challenge of signal integrity in deep-submicrometer cmos technology. *Proceedings of the IEEE*, 89(4):556 –573, Apr. 2001.

[11] G. Campobello, G. Patane, and M. Russo. Parallel crc realization. *IEEE Transactions on Computers*, 52:1312–1319, 2003.

[12] L. Chen, N. Sherwood-Droz, and M. Lipson. Compact bandwidth-tunable microring resonators. *Opt. Lett.*, 32(22):3361–3363, 2007.

[13] F. Chiaraluce and R. Garello. Extended hamming product codes analytical performance evaluation for low error rate applications. *Wireless Communications, IEEE Transactions on*, 3(6):2353 – 2361, nov. 2004.

[14] D. M. Chiarulli, S. P. Levitan, S. J. Dickerson, J. D. Bakos, and J. Martin. Efficient optical communications using multibit differential signaling. In A. M. Earman & R. T. Chen, editor, *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 6126 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pages 140–147, Mar. 2006.

[15] M. J. Cianchetti, J. C. Kerekes, and D. H. Albonesi. Phastlane: a rapid transit optical routing network. *SIGARCH Comput. Archit. News*, 37(3):441–450, 2009.

[16] C. Constantinescu. Trends and challenges in vlsi circuit reliability. *Micro, IEEE*, 23(4):14 – 19, 2003.

[17] B. Crothers. Intel says to prepare for 'thousands of cores'. Online Article, CNET News, July 2008. Available online (12 pages).

[18] W. Dally and B. Towles. *Principles and Practices of Interconnection Networks*. Morgan Kaufmann, San Francisco, 2004.

[19] J. Dionne, L. Sweatlock, M. Sheldon, A. Alivisatos, and H. Atwater. Silicon-based plasmonics for on-chip photonics. *Selected Topics in Quantum Electronics, IEEE Journal of*, 16(1):295 –306, jan.-feb. 2010.

[20] A. Ejlali, B. Al-Hashimi, P. Rosinger, S. Miremadi, and L. Benini. Performability/energy tradeoff in error-control schemes for on-chip networks. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 18(1):1 –14, 2010.

[21] D. Ernst, N. S. Kim, S. Das, S. Pant, R. Rao, T. Pham, C. Ziesler, D. Blaauw, T. Austin, K. Flautner, and T. Mudge. Razor: A low-power pipeline based on circuit-level timing speculation. *Microarchitecture, IEEE/ACM International Symposium on*, 0:7, 2003.

[22] M. Fiorentino. personal communication about trimming work being done by Andrei Faraon at HP., 2010.

[23] B. J. Frey, D. B. Leviton, and T. J. Madison. Temperature-dependent refractive index of silicon and germanium. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 6273 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, July 2006.

[24] B. Fu and P. Ampadu. A dual-mode hybrid arq scheme for energy efficient on-chip interconnects. In O. Akan, P. Bellavista, J. Cao, F. Dressler, D. Ferrari, M. Gerla, H. Kobayashi, S. Palazzo, S. Sahni, X. S. Shen, M. Stan, J. Xiaohua, A. Zomaya,

G. Coulson, and M. Cheng, editors, *Nano-Net*, volume 3 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 74–79. Springer Berlin Heidelberg, 2009.

[25] B. Fu and P. Ampadu. On hamming product codes with type-ii hybrid arq for on-chip interconnects. *Trans. Cir. Sys. Part I*, 56:2042–2054, September 2009.

[26] G. A. Gibson, L. Hellerstein, R. M. Karp, and D. A. Patterson. Failure correction techniques for large disk arrays. In *Proceedings of the third international conference on Architectural support for programming languages and operating systems*, ASPLOS-III, pages 123–132, New York, NY, USA, 1989. ACM.

[27] D. Goldberg. What every computer scientist should know about floating-point arithmetic. *ACM Computing Surveys*, 23(1):5–48, March 1991.

[28] B. Guha, B. B. C. Kyotoku, and M. Lipson. Cmos-compatible athermal silicon microring resonators. *Opt. Express*, 18(4):3487–3493, Feb 2010.

[29] R. Hamming. Error detecting and error correcting codes. *Bell System Tech.*, 29:147–160, 1950.

[30] A. Harke, M. Krause, and J. Mueller. Low-loss singlemode amorphous silicon waveguides. *Electronics Letters*, 41(25):1377–1379, Dec. 2005.

[31] M. Haurylau, G. Chen, H. Chen, J. Zhang, N. Nelson, D. Albonesi, E. Friedman, and P. Fauchet. On-chip optical interconnect roadmap: Challenges and critical directions. *Selected Topics in Quantum Electronics, IEEE Journal of*, 12(6):1699–1705, Nov.-dec. 2006.

[32] G. Hendry, S. Kamil, A. Biberman, J. Chan, B. G. Lee, M. Mohiyuddin, A. Jain, K. Bergman, L. P. Carloni, J. Kubiatowicz, L. Oliker, and J. Shalf. Analysis of photonic networks for a chip multiprocessor using scientific applications. *Networks-on-Chip, International Symposium on*, 0:104–113, 2009.

[33] J. L. Hennessy and D. A. Patterson. *Computer Architecture: A Quantitative Approach*. Morgan Kaufmann, San Francisco, 2002.

[34] R. Ho. *On-Chip Wires: Scaling and Efficiency*. PhD thesis, Stanford University, 2003.

[35] Y. Hoskote, S. Vangal, A. Singh, N. Borkar, and S. Borkar. A 5-ghz mesh interconnect for a teraflops processor. *Micro, IEEE*, 27(5):51–61, Sept.-Oct. 2007.

[36] W. Huang, S. Ghosh, K. Sankaranarayanan, K. Skadron, and M. R. Stan. Hotspot: Thermal modeling for cmos vlsi systems. *IEEE Transactions on Component Packaging and Manufacturing Technology*, 2005.

[37] W. Huang, K. Sankaranarayanan, K. Skadron, R. J. Ribando, and M. R. Stan. Accurate, pre-rtl temperature-aware design using a parameterized, geometric thermal model. *IEEE Transactions on Computers*, 57(9):1277–1288, 2008.

[38] R. R. Iraschko and W. D. Grover. A highly efficient path-restoration protocol for management of optical network transport integrity. *IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS*, 18:779–794, 2000.

[39] R. R. Iraschko, W. D. Grover, S. M. IEEE, M. H. Macgregor, M. IEEE, and M. IEEE. A distributed real time path restoration protocol with performance close to centralized multi-commodity max flow. In *in IEEE Journal on Selected Areas in Communications, Special Issue on Recent Advances in Network Management and Operations*, pages 17–20, 1998.

[40] A. Joshi, C. Batten, Y.-J. Kwon, S. Beamer, I. Shamim, K. Asanovic, and V. Stojanovic. Silicon-photonic clos networks for global on-chip communication. In *NOCS '09: Proceedings of the 2009 3rd ACM/IEEE International Symposium on Networks-on-Chip*, pages 124–133, Washington, DC, USA, 2009. IEEE Computer Society.

[41] A. Kahng, B. Li, L.-S. Peh, and K. Samadi. Orion 2.0: A fast and accurate noc power and area model for early-stage design space exploration. In *DATE*, pages 423–428, April 2009.

[42] Y.-H. Kao and H. J. Chao. Blocon: A bufferless photonic clos network-on-chip architecture. In *Networks on Chip (NoCS), 2011 Fifth IEEE/ACM International Symposium on*, pages 81 –88, may 2011.

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

[44] B. R. Koch, A. W. Fang, O. Cohen, and J. E. Bowers. Mode-locked silicon evanescent lasers. *Opt. Express*, 15(18):11225–11233, Sep 2007.

[45] T. Koch. Opportunities and challenges in silicon photonics. In *Lasers and Electro-Optics Society, 2006. LEOS 2006. 19th Annual Meeting of the IEEE*, pages 677 –678, oct. 2006.

[46] P. Koka, M. O. McCracken, H. Schwetman, X. Zheng, R. Ho, and A. V. Krishnamoorthy. Silicon-photonic network architectures for scalable, power-efficient multi-chip systems. In *Proceedings of the 37th annual international symposium on Computer architecture*, ISCA '10, pages 117–128, New York, NY, USA, 2010. ACM.

[47] Y. Kokubun. Vertically coupled microring resonator filter for integrated add/drop node. *IEICE TRANSACTIONS on Electronics*, E88-C(3):349–362, Mar 2006.

[48] Y. Kokubun, N. Kobayashi, and T. Sato. Uv trimming of polarization-independent microring resonator by internal stress and temperature control. *Opt. Express*, 18(2):906–916, 2010.

[49] S. P. Levitan, D. M. Chiarulli, S. J. Dickerson, J. D. Bakos, and J. R. Martin. Power efficient communication using multi-bit-differential signaling. In *16th Annual IEEE-LEOS Workshop on Interconnections within High-Speed Digital Systems*. IEEE, May 2005.

[50] L. Liao, D. R. Lim, A. M. Agarwal, X. Duan, K. K. Lee, and L. C. Kimerling. Optical transmission losses in polycrystalline silicon strip waveguides: effects of waveguide dimensions, thermal treatment, hydrogen passivation, and wavelength. *J. Electron. Mater.*, 29(12):1380–1386, 2000.

[51] S. Lin and D. J. Costello, Jr. *Error Control Coding*. Prentice Hall, Upper Saddle River, NJ, Jun 2004.

[52] M. Lipson. Guiding, modulating, and emitting light on silicon-challenges and opportunities. *Lightwave Technology, Journal of*, 23(12):4222–4238, Dec. 2005.

[53] M. Lipson. Compact electro-optic modulators on a silicon chip. *Selected Topics in Quantum Electronics, IEEE Journal of*, 12(6):1520–1526, Nov.-dec. 2006.

[54] B. Little, S. Chu, W. Pan, D. Ripin, T. Kaneko, Y. Kokubun, and E. Ippen. Vertically coupled glass microring resonator channel dropping filters. *Photonics Technology Letters, IEEE*, 11(2):215–217, Feb 1999.

[55] W. Lu and S. Wong. A fast crc update implementation. In *Proceedings of the 14th Annual Workshop on Circuits, Systems and Signal Processing,*, pages 113–120, November 2003.

[56] P. Magnusson, M. Christensson, J. Eskilson, D. Forsgren, G. Hallberg, J. Hogberg, F. Larsson, A. Moestedt, and B. Werner. Simics: A full system simulation platform. *Computer*, 35(2):50–58, Feb 2002.

[57] G. Maire, L. Vivien, G. Sattler, A. Kazmierczak, B. Sanchez, K. B. Gylfason, A. Griol, D. Marris-Morini, E. Cassan, D. Giannone, H. Sohlström, and D. Hill. High efficiency silicon nitride surface grating couplers. *Opt. Express*, 16(1):328–333, 2008.

[58] M. M. K. Martin, D. J. Sorin, B. M. Beckmann, M. R. Marty, M. Xu, A. R. Alameldeen, K. E. Moore, M. D. Hill, and D. A. Wood. Multifacet's general execution-driven multiprocessor simulator (gems) toolset. *SIGARCH Comput. Archit. News*, 33(4):92–99, 2005.

[59] D. A. B. Miller. Rationale and challenges for optical interconnects to electronic chips. *Proceedings of the IEEE*, 88(6):728–749, Jun 2000.

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

[61] D. A. B. Miller, A. Bhatnagar, S. Palermo, A. Emami-Neyestanak, and M. Horowitz. Opportunities for optics in integrated circuits applications. In *Solid-State Circuits Conference, 2005. Digest of Technical Papers. ISSCC. 2005 IEEE International*, pages 86–87 Vol. 1, Feb. 2005.

[62] G. Moore. Cramming more components onto integrated circuits. *Electronics*, 38(8):114–117, apr 1965.

[63] K. Murakami and H. S. Kim. Optimal capacity and flow assignment for self-healing atm networks based on line and end-to-end restoration. *IEEE/ACM Trans. Netw.*, 6(2):207–221, 1998.

[64] A. Naeemi, J. Xu, A. Mule', T. Gaylord, and J. Meindl. Optical and electrical interconnect partition length based on chip-to-chip bandwidth maximization. *Photonics Technology Letters, IEEE*, 16(4):1221 –1223, 2004.

[65] J. Nelson. *The Physics of Solar Cells*. Imperial College Press, London, 2003.

[66] C. Nitta, M. Farrens, and V. Akella. Addressing system-level trimming issues in on-chip nanophotonic networks. In *High Performance Computer Architecture, 2011. HPCA 2011. IEEE 17th International Symposium on*, Feb. 2011.

[67] C. Nitta, M. Farrens, and V. Akella. Resilient microring resonator based photonic networks. In *Micro-44: Proceedings of the 44th Annual IEEE/ACM International Symposium on Microarchitecture (to appear)*, Dec. 2011.

[68] C. Nitta, K. Macdonald, M. Farrens, and V. Akella. Inferring packet dependencies to improve trace based simulation of on-chip networks. In *Networks-on-Chip (NOCS), 2011 Fifth ACM/IEEE International Symposium on*, May 2011.

[69] S. Pae, T. Su, J. Denton, and G. Neudeck. Multiple layers of silicon-on-insulator islands fabrication by selective epitaxial growth. *Electron Device Letters, IEEE*, 20(5):194–196, May 1999.

[70] Y. Pan, J. Kim, and G. Memik. Flexishare: Channel sharing for an energy-efficient nanophotonic crossbar. In *High Performance Computer Architecture, 2010. HPCA 2010. IEEE 16th International Symposium on*, Jan. 2010.

[71] Y. Pan, P. Kumar, J. Kim, G. Memik, Y. Zhang, and A. Choudhary. Firefly: illuminating future network-on-chip with nanophotonics. *SIGARCH Comput. Archit. News*, 37(3):429–440, 2009.

[72] B. Parhami. A multi-level view of dependable computing. *Computers & Electrical Engineering*, 20(4):347 – 368, 1994.

[73] L.-S. Peh, N. Agarwal, N. Jha, and T. Krishna. Garnet: A detailed on-chip network model inside a full-system simulator. *International Symposium on Performance Analysis of Systems and Software (ISPASS)*, 0, April 2009.

[74] K. Preston, B. Schmidt, and M. Lipson. Polysilicon photonic resonators for large-scale 3d integration of optical networks. *Opt. Express*, 15(25):17283–17290, 2007.

[75] R. Pyndiah. Near-optimum decoding of product codes: block turbo codes. *Communications, IEEE Transactions on*, 46(8):1003 –1010, aug 1998.

[76] Y. Qian, S. Kim, J. Song, G. P. Nordin, and J. Jiang. Compact and low loss silicon-on-insulator rib waveguide 90° bend. *Opt. Express*, 14(13):6020–6028, Jun 2006.

[77] V. Raghunathan, W. N. Ye, J. Hu, T. Izuhara, J. Michel, and L. Kimerling. Athermal operation of silicon waveguides: spectral, second order and footprint dependencies. *Opt. Express*, 18(17):17631–17639, Aug 2010.

[78] A.-M. Rahmani, I. Kamali, P. Lotfi-Kamran, A. Afzali-Kusha, and S. Safari. Negative exponential distribution traffic pattern for power/performance analysis of network on chips. In *VLSID '09: Proceedings of the 2009 22nd International Conference on VLSI Design*, pages 157–162, Washington, DC, USA, 2009. IEEE Computer Society.

[79] J. Richard Shewchuk. Adaptive precision floating-point arithmetic and fast robust geometric predicates. *Discrete & Computational Geometry*, 18:305–363, 1997. 10.1007/PL00009321.

[80] J. Schrauwen, D. V. Thourhout, and R. Baets. Trimming of silicon ring resonator by electron beam induced compaction and strain. *Opt. Express*, 16(6):3738–3743, 2008.

[81] Semiconductor Industry Association. International technology roadmap for semiconductors 2009. 2009.

[82] A. Shacham, K. Bergman, and L. P. Carloni. The case for low-power photonic networks on chip. In *DAC '07: Proceedings of the 44th annual Design Automation Conference*, pages 132–135, New York, NY, USA, 2007. ACM.

[83] A. Shacham, K. Bergman, and L. P. Carloni. On the design of a photonic network-on-chip. In *NOCS '07: Proceedings of the First International Symposium on Networks-on-Chip*, pages 53–64, Washington, DC, USA, 2007. IEEE Computer Society.

[84] C. Shannon. A mathematical theory of communications. *Bell System Tech.*, 27:370–423, 1948.

[85] R. A. Soref and B. R. Bennett. Electrooptical effects in silicon. *IEEE Journal of Quantum Electronics*, 23:123–129, Jan. 1987.

[86] D. Taillaert, P. Bienstman, and R. Baets. Compact efficient broadband grating coupler for silicon-on-insulator waveguides. *Opt. Lett.*, 29(23):2749–2751, 2004.

[87] L. Tang and D. A. B. Miller. Commentary: Metallic nanodevices for chip-scale optical interconnects. *Journal of Nanophotonics*, 3(1):010302–+, Mar. 2009.

[88] J. Teng, P. Dumon, W. Bogaerts, H. Zhang, X. Jian, X. Han, M. Zhao, G. Morthier, and R. Baets. Athermal silicon-on-insulator ring resonators byoverlaying a polymer cladding on narrowedwaveguides. *Opt. Express*, 17(17):14627–14633, Aug 2009.

[89] S. Thoziyoor, N. Muralimanohar, J. H. Ahn, and N. P. Jouppi. Cacti 6.0: A tool to model large caches. Technical Report HPL-2009-85, HP Laboratories, Palo Alto, CA, USA, Apr. 2009.

[90] S. Vangal, J. Howard, G. Ruhl, S. Dighe, H. Wilson, J. Tschanz, D. Finan, A. Singh, T. Jacob, S. Jain, V. Erraguntla, C. Roberts, Y. Hoskote, N. Borkar, and S. Borkar. An 80-tile sub-100-w teraflops processor in 65-nm cmos. *Solid-State Circuits, IEEE Journal of*, 43(1):29–41, Jan. 2008.

[91] D. Vantrease, N. Binkert, R. Schreiber, and M. H. Lipasti. Light speed arbitration and flow control for nanophotonic interconnects. In *Micro-42: Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture*, pages 304–315, New York, NY, USA, 2009. ACM.

[92] D. Vantrease, R. Schreiber, M. Monchiero, M. McLaren, N. P. Jouppi, M. Fiorentino, A. Davis, N. Binkert, R. G. Beausoleil, and J. H. Ahn. Corona: System implications of emerging nanophotonic technology. In *ISCA '08: Proceedings of the 35th International Symposium on Computer Architecture*, pages 153–164, Washington, DC, USA, 2008. IEEE Computer Society.

[93] H.-S. Wang, X. Zhu, L.-S. Peh, and S. Malik. Orion: A power-performance simulator for interconnection networks. *Microarchitecture, IEEE/ACM International Symposium on*, 0:294, 2002.

[94] H. M. G. Wassel, M. Tiwari, J. K. Valamehr, L. Theogarajan, J. Dionne, F. T. Chong, and T. Sherwood. Towards chip-scale plasmonic interconnects. In *WINDS 2010 Workshop on the Interaction between Nanophotonic Devices and Systems*, Dec 2010.

[95] Y. Xiong and L. G. Mason. Restoration strategies and spare capacity requirements in self-healing atm networks. *IEEE/ACM Trans. Netw.*, 7(1):98–110, 1999.

[96] Q. Xu, D. Fattal, and R. G. Beausoleil. Silicon microring resonators with 1.5-μm radius. *Opt. Express*, 16(6):4309–4315, 2008.

[97] Q. Xu, B. Schmidt, S. Pradhan, and M. Lipson. Micrometre-scale silicon electro-optic modulator. *Nature*, 435(7040):325–327, 2005.

[98] T. Yin, R. Cohen, M. M. Morse, G. Sarid, Y. Chetrit, D. Rubin, and M. J. Paniccia. 31 ghz ge n-i-p waveguide photodetectors on silicon-on-insulator substrate. *Opt. Express*, 15(21):13965–13971, 2007.

[99] S. J. B. Yoo, V. Akella, R. Amirtharajah, B. Baas, K. Bergman, S. Fan, J. Harris, M. Lipson, D. A. B. Miller, and J. Shalf. Balanced computing with nanophotonic interconnects. In *IEEE Lasers and Electro-Optics Society, 2008. LEOS 2008. 21st Annual Meeting of the*, pages 368 –369, nov. 2008.

[100] X. Zheng, P. Koka, M. O. McCracken, H. Schwetman, J. G. Mitchell, J. Yao, R. Ho, K. Raj, and A. V. Krishnamoorthy. Energy-efficient error control for tightly coupled systems using silicon photonic interconnects. *J. Opt. Commun. Netw.*, 3(8):A21–A31, Aug 2011.

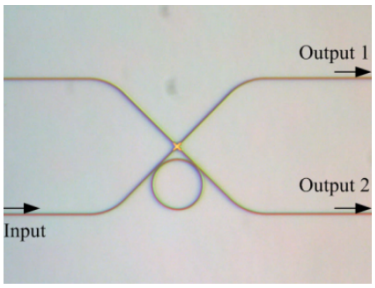
[101] L. Zhou, K. Okamoto, and S. Yoo. Athermalizing and trimming of slotted silicon microring resonators with uv-sensitive pmma upper-cladding. *Photonics Technology Letters, IEEE*, 21(17):1175–1177, Sept.1, 2009.

[102] L. Zhou and A. W. Poon. Silicon electro-optic modulators using p-i-n diodes embedded 10-micron-diameter microdisk resonators. *Opt. Express*, 14(15):6851–6857, 2006.

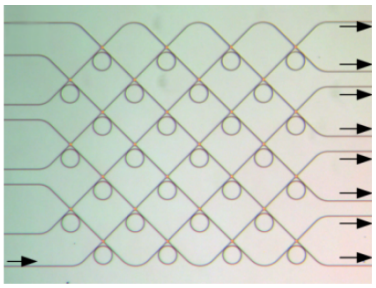
Appendix A

Device Fabrication

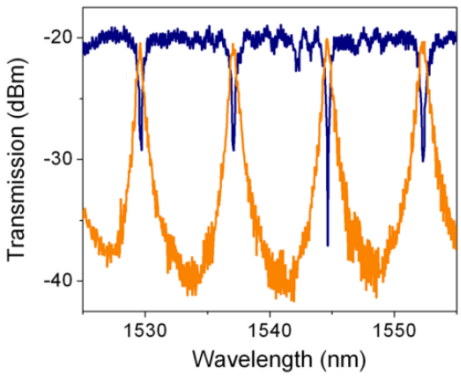
This appendix includes a detailed discussion of how vertically coupled microrings and multiple layers of photonics can be constructed. Micrographs and transmission spectra taken from microring based crossbar switches are shown in this appendix. These micror-



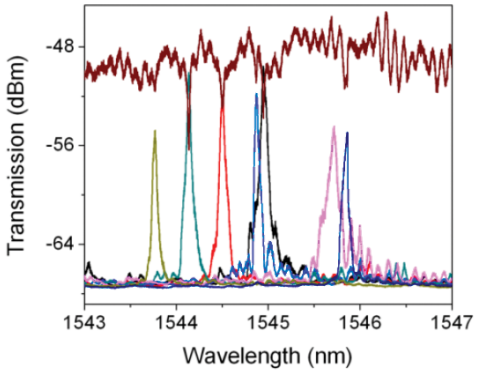
(a) Micrograph of 2x2 X-Bar



(b) Micrograph of 8x8 X-Bar



(c) Transmission Spectra of 2x2 X-Bar



(d) Transmission Spectra of 8x8 X-Bar

Figure A.1: Micrographs of 2x2 (a) and 8x8 (b) Optical Crossbars, and Transmission Spectra for Corresponding 2x2 (c) and 8x8 (d) Crossbars.

ing based crossbars have been fabricated and tested by researchers at the University of California, Davis.

Figure A.1(a) shows an example of some optical networks that have been built and tested in the lab. The waveguide width is 400nm, while the microring resonator radius is 10μm and the gap between the waveguides and microring resonator is 250nm. Figure A.1(c) shows the transmission spectra for the two complementary output ports. The transmission loss of the planar waveguide crossing is $\sim$1dB, higher than the theoretically predicted value of 0.1dB, which is mainly due to fabrication imperfections. The microring resonator is strongly coupled to the orthogonal crossing waveguides, with $\sim$0.3dB resonance transmission loss. The resonance FSR is 7.6nm, and the resonance bandwidth is $\sim$0.4nm. Figure A.1(b) and (d) show the 8x8 optical crossbar and its transmission spectra. Due to a fabrication error, two resonance channels deviated from their designed position. However, the phase error can be compensated for using post-fabrication techniques, such as e-beam and UV trimming [80].

Figure A.2 shows the steps involved in fabricating optical microrings. Vertically coupled microring resonators can be built on silica material with controlled coupling efficiency and signal routing flexibility [54, 47] – however, it is more difficult to realize them in silicon material because of its high index contrast and the lack of a deposition method for crystalline silicon. Therefore, it is assumed in this example that epitaxial growth is used to stack several layer of crystalline silicon as material platform for our microring resonator-based optical network [69].

The optical microring resonator fabrication begins with a SOI wafer. The SOI device layer thickness is 0.25μm, and the Buried Oxide (BOX) layer thickness is 2μm. The relatively thick BOX layer can effectively reduce the waveguide electric field leakage into the silicon substrate. The first layer of waveguides, together with small islands, is patterned on the silicon device layer of the SOI wafer using photolithography and Reactive Ion Etching (RIE). 600nm Plasma Enhanced Chemical Vapor Deposition (PECVD) is used to cover the whole waver with Silicon Dioxide (SiO_2). The deposited oxide layer follows the topology of the first silicon waveguide layer. To eliminate the surface fluctuation, Chemical Mechanical Polishing (CMP) is used to thin down the top SiO_2 layer to 450nm such that

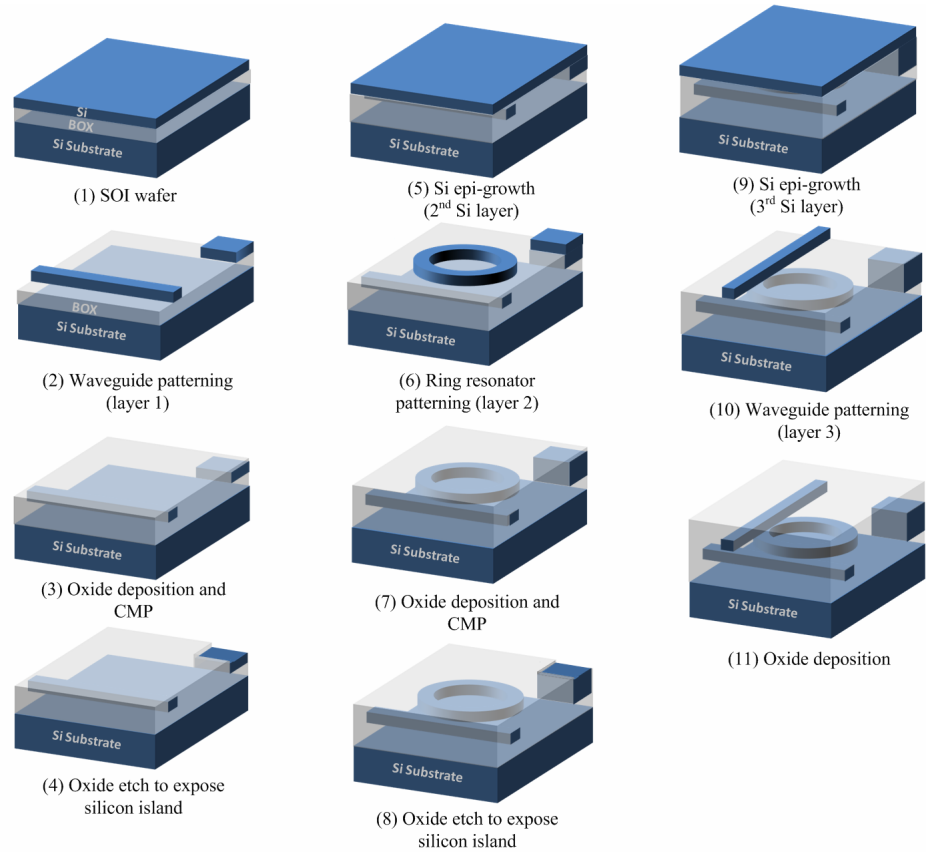


Figure A.2: Microring Resonator Fabrication Process

only 200nm of SiO_2 is left on top of the silicon waveguides.

Poly-crystalline or amorphous silicon material can be readily deposited on an oxide layer using a Low-Pressure Chemical Vapor Deposition (LPCVD) method. Although various annealing techniques have been tried to reduce the silicon grain boundary induced scattering loss [30, 74, 50], polysilicon waveguide loss is still relatively high (several dB/cm) compared to that of crystalline silicon waveguide [74]. Hence, crystalline silicon should be used for all the layers of our 3-dimensional stack devices. Crystalline silicon seeds are needed for epitaxial growth of a layer of crystalline silicon. The wafer is patterned to only expose the seeds region, and then the top oxide is etched off using a Buffered Oxide Etchant (BOE) solution.

Crystalline silicon is grown to cover the whole wafer, and the surface is again planarized by CMP [69]. The remaining silicon layer is 250nm thick, the same as the first waveguide layer. Microring resonators are patterned in this layer, and their alignment with

the waveguides beneath them is done using the alignment marks on the first waveguide layer. These two layers are vertically separated by a 200nm oxide layer. When the waveguide and ring resonators are close enough (within their evanescent tail range), lightwaves propagating in the first waveguide layer can couple into the microring resonator to form resonance if the resonance condition is satisfied.

To form the third waveguide layer, crystalline silicon is again grown from the seeds window, planarized by CMP, and patterned to form another layer of waveguides. The top waveguides are arranged perpendicular to the bottom waveguides to eliminate any cross-coupling between these two waveguide layers. Lightwave signals on the resonance wavelength can couple from the bottom waveguides to the top waveguides via the microring resonators sandwiched between them, and signals on the non-resonance wavelengths remain on the same layer without tunneling to the other layers.

For active microring resonators, the electronic devices (p-i-n diodes) can be fabricated using standard Complementary Metal–Oxide–Semiconductor (CMOS) processes [102]. The heavily-doped n^+ and p^+ electrodes can be positioned in the thin slab region around the microring resonator, and metal wires can connect to the electrodes through the contact holes inside the oxide layer.

Appendix B

Complete Thermal Results

This appendix presents the complete thermal results for the three networks analyzed. Figure B.1 shows the thermal results of DCAF simulation in K. Figures B.2 and B.3 show the similar thermal results for CrON and FCON respectively. These simulations were run at an ambient temperature of 318K. Figures B.4, B.5 and B.6 show the same thermal results for DCAF, CrON and FCON respectively, but the scales have been changed to enhance detail.

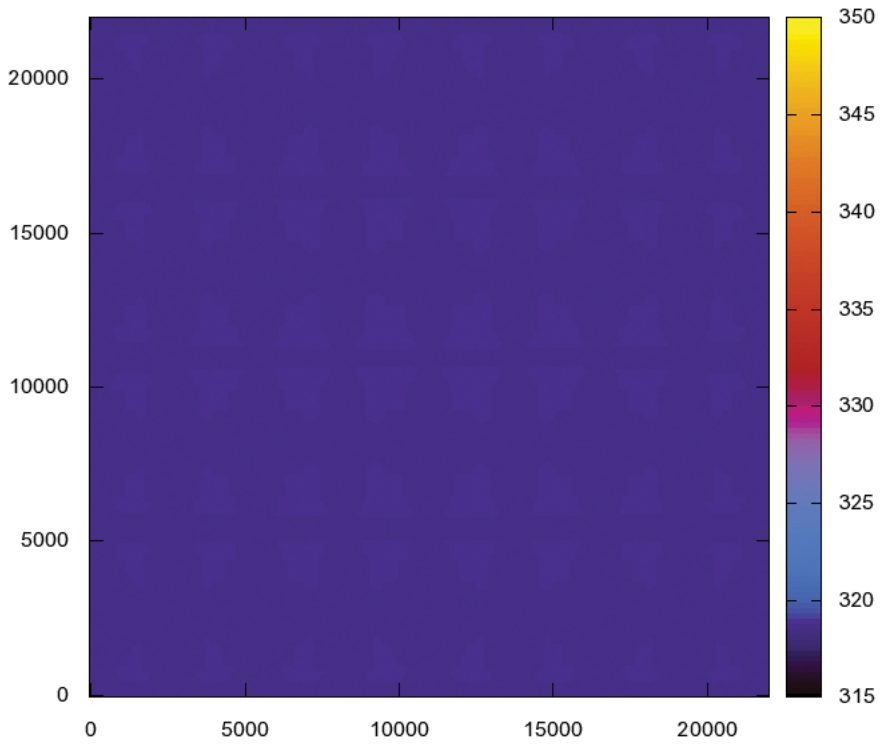


Figure B.1: Temperature (K) vs. X,Y (μm) for DCAF with 318K (45°C) Ambient

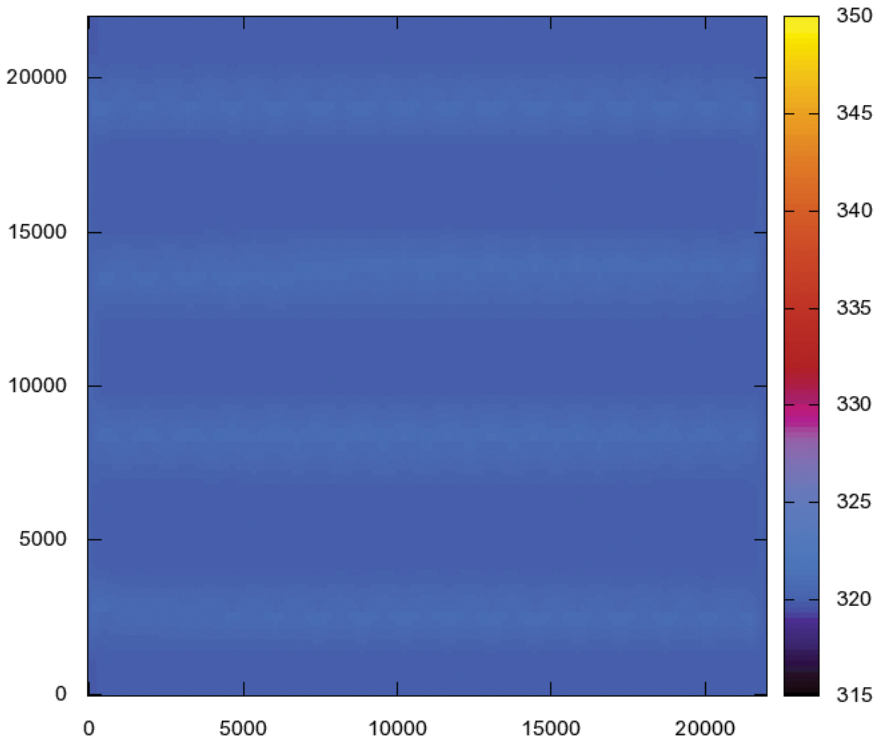


Figure B.2: Temperature (K) vs. X,Y (μm) for CrON with 318K (45°C) Ambient

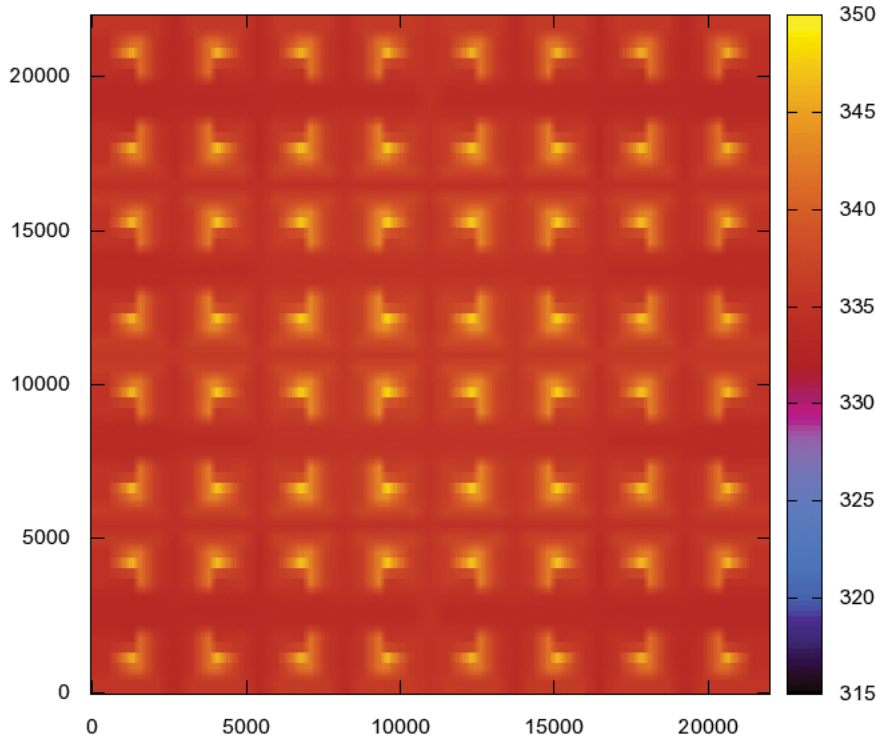


Figure B.3: Temperature (K) vs. X,Y (μm) for FCON with 318K (45°C) Ambient

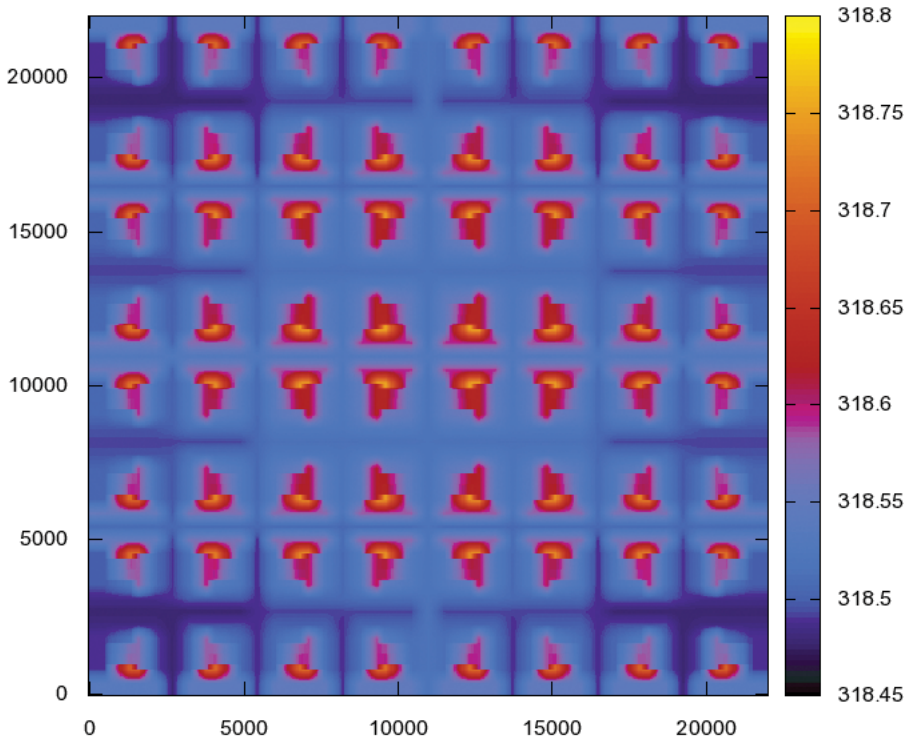


Figure B.4: Temperature (K) vs. X,Y (μm) for DCAF with 318K (45°C) Ambient

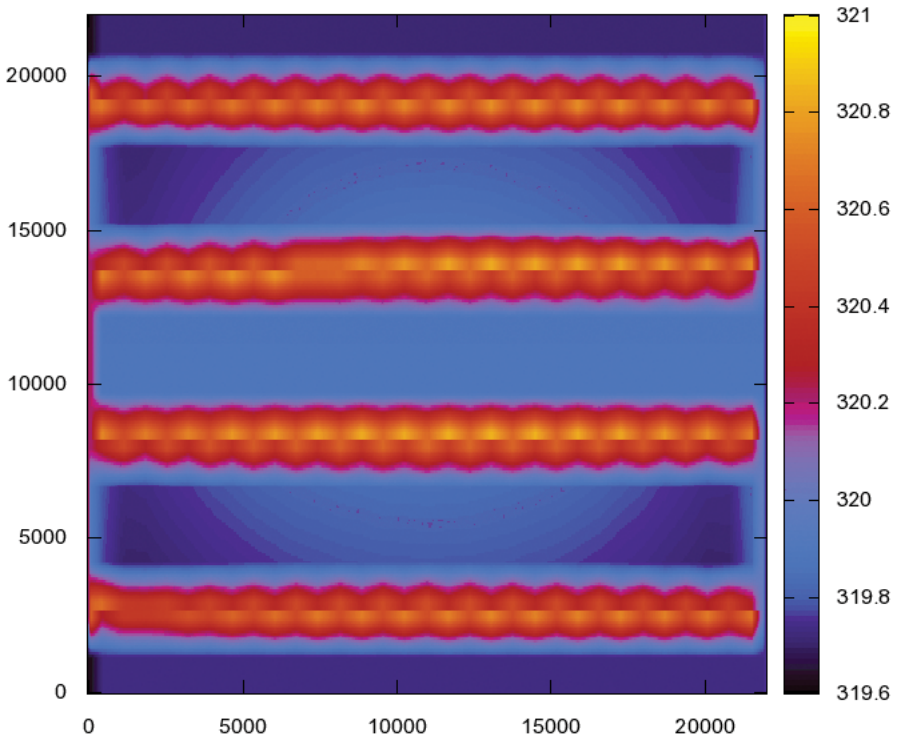


Figure B.5: Temperature (K) vs. X,Y (μm) for CrON with 318K (45°C) Ambient

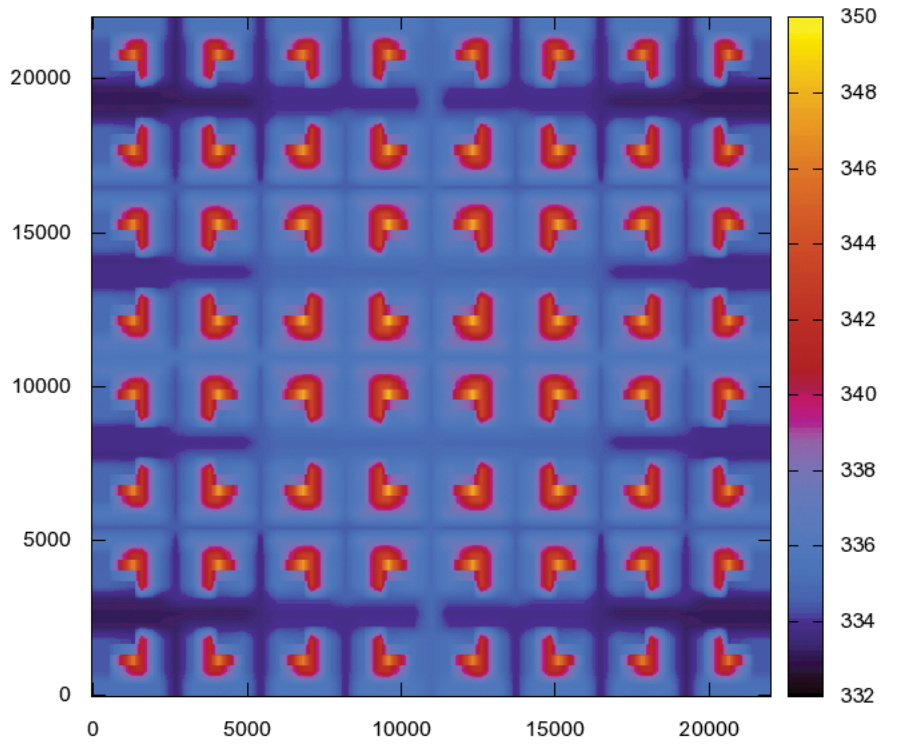


Figure B.6: Temperature (K) vs. X,Y (μm) for FCON with 318K (45°C) Ambient